The tiny eukaryote *Ostreococcus* provides genomic insights into the paradox of plankton speciation

Brian Palenik[a,b], Jane Grimwood[c], Andrea Aerts[d], Pierre Rouzé[e], Asaf Salamov[d], Nicholas Putnam[d], Chris Dupont[a], Richard Jorgensen[f], Evelyne Derelle[g], Stephane Rombauts[h], Kemin Zhou[d], Robert Otillar[d], Sabeeha S. Merchant[i], Sheila Podell[j], Terry Gaasterland[j], Carolyn Napoli[f], Karla Gendler[f], Andrea Manuell[k], Vera Tai[a], Olivier Vallon[l], Gwenael Piganeau[g], Séverine Jancek[g], Marc Heijde[m], Kamel Jabbari[m], Chris Bowler[m], Martin Lohr[n], Steven Robbens[h], Gregory Werner[d], Inna Dubchak[d], Gregory J. Pazour[o], Qinghu Ren[p], Ian Paulsen[p], Chuck Delwiche[q], Jeremy Schmutz[c], Daniel Rokhsar[d], Yves Van de Peer[h], Hervé Moreau[g], and Igor V. Grigoriev[b,d]

[a]Scripps Institution of Oceanography, University of California at San Diego, La Jolla, CA 92093-0202; [c]Joint Genome Institute and Stanford Human Genome Center, Stanford University School of Medicine, 975 California Avenue, Palo Alto, CA 94304; [d]U.S. Department of Energy Joint Genome Institute, 2800 Mitchell Drive, Walnut Creek, CA 94598; [e]Laboratoire Associé de l'Institut National de la Recherche Agronomique (France), Ghent University, Technologiepark 927, B-9052 Ghent, Belgium; [f]Department of Plant Sciences, University of Arizona, 303 Forbes Building, Tucson, AZ 85721-0036; [g]Observatoire Océanologique, Laboratoire Arago, Centre National de la Recherche Scientifique/Université Pierre et Marie Curie Paris 6, Unité Mixte de Recherche 7628, BP 44, 66651 Banyuls sur Mer Cedex, France; [h]Department of Plant Systems Biology, Flanders Interuniversity Institute for Biotechnology, Ghent University, Technologiepark 927, B-9052 Ghent, Belgium; [i]Department of Chemistry and Biochemistry, University of California, Box 951569, Los Angeles, CA 90095; [j]Scripps Genome Center, Scripps Institution of Oceanography, University of California at San Diego, 9500 Gilman Drive, La Jolla, CA 92093-0202; [k]Department of Cell Biology, The Scripps Research Institute, 10550 North Torrey Pines Road, La Jolla, CA 92037; [l]Institut de Biologie Physico-Chimique, Centre National de la Recherche Scientifique/Université Paris 6, Unité Mixte de Recherche 7141, 13, Rue Pierre et Marie Curie, 75005 Paris, France; [m]Département de Biologie, Ecole Normale Supérieure, Formation de Recherche en Evolution 2910, Centre National de la Recherche Scientifique, 46, Rue D'Ulm, 75230 Paris Cedex 05, France; [n]Institut für Allgemeine Botanik, Johannes Gutenberg-Universität, D-55099 Mainz, Germany; [o]Program in Molecular Medicine, University of Massachusetts Medical School, Suite 213, Biotech II, 373 Plantation Street, Worcester, MA 01605; [p]The Institute for Genomic Research, 9712 Medical Center Drive, Rockville, MD 20850; and [q]Cell Biology and Molecular Genetics, University of Maryland, H. J. Patterson Hall, Building 073, College Park, MD 20742-5815

Edited by Robert Haselkorn, University of Chicago, Chicago, IL, and approved March 13, 2007 (received for review December 12, 2006)

The smallest known eukaryotes, at ≈1-μm diameter, are *Ostreococcus tauri* and related species of marine phytoplankton. The genome of *Ostreococcus lucimarinus* has been completed and compared with that of *O. tauri*. This comparison reveals surprising differences across orthologous chromosomes in the two species from highly syntenic chromosomes in most cases to chromosomes with almost no similarity. Species divergence in these phytoplankton is occurring through multiple mechanisms acting differently on different chromosomes and likely including acquisition of new genes through horizontal gene transfer. We speculate that this latter process may be involved in altering the cell-surface characteristics of each species. In addition, the genome of *O. lucimarinus* provides insights into the unique metal metabolism of these organisms, which are predicted to have a large number of selenocysteine-containing proteins. Selenoenzymes are more catalytically active than similar enzymes lacking selenium, and thus the cell may require less of that protein. As reported here, selenoenzymes, novel fusion proteins, and loss of some major protein families including ones associated with chromatin are likely important adaptations for achieving a small cell size.

green algae | picoeukaryote | genome evolution | selenium | synteny

Phytoplankton living in the oceans perform nearly half of total global photosynthesis (1). Eukaryotic phytoplankton exhibit great diversity that contrasts with the lower apparent diversity of ecological niches available to them in aquatic ecosystems. This observation, know as the "paradox of the plankton," has long puzzled biologists (2). By providing molecular level information on related species, genomics is poised to provide new insights into this paradox.

Picophytoplankton, with cell diameters <2 μm, play a significant role in major biogeochemical processes, primary productivity, and food webs, especially in oligotrophic waters. Within this size class, the smallest known eukaryotes are *Ostreococcus tauri* and related species. Although more similar to flattened spheres in shape, these organisms are ≈1 μm in diameter (3, 4) and have been isolated or detected from samples of diverse geographical origins (5–8). They belong to the Prasinophyceae, an early diverging class within the green plant lineage, and have a strikingly simple cellular organization, with no cell wall or flagella, and with a single chloroplast and mitochondrion (4). Recent work has shown that small-subunit rDNA sequences of *Ostreococcus* from cultures and environmental samples cluster into four different clades that are likely distinct enough to represent different species (6, 9).

Here we report on the gene content, genome organization, and deduced metabolic capacity of the complete genome of *Ostreococcus* sp. strain CCE9901 (7), a representative of surface–ocean adapted *Ostreococcus*, referred to here as *Ostreococcus lucimarinus.* We compare it to the analogous features of the related species *O. tauri* strain OTH95 (10). Our results show that many processes have been involved in the evolution and speciation of even these sister organisms, from dramatic changes in genome structure to significant differences in metabolic capabilities.

Results

Gene Content. *O. lucimarinus* is the first closed and finished genome of a green alga and as such will provide a great resource for in-depth analysis of genome organization and the processes of eukaryotic genome evolution. *O. lucimarinus* has a nuclear genome size of 13.2 million base pairs found in 21 chromosomes, as compared with a genome size for *O. tauri* of 12.6 million base pairs found in 20

Author contributions: J.G. and A.A. performed research; B.P., P.R., A.S., N.P., C. Dupont, R.J., E.D., S. Rombauts, K.Z., R.O., S.S.M., S.P., T.G., C.N., K.G., A.M., V.T., O.V., G.P., S.J., M.H., K.J., C.B., M.L., S. Robbens, G.W., I.D., G.J.P., Q.R., I.P., C. Delwiche, J.S., D.R., Y.V.d.P., H.M., and I.V.G. analyzed data; and B.P., P.R., C. Dupont, R.J., K.Z., S.S.M., M.L., H.M., and I.V.G. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

Abbreviations: Chr *n*, chromosome *n*; GC, guanine plus cytosine.

Data deposition: The *O. lucimarinus* genome sequence, predicted genes, and annotations reported in this paper have been deposited in the GenBank database (accession nos. CP000581–CP000601 for Chr 1 through Chr 21). The *O. lucimarinus* strain (CCE9901) used here has been deposited in the Provasoli-Guillard Culture Collection of Marine Phytoplankton (accession no. CCMP2514).

[b]To whom correspondence may be addressed. E-mail: bpalenik@ucsd.edu or ivgrigoriev@lbl.gov.

This article contains supporting information online at www.pnas.org/cgi/content/full/0611046104/DC1.

Table 1. Summary of predicted genes in *Ostreococcus* sp. genomes

Properties	*O. lucimarinus*	*O. tauri*
Genome size, Mbp	13.2	12.6
Chromosomes	21	20
No. of genes	7,651	7,892
Multiexon genes, %	20	25
Supported by, %		
Multiple methods	28	19
Genome conservation	65	73
Homology to another strain	93	92
Homology to SwissProt	84	79
ESTs	28	21
Peptides	13	N/D
Average gene size, bp	1,284	1,245
Transcript size, bp	1,234	1,175
No. of exons per gene	1.27	1.57
Exon size, bp	970	750
Intron size, bp	187	126

N/D, not determined.

chromosomes (10) (Table 1). For comparison here, both genomes were annotated by using the same tools, as described in *Methods*.

We predicted and annotated 7,651 genes in the genome of *O. lucimarinus*, and 7,892 genes are found in the genome of *O. tauri*. Overall gene content is similar between the genomes (Table 1). Approximately one-fifth of all genes in both genomes have multiexon structure, most of which belong to chromosome 2 (Chr 2), and have the introns of unusual size and structure that were reported earlier for *O. tauri* (10). A total of 6,753 pairs of orthologs have been identified between genes in the two *Ostreococcus* species with an average coverage of 93% and an average amino acid identity of 70%. A comparison of the amino acid identity between other sister taxa shows that they are more divergent than characterized species of *Saccharomyces* with similar levels of overall synteny [supporting information (SI) Table 2].

Approximately 5–6% of gene models are genome-specific and do not display homology to the other species (SI Table 3). These are mostly due to lineage-specific gene loss or acquisition or remaining gaps in the *O. tauri* sequence. The number of lineage-specific duplications is also low, 9% for *O. lucimarinus* and 4% for *O. tauri*, mostly because of several segmental duplications.

Genome Structure. Based on analysis of gene content, orthology, and DNA alignments, 20 chromosomes in each genome have a counterpart in the other species. Eighteen of these 20 are highly syntenic (Fig. 1) and formed the core of the ancestral *Ostreococcus* genome. The remaining two chromosomes of *O. tauri* (Chr 2 and Chr 19) and three chromosomes of *O. lucimarinus* (Chr 2, Chr 18,

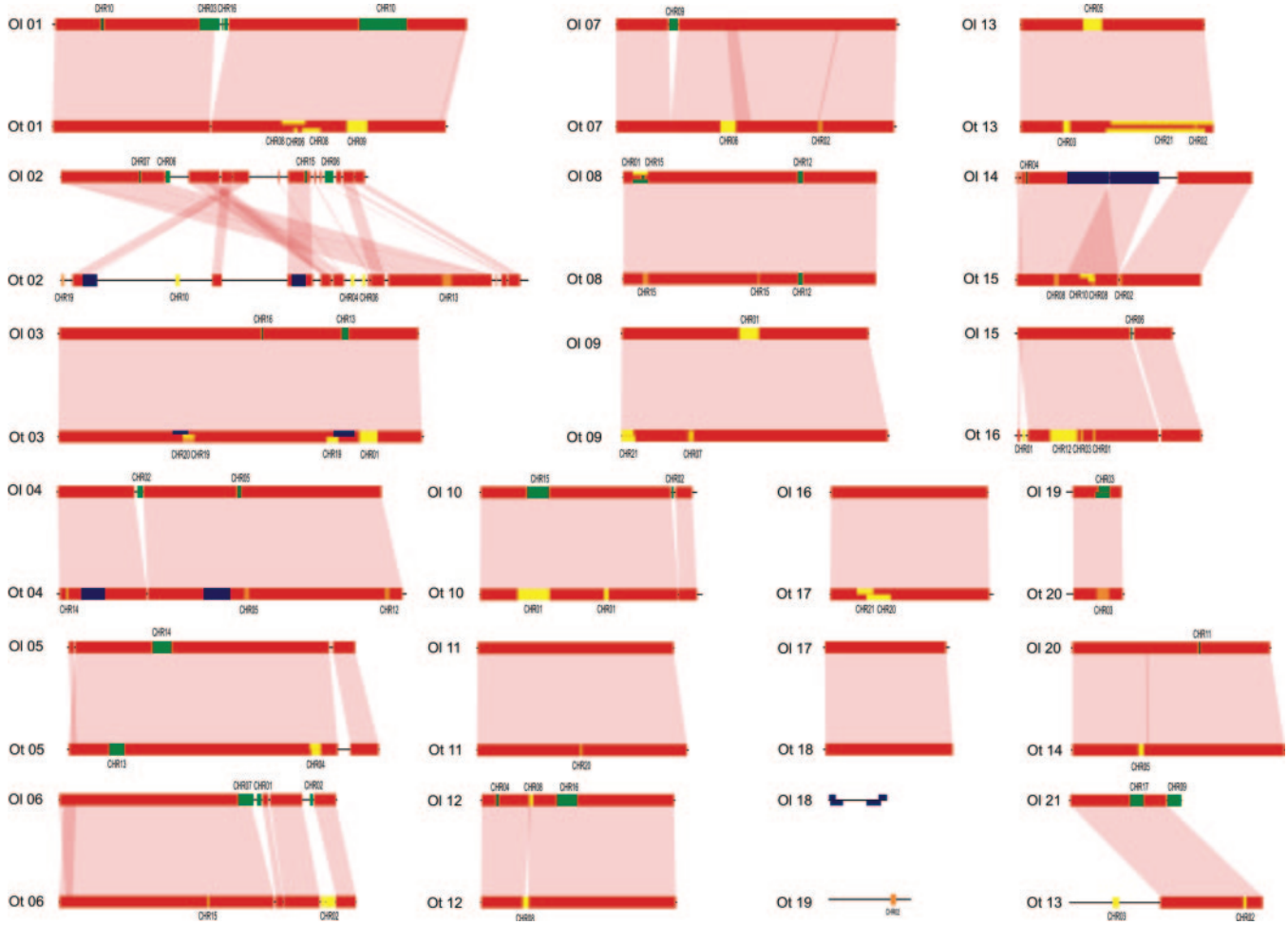

Fig. 1. Synteny between the chromosomes of *O. tauri* (Ot) and *O. lucimarinus* (Ol). Depicted areas in red show collinear regions (conserved gene order and content) as described in *Methods*. Blocks of different colors denote different sorts of duplications: blue, an internally duplicated segment; green, a duplicated segment that is collinear with a segment on a different chromosome in both Ot and Ol; yellow, a duplicated segment that is collinear with a segment on a different chromosome in Ol; orange, a duplicated segment that is collinear with a segment on a different chromosome in Ot.

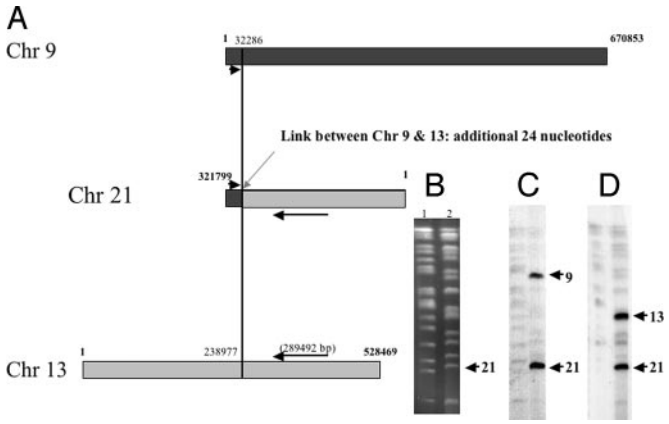

Fig. 2. Origin of the new *O. lucimarinus* chromosome, Chr 21. This chromosome was recently formed from pieces of Chr 9 and Chr 13. (*A*) Map of Chr 21. (*B*) Pulsed-field gel electrophoresis analysis of the *Ostreococcus* sp. genome migration for 72 h. Lane 1, *O. tauri* genome; lane 2, *O. lucimarinus* genome. (*C*) Results of hybridization with a probe from Chr 9. (*D*) Results of hybridization with a probe from Chr 13.

and Chr 21) (Figs. 1 and 2) are very distinct, not only from the core genome but also between the species.

Chr 2. In contrast to most other chromosomes, genes on Chr 2 are greatly rearranged between the two species as indicated by the absence of synteny (Fig. 1, synteny coded in red). These rearrangements are largely localized to regions of Chr 2 with distinctly lower guanine plus cytosine (GC) content, ≈15% less than coding sequence in the rest of the genome (SI Fig. 4). The genes found in the low-GC region of both species are still very closely related. This suggests that, although the rate of intrachromosomal rearrangement has been greatly increased in this part of the genome, the mutation rate remains the same. Small differences in rates of intrachromosomal rearrangement have been noted, for example in *Drosophila* (11), but not as dramatically as shown here. Transposons, which were found in higher abundance in Chr 2, may play an important role in these rearrangements. Interestingly, there are more types and absolute numbers of transposons in *O. tauri* than in *O. lucimarinus*.

Remarkably, pairs of converging genes, i.e., on opposite strand and sharing their 3′ side, are conserved in the low-GC region. Of the 174 genes found in both species, 122 are in such a "convergent pair" situation. When there are ESTs representing one or both transcripts in such pairs, they always show a large overlap of the transcripts on their 3′ side, not only 3′ UTRs but often significant parts of the coding sequences (e.g., Apm1/Cug1, Sen1/Pwp2, Coq4/Cup62, HecR/Cup201, and SufE/Spt4). This may indicate an interaction between the genes at the expression level, such as a RNAi-like down-regulation of one gene by the expression of the other. Some of these pairs may be recent ad hoc interactions recruited in *Ostreococcus* and nearby lineages, but others may be more ancient, and these will help in understanding gene networks in organisms such as land plants.

Contrary to the rest of the genome, most of the genes in Chr 2 are split by many introns (up to 15). Of the 180 genes in *O. lucimarinus*, 108 are split with a total of 419 introns. Most of the introns (395) form a special class, which differs from the "canonical introns" found in the rest of the genome (see also ref. 10), being smaller (40–65 bp), with poorly conserved splice-site motifs and no clear branch-point motif. A few canonical introns (24 of 419) occur in some genes, sometimes in combination with small introns. In most cases, positions of introns are conserved between the orthologs. However, a few genes have many small introns in one strain but either none or far fewer introns in another. The comparative analysis of the two species of *Ostreococcus* is casting some light on "raison d'être" of the low-GC region of Chr 2. The striking correlation between low GC content, high transposon density, and increased shuffling rate suggests a mechanism by which a local compositional bias is responsible for an enhanced activity of transposons and faster loss of synteny. A direct effect of this is to forbid interstrain crossing, because pairing of Chr 2 would not be possible, and eventual aneuploid offspring of such crossing would not be viable. The genes for meiosis have been noted in *O. tauri* (10) and are present in *O. lucimarinus* as well. In this view, Chr 2 would be a speciation chromosome, maintaining the strain in genetic isolation from its relatives.

Chr 18 of* O. lucimarinus *(Chr 19 of* O. tauri*). Chr 18 and Chr 19 are the smallest chromosomes of *O. lucimarinus* and *O. tauri*, with 83 and 131 predicted genes, respectively. Only 30 genes in *O. lucimarinus* Chr 18 have an ortholog in the *O. tauri* genome, including eight in Chr 19. Using VISTA (12) only 15% of the *O. lucimarinus* Chr 18 nucleotide sequence can be aligned with *O. tauri* genome including 5% aligned with Chr 19. For comparison, 80–90% of other *O. lucimarinus* chromosomes including Chr 2 can be aligned with their counterparts in *O. tauri* (SI Fig. 5).

Functions of two-thirds of Chr 18 genes are unknown while more than a half of them are supported by either ESTs or DNA conservation with the *O. tauri* genome. Many of the functionally annotated genes on Chr 18 of *O. lucimarinus* are related to sugar biosynthesis, modification, or transport, which suggests that Chr 18 may take part in a specific process.

Several of the Chr 18 genes are *O. lucimarinus*-specific, which suggests ongoing adaptation. One interesting example is gene OSTLU 28425. This is predicted to be similar to a UDP-*N*-acetylglucosamine 2-epimerase, which would produce UDP-*N*-acetylmannosamine. It is phylogenetically related to similar enzymes in bacteria only, and one of the top BLASTp hits is to the marine bacterium *Microscilla marina* ATCC 23134 (e-92). This seems a likely candidate for recent horizontal gene transfer into *O. lucimarinus*, as well as the majority of genes on Chr 18 that do not show homology to any other known proteins.

Similar sugar-related differences have been seen in the genomes of marine cyanobacterial species that coexist with *Ostreococcus*. It has been shown that apparently horizontally transferred genes in cyanobacteria are often glycosyltransferases (13). It was hypothesized that horizontal gene transfer makes available genes for the constant alteration of cell-surface glycosylation that would help the phytoplankton "disguise" itself from phages or grazers (13), and the results reported here suggest that this is an emerging theme in phytoplankton speciation.

Chr 18 and Chr 2 in *O. lucimarinus* have lower GC content than the rest of the genome as reported earlier for *O. tauri* (10). Principal component analysis of codon usage in both genomes shows that most of the chromosomes in each of the genomes are clustered together (Fig. 3). Within each genome, significant differences in codon usage have been observed between the core genome, Chr 2 (in particular, low-GC regions), and Chr 18 of *O. lucimarinus* (Chr 19 of *O. tauri*). The pattern of the segregation of chromosomes along the first principal component on Fig. 3 correlates with their GC content. A parallel shift along the first two components for all chromosomes except Chr 18 of *O. lucimarinus* and Chr 19 of *O. tauri* can describe differences in codon usage between the genomes and may reflect a general adaptation process. It is impossible to explain both the low similarity on the DNA and protein level between Chr 18 and Chr 19 and the differences in codon usage bias by classical evolutionary paradigms. Rather, they can best be explained by acquisition of genetic material for these two chromosomes from external sources after the divergence of the two species. With the exception of some examples as noted above, however, weak or undetected similarities between genes on these chromosomes and other known genes make it difficult to prove this with phylogenetic analysis.

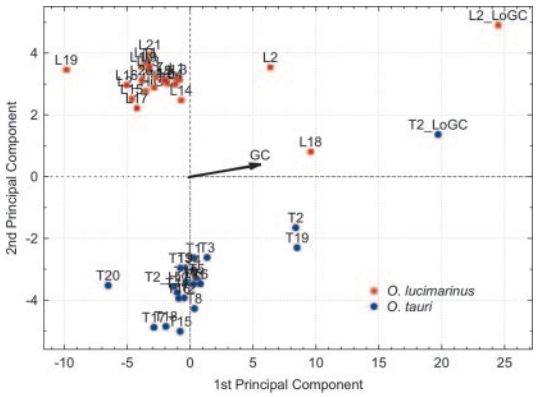

Fig. 3. Principal component analysis of *Ostreococcus* genomes.

Chr 21. Chr 21 is present only in *O. lucimarinus* and corresponds to a fusion between a small fragment of Chr 9 and a bigger fragment of Chr 13, with a short intervening sequence of 24 nt (Fig. 2). The recent origin is indicated by the fact that duplicated regions are almost 100% identical, with only 5 nt differing from the original chromosome. The existence of this chromosome has been experimentally confirmed (Fig. 2 *B–D*).

Intrachromosomal rearrangements. There are several internal duplications on Chr 2, 3, 4, and 8 of *O. tauri* and a large block of 142 kbp duplicated on Chr 14 of *O. lucimarinus* (Fig. 1). Spontaneous duplication of large chromosomal segments has been observed in yeast (14), and a similar process appears to be occurring at a significant rate during speciation of *Ostreococcus*. Surprisingly, almost all of these duplications are recent changes because none are observed on the corresponding chromosomes of the counterpart species (except Chr 8 and 12). Because gene sequence and order are so well conserved in the genus, this suggests that large chromosomal duplications were infrequent in the period preceding separation of the two species. It is unfortunately not possible yet to understand whether these duplications could have helped cause the speciation or occurred much later.

As seen in these three major chromosomal differences between the *O. tauri* and *O. lucimarinus* genomes, as well as some smaller intrachromosomal duplications, the speciation of these sister organisms is not accompanied by a single type of genome structural divergence, but multiple types, likely occurring at different time scales.

Environmental Adaptations. Most of the characterization of phytoplankton diversity traditionally has focused on pigment and morphological characteristics, and occasionally the utilization of nutrients, for example (15). The availability of the predicted proteomes of two closely related species of photosynthetic eukaryotes from different ecological niches allows some new insights into the role of micronutrients (metals and vitamins) in their ecological strategies and speciation relative to each other and other phytoplankton.

Selenoproteins. *Ostreococcus* has genes for a surprising number of selenocysteine-containing proteins relative to its genome size. Selenoproteins are encoded by coding sequences in which TGA, instead of being read as a stop codon, is recoded to selenocysteine if a control element (called SECIS) is encountered downstream in the 3′ UTR of the transcript in eukaryotes. We found 20 candidate selenocysteine-encoding genes in *O. lucimarinus*, all containing a putative SECIS element at their 3′ end; 19 are shared with *O. tauri*, and one is a recent duplication in *O. lucimarinus* only (SI Table 4). *O. tauri* has an additional selenocysteine-encoding candidate gene as discussed below. In contrast, *Chlamydomonas* is predicted to have 10 selenoproteins (16) despite having a 10 times larger genome size of ≈120 million base pairs (www.jgi.doe.gov/chlamy). One major category of the selenoproteins in *Ostreococcus* includes the glutathione peroxidases, for which five of six gene models are predicted selenoproteins. These results suggest possibly a functional tuning to the origin of the stress or subcellular compartment for each member of the glutathione peroxidase family (17). The greater catalytic efficiency of a selenocysteine-containing enzyme relative to a cysteine-containing homolog [e.g., recently reported 10- to 50-fold increase for a *Chlamydomonas* selenoprotein (18)] allows an organism to "save" on nutrient resources like nitrogen for protein production, particularly if the relevant activity is highly expressed.

Of particular interest to understanding the speciation of phytoplankton, *O. tauri* has a predicted gene for a selenoprotein (SelA) that is conserved in *O. lucimarinus*, but it is not a selenoprotein, the three selenocysteines being replaced by Cys (two) or Ser (one). This suggests that selenium availability may be acting as a force on the speciation of these and other phytoplankton, a hypothesis that has not been suggested previously.

Iron and other metals. Iron is also likely to affect phytoplankton diversity and speciation, because it has been demonstrated to be limiting in some ecosystems (19). In unicellular free-living eukaryotes a common system for iron acquisition has been proposed involving the coupled activity of a ferric reductase, multicopper oxidase, and a ferric permease (20–22). This system is found in marine diatoms and *Chlamydomonas*, a relative of *Ostreococcus* in the green algal lineage. *Ostreococcus* in stark contrast appears to lack all of these iron transport components, with the possible exception of a multicopper oxidase found only in *O. tauri*, as well as lacking any genes related to phytosiderophore uptake (23, 24). This implies that *Ostreococcus* has a novel system of Fe acquisition for a eukaryote that is mechanistically different from those of major competitors such as diatoms. Both strains of *Ostreococcus* have genes coding for proteins with significant sequence similarity to prokaryotic siderophore-iron uptake. Given the lack of any clear system of Fe acquisition in an organism isolated from an environment typified by low Fe concentrations, it is tempting to suggest that this organism may be able to acquire Fe-siderophore complexes. These complexes may be present in solution when bacteria in the same environment produce and export siderophores. We cannot rule out the possibility that *Ostreococcus* may be able to make its own siderophores. We found the biosynthesis pathway for catecholates in *O. lucimarinus* only, and these could be involved in siderophore biosynthesis.

Ostreococcus does appear to have genetic adaptations that reduce Fe requirements and allow Fe storage. *O. tauri* has a single copy of ferritin, and *O. lucimarinus* has a second copy that may be related to adaptations to continuous high light stress. Cytochrome c_6 (the iron-containing replacement of plastocyanin) is missing, and the use of plastocyanin as the sole electron carrier between the Cyt b_6/f complex and photosystem I, while reducing Fe quotas, imposes an absolute requirement for copper in this organism. Additionally, both genomes contain a copy of a small flavodoxin that may replace ferrodoxin in the photosynthetic electron transfer chain, further reducing iron requirements. Finally, both strains have genes for Cu/Zn- and Mn-containing superoxide dismutases, possibly a Ni-containing SOD, but not a Fe-SOD (25).

Copper concentrations have been shown to affect community composition in coastal ecosystems (26); therefore, it came as some surprise to find that *Ostreococcus* lacks a gene for phytochelatin synthase for ameliorating copper toxicity (27, 28). Instead, this organism contains tesmin-like metallothionein sequences and several Cu-efflux proteins. Arguably, the obligate use of Cu in photosynthesis (plastocyanin), respiration (cytochrome *c* oxidase), and oxidative defense (Cu/Zn SOD) may necessitate higher than typical Cu quotas in the organism.

Vitamins. The *Ostreococcus* genomes suggest that the organic and organometallic micronutrients thiamine and B_{12} must be acquired from the extracellular environment for growth. Unlike the *Chlamydomonas* genome, which encodes both B_{12}-dependent and -independent methionine synthases, the *Ostreococcus* genome con-

tains only the B_{12}-dependent form and hence has a strict dependence on B_{12}. Because the genome does not encode a B_{12} biosynthetic pathway, this implies that *Ostreococcus* acquires B_{12} or a precursor from seawater or associated bacteria (29).

The *Ostreococcus* genomes also lack a complete pathway for thiamine biosynthesis. In addition, thiamine pyrophosphate riboswitches, metabolite-sensing conserved RNA secondary structures, were found in UTRs of genes (30). Although mostly common to prokaryotes, a few riboswitches have been documented in eukaryotes. In the *O. tauri* and *O. lucimarinus* genomes these elements were found upstream of coding sequences with similarity to bacterial sodium:solute symporters. Although there is no indication for the specificity of a transporter located on Chr 4, PanF located on Chr 12 is clearly related to pantothenate transporters. The orthologous genes and thiamine pyrophosphate riboswitch were also found in a Sargasso Sea metagenomics data set, which is thought to contain *Ostreococcus* DNA (31). Altogether this strongly suggests that thiamine pyrophosphate regulates the expression of these two genes.

Evolution of the Genus *Ostreococcus*. The *Ostreococcus* genomes provide insights into evolutionary processes other than speciation including the evolution of a uniquely small cell size and the evolution of the green plant lineage that includes terrestrial plants.

Gene loss. In the evolution of its small size, *Ostreococcus* has lost a number of genes involved in flagellum biosynthesis and is missing cell wall proteins that are found in *Chlamydomonas*. Many characterized transcription factors in *Arabidopsis* are rare or absent in *O. tauri* and *O. lucimarinus* (e.g., ERF, MADS-box, basic helix–loop–helix, and NAM) (SI Table 5). Like in plants, the ERF and basic helix–loop–helix factors are common in *Chlamydomonas*, suggesting their loss in *Ostreococcus*. *Chlamydomonas* also has two plant-specific classes, AUX-IAA and SBP, that *Ostreococcus* does not have.

Peroxisomes have not been described in *Ostreococcus*, and we therefore expected to find the loss of peroxisome-specific genes. However, a comparison of the *Ostreococcus* proteomes with those of land plants, *Chlamydomonas*, and diatoms revealed the presence of sufficient peroxisomal proteins (PEX genes) needed to create a functioning peroxisome even in an organism of this small cell size. In some phytoplankton the size of the peroxisome greatly increases when the organism is grown on purines as a nitrogen source (32). The pathways for purine degradation that occur in the peroxisome were not found in *Ostreococcus*, which is consistent with selection for a small cell size.

Unique gene transfer to the nucleus. The *Ostreococcus* genome encodes heme-handling components like CcsA and Ccs1 and thiol-metabolizing components like CcdA (33). Interestingly, CcsA, which is encoded on the organelle genome in all other plant and algal genomes, is found in the nuclear genome in both *Ostreococcus* species. CcsA is a polytopic, hydrophobic protein that is the defining "core" component, presumably a heme-ligating molecule, of the system II cytochrome biogenesis pathway (34), and its occurrence in *Ostreococcus* nuclear genomes is the first example of the transfer of this gene from the organelle to the nucleus.

Gene fusions. Possibly because of evolutionary pressure toward a smaller cell and genome size where intergenic DNA and intron DNA would be spared, the *Ostreococcus* genomes show some unique examples of apparent fusion proteins. We have identified 330 and 348 potential gene fusions from *O. tauri* and *O. lucimarinus*, respectively, 137 of which were found in both species (SI Table 6). Although some may be chimeric gene predictions, 49 potential gene fusions have single-exon gene models and combine functions of two metabolic or redox enzymes. Some fusions involve important metabolic pathways such as pigment biosynthesis and nitrate reduction (SI Table 6).

Chromatin proteins. The most striking fact about the complement of chromatin proteins encoded by the *Ostreococcus* genome is that it lacks quite a few proteins found widely in plants, animals, and fungi. We searched the *Ostreococcus* genome for 104 chromatin proteins that existed in the most recent common ancestor of plants and animals (www.chromdb.org); 76 of these were found, but 28 were not. Similarly, budding yeasts (*Saccharomyces cerevisiae* and *Candida glabrata*) retained 70 of these proteins and dispensed with 34 of them. Eighteen of the 28 proteins not found in *Ostreococcus* were also not found in budding yeasts. However, both yeasts and *Ostreococcus* do possess a basic complement of all types of histone chaperones and histone-modifying enzymes. Ten chromatin-associated genes not found in *Ostreococcus* that are found in yeasts appear largely to be involved in the homologous recombination mode of double-strand break DNA repair.

Although *Ostreococcus* lacks both major eukaryotic DNA methyltransferase types (Dnmt1 and Dnmt3), it does possess two bacterial 5-cytosine DNA methyltransferases, both fused to a chromatin domain. Interestingly, *Ostreococcus* also possesses a DNA glycosylase that is a member of a clade of plant DNA glycosylases that mediate DNA demethylation via a DNA repair-like process. Thus, *Ostreococcus* may possess a unique DNA methylation/demethylation system whose function could be involved in defense against foreign DNA.

Conclusion

Comparative analysis of the genomes of two *Ostreococcus* species has revealed major differences in genome organization between them. While the core set of 18 chromosomes is conserved between the genomes, the remaining chromosomes (2, 18, 19, and 21) evolve in a number of different ways and may reflect ongoing adaptation and speciation processes. Small differences in proteomes such as the gain or loss of metal using genes not only illustrate the divergence of these two sister organisms but may be especially important in defining the ecological niche of each species. In addition, both *Ostreococcus* species employ similar mechanisms for optimization of genome and cell size, including gene loss, gene fusion, utilization of selenocysteine-containing proteins, chromatin reduction, and others. As genomes of other phytoplankton species become available, the relative importance of the processes outlined here in creating or maintaining phytoplankton diversity will become clearer.

Methods

Data and Strain Availability. Gene predictions, annotations, supporting evidence, and analyses are available through JGI Genome Portals on www.jgi.doe.gov/Olucimarinus and www.jgi.doe.gov/Otauri. *O. lucimarinus* genome sequence, predicted genes, and annotations were deposited in the GenBank database under accession numbers CP000581–CP000601 for Chr 1 through Chr 21. The *O. lucimarinus* strain (CCE9901) used here was isolated by B.P. from 32.9000 N 117.2550 W (Scripps Institution of Oceanography Pier, La Jolla, CA) and was grown as reported previously (7). This strain has been deposited in the Provasoli-Guillard Culture Collection of Marine Phytoplankton as CCMP2514.

Genome Sequencing and Finishing. Whole-genome shotgun sequencing was performed as in refs. 35 and 36. To perform finishing, initial read layouts from the *O. lucimarinus* whole-genome shotgun assembly were converted into our Phred/Phrap/Consed pipeline (37). After manual inspection of the assembled sequences, finishing was performed by resequencing plasmid subclones and by walking on plasmid subclones or fosmids using custom primers. All finishing reactions were performed with 4:1 BigDye to dGTP BigDye terminator chemistry (Applied Biosystems, Foster City, CA). Because of the high GC content of this genome, primer walks failed to resolve a large number of the gaps; these were resolved by generating pooled small insert shatter libraries from 3-kb plasmid clones. Repeats were resolved by transposon-hopping 8-kb plasmid clones. Fosmid clones were shotgun-sequenced and finished to fill

large gaps, resolve large repeats, or resolve chromosome duplications and extend into chromosome telomere regions. Finished chromosomes have no gaps, and the sequence has less than one error in 100,000 bp.

Pulsed-Field Gel Electrophoresis and Radiolabeled Hybridization. The two *Ostreococcus strains* (2–5 $\times$ 10^7 cells) were agarose-embedded and analyzed by pulsed-field gel electrophoresis as described previously (9, 38, 39). The sequences of the primers specifically designed from the two duplicated parts of the *O. lucimarinus* Chr 21 sequence were (*i*) 5′-AACGCGCGATTAAGTCGTAC-3′ and 5′-CATCCGTCAACTTGTCTTCG-3′ for Chr 9 duplication and (*ii*) 5′-TTCGCCGTTACTATCGGATC-3′ and 5′-GGAGGTCATAGCAACATCGT-3′ for Chr 13 duplication. Using these primers, DNA fragments of 600 and 820 bp, respectively, were amplified by standard PCR, purified, and radiolabeled with [α-^{32}P]dCTP by random priming (Prime-a-gene kit; Promega, Madison, WI).

Genome Annotation. Gene prediction methods used for annotation of two *Ostreococcus* genomes included *ab initio* Fgenesh (40), homology-based Fgenesh+ (SoftBerry), Genewise (41), MAGPIE (42), EST-based estExt (I.V.G., unpublished data), and a combined-approach EuGene (43). Predicted genes were annotated by using double-affine Smith-Waterman (TimeLogic) alignments against proteins from the National Center for Biotechnology Information nonredundant protein database, protein domain predictions using InterProScan (44), and their mappings to Gene Ontology (45), eukaryotic clusters of orthologous groups [KOGs (46)], and KEGG metabolic pathways (47). The available functional annotation of *O. tauri* (GenBank accession nos. CR954201–CR954220) was also used for annotation of the genome of *O. lucimarinus*.

All predicted models were combined into a nonredundant set of models, filtered models, in which the best model per locus was selected based on homology to other proteins and EST support. The predicted set of gene models has been validated by using available experimental data and computational analysis. Nineteen percent to 28% of genes in the final set are the same models produced by at least two different methods. Sixty-five percent to 73% of gene models are supported by conservation with the related *Ostreococcus* genome at the DNA level using VISTA analysis. Twenty-one percent to 28% of predicted genes are supported by ESTs mapped to corresponding genomes using BLAT (48). Seventy-nine percent to 84% of *Ostreococcus* genes have shown homology to a nonredundant set of proteins from National Center for Biotechnology Information and 92–93% to each other as detected by BLAST (49) ($e < 1e\text{-}8$). Less than 5% of the models are not supported by either of these lines of evidence. Predicted genes and their coordinates and functional assignments are also being manually curated by the community of annotators.

Whole-Genome Alignments. Chromosome-scale synteny between both *Ostreococcus* species was analyzed with i-ADHoRe, which identifies runs of collinear predicted proteins between genomic regions (50). We used gap size of 25 genes, a Q value of 0.9, and a minimum of three homologs to define a collinear block. In addition, we used the VISTA framework (12) with the constructed genome-wide pairwise alignments accessible from http://pipeline.lbl.gov.

Analysis of Codon Usage. For each chromosome of each species, frequencies for each of the 64 codons and GC frequency were calculated by using the genomic sequence for the all predicted protein coding regions on that chromosome as input to the "cusp" program from the EMBOSS 3.0 bioinformatics suite (51). Codon-frequency principal components, using correlations, were then calculated with each chromosome as a case and each codon frequency as a variable (52). Similarities between GC content and codon usage were evaluated by projecting each case onto the first and second principal components and then calculating the correlation between each principal component's projections and GC frequency.

We are grateful to J. Bristow of the Joint Genome Institute for critical reading of the manuscript. B.P. and I.P. were supported by Department of Energy Grant DE-FG03-O1ER63148 for transporter annotation. E.D., S.J., H.M., and G.P. were supported by the European network "Marine Genomics Europe" (GOCE-20040505403). This work was performed under the auspices of the U.S. Department of Energy's Office of Science, Biological and Environmental Research Program, and by the University of California, Lawrence Livermore National Laboratory under Contract W-7405-Eng-48, Lawrence Berkeley National Laboratory under Contract DE-AC02-05CH11231, Los Alamos National Laboratory under Contract DE-AC52-06NA25396, and Stanford University under Contract DEFC02-99ER62873.

1. Behrenfeld MJ, Falkowski PG (1997) *Limnol Oceanogr* 42:1–20.
2. Hutchinson GE (1961) *Am Nat* 95:137–145.
3. Courties C, Vaquer A, Trousselier M, Lautier J, Chretiennotdinet MJ, Neveux J, Machado C, Claustre H (1994) *Nature* 370:255–255.
4. Chretiennot-Dinet MJ, Courties C, Vaquer A, Neveux J, Claustre H, Lautier J, Machado MC (1995) *Phycologia* 34:285–292.
5. Diez B, Pedros-Alio C, Massana R (2001) *Appl Environ Microbiol* 67:2932–2941.
6. Guillou L, Eikrem W, Chretiennot-Dinet MJ, Le Gall F, Massana R, Romari K, Pedros-Alio C, Vaulot D (2004) *Protist* 155:193–214.
7. Worden AZ, Nolan JK, Palenik B (2004) *Limnol Oceanogr* 49:168–179.
8. Countway PD, Caron DA (2006) *Appl Environ Microbiol* 72:2496–2506.
9. Rodriguez F, Derelle E, Guillou L, Le Gall F, Vaulot D, Moreau H (2005) *Environ Microbiol* 7:853–859.
10. Derelle E, Ferraz C, Rombauts S, Rouze P, Worden AZ, Robbens S, Partensky F, Degroeve S, Echeynie S, Cooke R, *et al.* (2006) *Proc Natl Acad Sci USA* 103:11647–11652.
11. Gonza'lez J, Ranz JM, Ruiz A (2002) *Genetics* 161:1137–1154.
12. Frazer KA, Pachter L, Poliakov A, Rubin EM, Dubchak I (2004) *Nucleic Acids Res* 32:W273–W279.
13. Palenik B, Brahamsha B, Larimer FW, Land M, Hauser L, Chain P, Lamerdin J, Regala W, Allen EE, McCarren J, *et al.* (2003) *Nature* 424:1037–1042.
14. Koszul R, Caburet S, Dujon B, Fischer G (2004) *EMBO J* 23:234–243.
15. Peers G, Price NM (2006) *Nature* 441:341–344.
16. Novoselov SV, Rao M, Onoshko NV, Zhi H, Kryukov GV, Xiang Y, Weeks DP, Hatfield DL, Gladyshev VN (2002) *EMBO J* 21:3681–3693.
17. Gladyshev VN, Kryukov GV (2001) *Biofactors* 14:87–92.
18. Kim H-Y, Fomenko DE, Yoon Y-E, Gladyshev VN (2006) *Biochem Mol Biol Int* 45:13697–13704.
19. Martin JH, Coale KH, Johnson KS, Fitzwater SE, Gordon RM, Tanner SJ, Hunter CN, Elrod VA, Nowicki JL, Coley TL, *et al.* (1994) *Nature* 371:123–129.
20. Askwith CC, de Silva D, Kaplan J (1996) *Mol Microbiol* 20:27–34.
21. Armbrust EV, Berges JA, Bowler C, Green BR, Martinez D, Putnam NH, Zhou SG, Allen AE, Apt KE, Bechner M, *et al.* (2004) *Science* 306:79–86.
22. La Fontaine S, Quinn JM, Nakamoto SS, Page MD, Göhre V, Moseley JL, Kropat J, Merchant S (2002) *Eukaryot Cell* 1:736–757.
23. Curie C, Briat J-F (2003) *Annu Rev Plant Biol* 54:183–206.
24. Kosman DJ (2003) *Mol Microbiol* 47:1185–1197.
25. Kliebenstein DJ, Monde RA, Last RL (1998) *Plant Physiol* 118:637–650.
26. Moffett JW, Brand LE, Croot PL, Barbeau KA (1997) *Limnol Oceanogr* 42:789–799.
27. Ahner BA, Kong S, Morel FMM (1995) *Limnol Oceanogr* 40:649–657.
28. Cobbett CS (1999) *Trends Plants Sci* 4:335–337.
29. Croft MT, Lawrence AD, Raux-Deery E, Warren MJ, Smith AG (2005) *Nature* 438:90–93.
30. Mandal M, Breaker RR (2004) *Nat Rev Mol Cell Biol* 5:451–463.
31. Venter JC, Remington K, Heidelberg JF, Halpern AL, Rusch D, Eisen JA, Wu DY, Paulsen I, Nelson KE, Nelson WC, *et al.* (2004) *Science* 304:66–74.
32. Oliveira L, Huynh H (1990) *Can J Fish Aquat Sci* 47:351–356.
33. Kranz R, Lill R, Goldman B, Bonnard G, Merchant S (1998) *Mol Microbiol* 29:383–396.
34. Hamel PP, Dreyfuss BW, Xie Z, Gabilly ST, Merchant S (2003) *J Biol Chem* 278:2593–2603.
35. Myers EW (1999) in *Proceedings of the Seventh International Conference on Intelligent Systems for Molecular Biology*, eds Lengauer T, Schneider R, Bork P, Brutlad D, Glasgow J, Mewes H-W, Zimmer R (AAAI Press, Menlo Park, CA), pp 202–210.
36. Aparicio S, Chapman J, Stupka E, Putnam N, Chia J, Dehal P, Christoffels A, Rash S, Hoon S, Smit A, *et al.* (2002) *Science* 297:1301–1310.
37. Gordon D, Abaijian C, Green P (1988) *Genome Res* 8:195–202.
38. Mead JR, Arrowood MJ, Current WL, Sterling CR (1988) *J Parasitol* 74:366–369.
39. Wöhl T, Brecht M, Lottspcich F, Ammer H (1995) *Electrophoresis* 16:739–741.
40. Salamov AA, Solovyev VV (2000) *Genome Res* 10:516–522.
41. Birney E, Clamp M, Durbin R (2004) *Genome Res* 14:988–995.
42. Gaasterland T, Sensen CW (1996) *Biochimie* 78:302–310.
43. Schiex T, Moisan A, Rouze P (2001) *Lect Notes Comput Sci* 2066:111–125.
44. Mulder NJ, Apweiler R, Attwood TK, Bairoch A, Bateman A, Binns D, Bradley P, Bork P, Bucher P, Cerutti L, *et al.* (2005) *Nucleic Acids Res* 33:D201–D205.
45. Gene Ontology Consortium (2001) *Genome Res* 11:1425–1433.
46. Koonin EV, Fedorova ND, Jackson JD, Jacobs AR, Krylov DM, Makarova KS, Mazumder R, Mekhedov SL, Nikolskaya AN, Rao BS, *et al.* (2004) *Genome Biol* 5:R7.
47. Kanehisa M, Goto S, Kawashima S, Okuno Y, Hattori M (2004) *Nucleic Acids Res* 32:D277–D280.
48. Kent WJ (2002) *Genome Res* 12:656–664.
49. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) *J Mol Biol* 215:403–410.
50. Simillion C, Vandepoele K, Saeys Y, Van de Peer Y (2004) *Genome Res* 14:1095–1106.
51. Rice P, Longden I, Bleasby A (2000) *Trends Genet* 16:276–277.
52. Venables WN, Ripley BD (2002) *Modern Applied Statistics with S* (Springer, New York).