Clues about the Genetic Basis of Adaptation Emerge from Comparing the Proteomes of Two *Ostreococcus* Ecotypes (Chlorophyta, Prasinophyceae)

Séverine Jancek,† Sébastien Gourbière,‡ Hervé Moreau,*† and Gwenaël Piganeau*†

*UPMC University of Paris 06, UMR 7628, MBCE, Observatoire Océanologique, Banyuls/mer, France; †CNRS, UMR 7628, MBCE, Observatoire Océanologique, Banyuls/mer, France; and ‡Université de Perpignan Via Domitia, Laboratoire de Mathématiques, Physiques et Systèmes, EA 4217, Perpignan, France

We compared the proteomes of two picoplanktonic *Ostreococcus* unicellular green algal ecotypes to analyze the genetic basis of their adaptation with their ecological niches. We first investigated the function of the species-specific genes using Gene Ontology databases and similarity searches. Although most species-specific genes had no known function, we identified several species-specific functions involved in various cellular processes, which could be critical for environmental adaptations. Additionally, we investigated the rate of evolution of orthologous genes and its distribution across chromosomes. We show that faster evolving genes encode significantly more membrane or excreted proteins, consistent with the notion that selection acts on cell surface modifications that is driven by selection for resistance to viruses and grazers, keystone actors of phytoplankton evolution. The relationship between GC content and chromosome length also suggests that both strains have experienced recombination since their divergence and that lack of recombination on the two outlier chromosomes could explain part of their peculiar genomic features, including higher rates of evolution.

Introduction

Unicellular photosynthetic organisms are responsible for most of the primary biomass production in oceans. Their diversity is amazingly high, including organisms belonging to most lineages of the tree of life. For practical reasons, they are referred to by their size, from microplankton (10–100 μm) to picoplankton species (below 2–3 μm). Picoplankton is constituted both by prokaryotic and eukaryotic cells, which can be either heterotroph or autotroph. Although picoeukaryotes are a minor component of picoplankton in terms of cell number, the photosynthetic species of these organisms are known to play a significant role in primary productivity in oligotrophic areas, where they represent up to 80% of the autotrophic biomass (Li 1994). They may account for 75% of net carbon assimilation in some coastal areas (Worden et al. 2004). Although some work on picoeukaryotes has been done, the knowledge about its diversity remains far behind our understanding of prokaryotic diversity (Moreira and Lopez-Garcia 2002; Vaulot et al. 2002; Guillou et al. 2004; Piganeau et al. 2008). Some quantitative studies based on in situ hybridization experiments show that among these groups, Prasinophytes apparently dominate picoeukaryotes in different oceanic areas (Not et al. 2004).

Prasinophytes are primary endosymbionts that probably diverged very early from the ancestor of all chloroplast-containing green plants and algae. Discovered in 1994 (Courties et al. 1994), *Ostreococcus* is the smallest (diameter 0.9–1.0 μm) such free-living eukaryotic organism described to date. It has a minimal cellular organization (one chloroplast and one mitochondrion), a small genome (between 12 and 15 Mb) (Derelle et al. 2002) and is widespread, having been found in coastal and oligotrophic North Atlantic waters, in the Mediterranean, Indian, and Pacific Oceans (Worden et al. 2004; Zhu et al. 2005; Countway and Caron 2006; Piganeau and Moreau 2007). Different *Ostreococcus* ecotypes from surface or deeper layers of waters provide evidence of niche adaptation (Rodriguez et al. 2005), similar to the ecotypic differentiation and consequent adaptative success illustrated by *Prochlorococcus*, the most abundant marine prokaryotic picophytoplankter (Moore et al. 1998). *Ostreococcus tauri* and *Ostreococcus lucimarinus* are two surface strains, *O. lucimarinus* being a high-light-adapted species found in the Pacific ocean and *O. tauri* being a high-light-adapted and low-light-adapted species found in the Mediterranean lagoons.

The recent availability of two complete *Ostreococcus* genome sequences (Derelle et al. 2006; Palenik et al. 2007) of these two species opens new approaches for finding biological functions that may be important for niche adaptation.

The comparison of the genomes of these two ecotypes already provided insight about how they achieve such a small cell size and unraveled multiple mechanisms implied in the divergence of these two species (Palenik et al. 2007). Strikingly, two chromosomes in both species show both lower levels of between-strain synteny and different base composition and gene densities to the other chromosomes. Furthermore, most genes on these chromosomes are species specific, and some of them are good candidates for recent horizontal gene transfer from bacteria into *Ostreococcus* (Palenik et al. 2007). These chromosomes could therefore be involved in speciation by maintaining the strains in genetic isolation from their relatives (Palenik et al. 2007).

In this study, we analyzed the features of the species-specific genes in both strains and the mode and tempo of evolution of their orthologous genes to investigate further the genetic basis of the adaptation of these two ecotypes to their ecological niches.

Methods

Data

Gene predictions, KOG (eukaryotic cluster of orthologous groups), KEGG (Kyoto Encyclopedia of Genes and Genomes), and GO (Gene Ontology) databases were downloaded from JGI (Joint Genome Institute) Genome Portals at http://www.jgi.doe.gov/Olucimarinus for *O. lucimarinus* and http://www.jgi.doe.gov/Otauri for *O. tauri*. The *O. tauri* (Ot) strain was isolated in the Thau lagoon (France)

Key words: picoeukaryotes, genome comparison, GC content, selection, adaptation.

E-mail: gwenael.piganeau@obs-banyuls.fr.

Mol. Biol. Evol. 25(11):2293–2300. 2008
doi:10.1093/molbev/msn168
Advance Access publication August 4, 2008

from 43°24′N 3°36′E (Chretiennot-Dinet et al. 1995). The *O. lucimarinus* (Ol) strain was isolated by B. Palenik from 32.9000N 117.2550W (Scripps Institution of Oceanography Pier, La Jolla, CA).

Search for Species-Specific Functions

GO (Ashburner et al. 2000; Gene Ontology Consortium 2001), KOG (Tatusov et al. 2003), and KEGG (Kanehisa et al. 2006) databases were used to find species-specific functions. We screened automatically each database for accession numbers present in only one of the two strains. We then checked that each identified species-specific function did not arise from annotation errors by searching for homologs of the genes corresponding to that function in the genome of the other species with tblastn (Altschul et al. 1990). We used a very stringent criterion for our search for species-specific functionbecause we dismissed all orthologous genes. Given the average divergence between orthologs of the two species (70.5%), some orthologous genes may have evolved into different functions. Alternatively, this criterion may also lead to some false-positive species-specific functions, because cases of nonhomologous genes sharing the same function, as a consequence of convergent evolution, are also known. However, given the high percentage of orthologous genes between the two genomes (79–82%), we think that convergent evolution in species-specific genes is an unlikely scenario.

Species-Specific Genes, Orthologs, and Duplications

To assess distribution of orthologous genes between chromosomes, we corrected the total number of genes per chromosome by the number of genes present in two identical copies due to recent segmental duplication on chromosomes 14, 18, and 21 in *O. lucimarinus* (247 genes). We also corrected the total number of genes per chromosome by removing 20 exact duplications scattered on nine chromosomes in *O. tauri*. We used these corrected gene numbers to estimate the percentage of specific genes and the repartition of orthologs between chromosomes, that is, we did not consider the products of species-specific duplication as species-specific genes. We considered that a gene had a nearly exact duplicate in a genome when the average nucleotide identity between the two genes was over 97% for 98% of their length. When a chromosome contained massive nearly exact duplicates from another chromosome, we removed the genes on the smaller chromosome (e.g., chromosome 21 for Ol). Otherwise, we randomly excluded one of the two copies.

Genes on chromosome 2 presenting a heterogeneous structure regarding GC content and intron structure in both strains were split into low GC and high GC regions—*O. tauri*: Chr2A (low GC) and Chr2B (high GC) and *O. lucimarinus*: Chr2a (high GC), Chr2b (low GC), and Chr2c (high GC).

Estimation of Substitution Rates

We assessed orthologous pairs of genes by using reciprocal blastp hits (RBH) with an *e* value threshold of 0.01 between the predicted protein sequences of each strain. This is a more stringent criterion than previously used (Palenik et al. 2007) and we thus retrieved 6,270 pairs of orthologs. Each pair was aligned with ClustalW 1.7 (Thompson et al. 1994) with default parameters. The average amino acid identity between orthologs is 70.5%. Pairwise estimates of the synonymous (dS) and nonsynonymous (dN) substitution rates were obtained from PAML 3.15 (Yang 1997) (runmode −2) with default parameters and with the codon frequency model F3×4 that assumes that the equilibrium codon frequencies are calculated from the average nucleotide frequencies at the three codon positions. We performed additional PAML analysis (Yang 2006) using the likelihood ratio test to discriminate between a model considering one dN/dS ratio, M0, and a model considering three types of sites: neutrally evolving, under positive selection, and under purifying selection M2. The dN/dS ratio estimates under the M0 and M2 models were highly correlated, and we thus used dN/dS estimated with the simplest model of protein evolution M0. To reduce estimation biases, the dN/dS ratio was calculated for sequences longer than 300 bp, and only sequences with values of $d\text{S} < 2$ and/or $d\text{N} < 5$ were kept for the analysis, leading to a further reduction of the data set to 1,305 orthologs.

Peptide Signal, Transmembrane Region Prediction, and Protein Localization

We used the Neural Network (Bendtsen et al. 2004) of SignalP 3.0 (http://www.cbs.dtu.dk/services/SignalP/) (Nielsen et al. 1997; Bendtsen et al. 2004) and TMHMM 2.0 (http://www.cbs.dtu.dk/services/TMHMM-2.0/) (Sonnhammer et al. 1998; Krogh et al. 2001) to identify putative excreted proteins and membrane proteins.

Wolf PSORT 0.2 (http://wolfpsort.seq.cbrc.jp/) (Nakai and Horton 1999) was used to predict protein subcellular localization. We considered that the localization was correctly inferred when the first localization's score was equal or superior to seven and the second localization, if present, was at least inferior to half the first score.

Statistical Analysis

We used the linear regression model to test the relationships between GC content, chromosome length, and substitution rates. Because the substitution rates (dN, dS, and dN/dS) are not normally distributed, we used nonparametric analysis of variance to check substitution rate heterogeneity between chromosomes.

All the statistical analysis was performed with R software (http://www.R-project.org).

Results

Identification of Species-Specific Genes and Functions

We identified 6,270 pairs of orthologs (reciprocal best hit, RBH) between the two *Ostreococcus* genomes. After removing duplicated genes among nonorthologs, we found 1,340 (17% of total genes) and 1,134 (15% of total genes)

Table 1
Genomic Features of *Ostreococcus tauri* and *Ostreococcus lucimarinus*

	Ostreococcus tauri	*Ostreococcus lucimarinus*
Genome size, Mbp	12.6	13.2
Chromosomes	20	21
Number of genes	7,725	7,651
Number of duplicated genes	20	247
Orthologous genes (% of total)	6,270 (81%)	6,270 (82%)
KOG characterized genes (%)	3393 (54%)	3435 (55%)
KOG poorly characterized genes (%)	933 (15%)	937 (15%)
KEGG indexed genes (%)	21 (<2%)	35 (3%)
GO indexed genes (%)	3,359 (53%)	3,419 (54%)
Strain-specific genes (% of total)	1,340 (17%)[a]	1,134 (15%)
No hit against GenBank (%)	77	58
KOG characterized genes (%)	178 (13%)	364 (32%)
KOG poorly characterized genes (%)	52 (4%)	146 (13%)
KEGG indexed genes (%)	21 (<2%)	35 (3%)
GO indexed genes (%)	225 (17%)	360 (32%)

[a] We excluded the 95 genes not assigned to any chromosome.

specific genes in *O. tauri* (Ot) and *O. lucimarinus* (Ol), respectively (table 1).

Comparisons with the total percent of genes assigned in the KOG database for seven other sequenced organisms (*Homo sapiens*, *Drosophila melanogaster*, *Caenorhabditis elegans*, *Saccharomyces cerevisiae*, *Arabidopsis thaliana*, *Cyanidioschyzon merolae*, and *Thalassiosira pseudonana*) revealed that the coverage of assigned functions for *O. tauri* and *O. lucimarinus* is among the highest, with 61% and 66%, respectively, a little less than in the yeast *S. cerevisiae* (69%). The percentage of genes assigned in the classes "unknown functions" and "uncharacterized functions" is also lower in *O. tauri* and *O. lucimarinus*, with 13% and 15%, respectively, than in the red algae *C. merolae* (17%) and the diatom *T. pseudonana* (19%) (Armbrust et al. 2004). A fraction of these species-specific genes could be indexed in KOG, KEGG, or GO databases, ranging from 32% to 13% using the KOG database to less than 3% for *O. lucimarinus* and to 2% for *O. tauri* genes using the GO database (table 1). When functional classes represented by at least five functions in the largest database (KOG) are taken into account, cell functions concerned by these species-specific genes in *O. lucimarinus* were mainly found in extracellular structure (33%), secondary metabolites biosynthesis (13%), carbohydrate metabolism (10%), and defense mechanisms (9%). For *O. tauri*, cell functions of species-specific genes were found in secondary metabolites biosynthesis (12%), defense mechanisms (10%), amino acid transport and metabolism (7%), carbohydrate metabolism (7%), and cell wall biogenesis (7%) (supplementary tables A1 and A2, Supplementary Material online). Some functions were specific to one strain because the reciprocal function in the other strain was missing in the database, despite the presence of well-conserved orthologs for the genes concerned. In such cases, the function was not considered as species specific. We estimated that this kind of mistake occurs for 3% of all assigned functions of the KOG database.

However, most of these species-specific genes have unknown functions (87% [Ot] to 68% [Ol] and 77% [Ot] to 58% [Ol] have no hit against GenBank [table 1]) so that we have no clue about their role in the adaptation process of these two species. This is much higher than the percentage of no hits observed for the total number of genes in both strains (about 8% in both strains) (Derelle et al. 2006; Palenik et al. 2007) and merely reflects that most unknown genes are species specific.

This high percentage of unknown genes might arise because of the absence of complete genomes of close relatives of *Ostreococcus* in GenBank or a higher rate of evolution of the species-specific genes. Species-specific genes also contain fewer green lineage-specific genes. Indeed, the global annotation of both genomes showed that around 40% of the genes had a green lineage origin (Viridiplantae), whereas only 20% of *O. lucimarinus* and 8% of *O. tauri*-specific genes gave a significant Blast score with Viridiplantae genes, a similar percentage to that seen with genes of bacterial or metazoan origin (fig. 1).

A potentially important cell pathway for the adaptation of phytoplankton to environment is the iron metabolism (Strzepek and Harrison 2004). Interestingly, we confirmed and extended the specificity of iron uptake in *O. tauri* in this analysis. Indeed, we confirmed the presence of a multicopper oxidase in *O. tauri* already described in the original genome comparison study (Palenik et al. 2007) and also found two ferric reductase genes not previously described in *O. tauri*, which are absent in *O. lucimarinus*. No expressed sequence tags (ESTs) have been obtained for the multicopper oxidase, whereas an EST is present for at least one of the two *O. tauri*-specific ferric reductases. These genes are known to be involved in the iron uptake in diatoms and *Chlamydomonas*, usually in combination with iron transporters. No clear iron transporter could be identified in either of the *Ostreococcus* strains, although some genes showed a loose similarity.

Analysis of Substitution Rates

The two *Ostreococcus* are very divergent, with 70% average amino acid identity between orthologs, making them the most divergent species within the same genus among sequenced eukaryotes (Palenik et al. 2007). Because the synonymous substitution rate is saturated in 79% of the orthologs, we restricted our analysis to the remaining 1,305 genes, where average dN/dS ratio was 0.07, consistent with the notion that purifying selection acts on most nonsynonymous mutations in these organisms (fig. 2).

The distribution of orthologs is not homogenous between chromosomes (table 2, chi-square test, Ot: degrees of freedom [df] = 20, P value $< 10^{-16}$; Ol: df = 22, P value $< 10^{-16}$). When chromosomes containing extremely low proportions of orthologs and presenting an unusual structure are excluded (outlier chromosomes Ch18 and Ch2b for *O. lucimarinus*; Ch19 and Chr2A for *O. tauri*), the distribution of orthologs per chromosome remains significantly different (chi-square, Ot: P value = 0.021; Ol: P value = 0.0003). Consistent with this trend, the rate of nonsynonymous substitution, dN, is significantly different between chromosomes (Kruskal–Wallis, Ot: 47.2, df = 20, P value = 0.0005, n = 6,217; Ol: 48.7, df = 22, P value = 0.0009, n = 6,242). Faster evolving genes also show a heterogeneous distribution, and most of them are

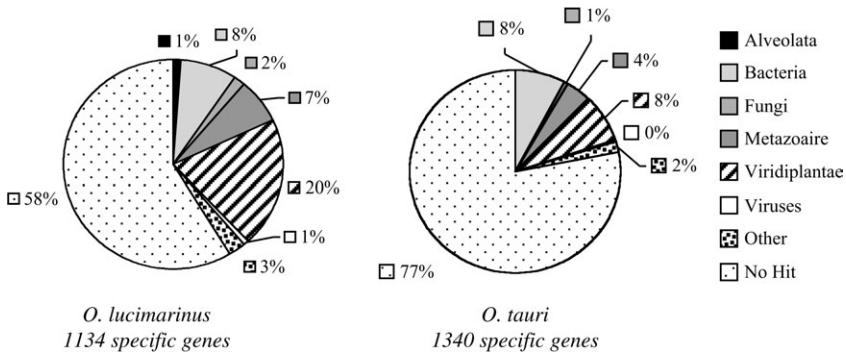

FIG. 1.—Taxonomic distribution of species-specific genes for both species.

on chromosomes 18 (*O. lucimarinus*) and 19 (*O. tauri*). There is still a significant difference between chromosomes when these two fast-evolving chromosomes are excluded from the analysis (Kruskal–Wallis, P value < 0.05 for both Ot and Ol). Although no significant difference in dS between chromosomes could be observed (reducing the number of genes to $d\text{S} < 2$; Kruskal–Wallis, P value > 0.4 for both Ot, $n = 1{,}314$, and Ol, $n = 1{,}315$), a marginally significant difference in the dN/dS ratio was found in *O. lucimarinus* (P value 50.055), where chromosomes 18 and 02b have a higher dN/dS ratio.

The percentage of species-specific genes does not correlate with the average rate of protein evolution per chromosome if the outlier chromosomes are excluded. However, the greater abundance of species-specific genes on these outlier chromosomes could be seen either as a consequence of faster sequence evolution on these chromosomes due to a higher mutation rate and/or as relaxed functional constraints.

Because faster evolving proteins are likely candidates for adaptation (Yang and Bielawski 2000), we investigated the function of 50 fastest evolving genes, as measured by dN/dS. In all, 50% of the fastest evolving genes had no hit against GenBank and 26% of them had unknown functions. In contrast, the 50 most highly constrained genes had less

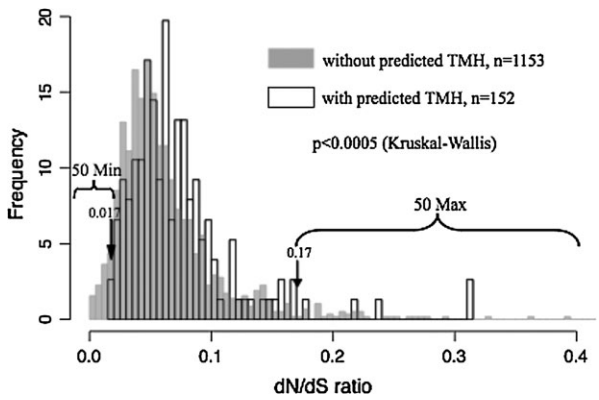

FIG. 2.—Distribution of dN/dS ratio of 1,305 RBH between *Ostreococcus tauri* and *Ostreococcus lucimarinus* ($d\text{S} < 2$ and length >300 bp) among proteins with and without predicted transmembrane helix domain (TMH).

than 6% genes with unknown function and no "no hit" (table 4). Most of these genes are housekeeping genes involved in basal metabolism or in chromatin structure and genome dynamics. In contrast, no clear cell pathway could be identified among the fastest evolving genes having a significant hit, except for certain genes involved in metal metabolism, such as a metal ion binding or a zinc ribbon protein and two urease accessory proteins (table 3). These two urease accessory proteins act as urease-specific chaperones by incorporating Nickel into the urease protein and are required for assembling an active urease (Sirko and Brodzik 2000). These urease accessory proteins interact sterically to form a stable complex (Witte et al. 2005), consistent with their concomitantly high dN/dS ratio.

TMHMM and other softwares (see Methods) predicted that 14% (7/50) of faster evolving proteins have one or several transmembrane domains that are either membrane localized or excreted proteins (table 3). This is significantly more than the 2% of transmembrane proteins observed in slow evolving proteins (Fisher's exact test, P value $= 0.03$) (table 4). Thus, proteins predicted to have a transmembrane domain evolve faster than other proteins (fig. 2) (Kruskal–Wallis, df $= 2$, P value $= 0.00002$ for dN and P value $= 0.0007$ for dN/dS), what is consistent with recent findings in yeast (Julenius and Pedersen 2006). TMHMM searches of species-specific genes predicted that 7% (Ot) and 11% (Ol) of specific genes have one or several transmembrane helices. This is significantly less than that found for the orthologs, with around 14% in both strains (Fisher's exact test, Ot: P value $= 10^{-14}$; Ol: P value $= 0.001$). However, specific genes are on average 150 bp shorter than orthologous genes (Student's t-test $P < 10^{-5}$), and the fraction of proteins with predicted transmembrane domains increases with gene length. To correct for this bias, we split the data into gene length classes: specific genes show a lower proportion of transmembrane proteins for genes shorter than 1,200 bp and a higher proportion of transmembrane proteins for genes longer than 1,200 bp, but none of the differences between species-specific and orthologous genes were significant (Fisher's exact test, Ot: short genes, P value $= 0.91$, long genes, P value $= 0.75$; Ol: short genes, P value $= 0.67$, long genes, P value $= 0.12$).

We then used the PSORT software to determine subcellular localization in faster evolving and highly

Table 2
Distribution of Orthologs per Chromosomes in Both Species

Ostreococcus tauri				*Ostreococcus lucimarinus*			
Chromosomes	Total genes	nb RBH	% RBH	Chromosomes	Total genes	nb RBH	% RBH
Chr 01	660	584	88.5	Chr 01	684	597	87.3
Chr 02A	297	145	**48.8**	Chr 02b	173	125	**72.3**
Chr 02B	287	232	80.8	Chr 02a	186	165	88.7
				Chr 02c	120	103	85.8
Chr 03	577	506	87.7	Chr 03	589	498	84.6
Chr 04	554	460	83.0	Chr 04	525	466	88.8
Chr 05	504	429	85.1	Chr 05	494	423	85.6
Chr 06	468	403	86.1	Chr 06	456	406	89.0
Chr 07	456	391	85.7	Chr 07	463	398	86.0
Chr 08	404	341	84.4	Chr 08	424	351	82.8
Chr 09	438	362	82.6	Chr 09	415	344	82.9
Chr 10	363	310	85.4	Chr 10	358	310	86.6
Chr 11	339	288	85.0	Chr 11	328	292	89.0
Chr 12	315	269	85.4	Chr 12	318	267	84.0
Chr 13	308	270	87.7	Chr 13	300	265	88.3
Chr 14	336	283	84.2	Chr 20	332	286	86.1
Chr 15	298	249	83.6	Chr 14	327	246	**75.2**
Chr 16	277	233	84.1	Chr 15	266	226	85.0
Chr 17	269	226	84.0	Chr 16	263	229	87.1
Chr 18	204	170	83.3	Chr 17	209	173	82.8
Chr 19	128	15	**11.7**	Chr 18	82	19	**23.2**
Chr 20	103	79	**76.7**	Chr 19	92	81	88.0
				Chr 21[a]	182	109	59.9

NOTE.—Parts of chromosomes 2 in both strains have been defined from their GC content. Chr_02A (*O. tauri*) and Chr_02b (*O. lucimarinus*): low GC content region of Chromosome 2 and Chr_02B (*O. tauri*), Chr_02a, and Chr_02c (*O. lucimarinus*): high/normal GC content.

[a] Excluded from statistical analysis (see Methods).

constrained genes. Interestingly, 35% of fast-evolving proteins are potentially targeted to the chloroplast, in contrast to the most highly constrained genes, of which only 5% are targeted to this organelle. The main sublocalizations of constrained proteins are the cytoplasm (39%) and the nucleus (20%) and are consistent with the main functions found among these genes.

Base Composition Variation as a Function of Chromosome Length

In many species including mice, rats, and humans, recombination rates vary between chromosomes, and there is a strong negative relationship between chromosome size and chromosome recombination rate: large chromosomes

Table 3
Features of 51 Fastest Evolving Genes in Both Species (accession numbers of corresponding genes are provided as supplementary table A3, Supplementary Material online)

			TMHMM		PSORT	
Category	Nb	*d*N/*d*S value	*Ostreococcus tauri*	*Ostreococcus lucimarinus*	*Ostreococcus tauri*	*Ostreococcus lucimarinus*
APC10	1	0.2	No	No	cyto	cyto
Metal ion binding	1	0.31	Yes (1)	Yes (1)[a]	chlo	extr
Oxydoreductase	1	0.25	No	No	n.s.	mito
Oxysterol	1	0.18	No	No	nucl	cyto
Phosphate starvation	1	0.21	No	No	nucl	n.s.
Photosynthesis	3					
PSI subunit N (PsaN)		0.29	No	No	chlo	chlo
Psb3 of PSII		0.17	No	No	chlo	chlo
Ferredoxin PetF		0.22	No	No	chlo	chlo
Protease	1	0.47	No	No	nucl	nucl
Transamidase	1	0.24	No	Yes (1)[a]	n.s.	plas
Urease	2	00.19 (0.16–0.22)[b]	No	No	cyto	cyto
Zinc ribbon	1	0.19	No	No	chlo	chlo
Unknown	13	0.27 (0.17–0.59)[b]	Yes (1)[a]	Yes (2)[a]	chlo (5), n.s. (5), plas (1), cyto (1), nucl (1)	chlo (3), n.s. (3), cyto (3), plas (2), nucl (1), mito (1)
No hit	25	0.26 (0.17–0.72)[b]	Yes (5)[a]	Yes (3)[a]	chlo (8), n.s. (8), nucl (4), cyto (2), plas (2), mito (1)	chlo (10), n.s. (9), nucl (5)

NOTE.—n.s., not significant; chlo, chloroplast; nucl, nucleus; cyto, cytoplasm; plas, plasmic membrane; mito, mitochondria; and extr, extracellular.

[a] Peptide signal predicted with SignalP.

[b] Average (minimum–maximum).

Table 4
Features of Top 52 Highly Constrained Genes in Both Species (accession numbers of corresponding genes are provided as supplementary table A4, Supplementary Material online)

Category	Nb	*d*N/*d*S value	TMHMM	PSORT	
				Ostreococcus tauri	*Ostreococcus lucimarinus*
ABC transporter	1	0.017	No	cyto	cyto
APC	1	0.015	No	nucl	nucl
Calmodulin	1	0.001	No	n.s.	n.s.
Cytoskeleton	5	0.010 (0.006–0.015)[a]	No	n.s. (4), cysk (1)	n.s. (4), cysk (1)
Dehydrogenase	1	0.016	No	cyto	cyto
Ethylene	1	0.01	No	cyto	cyto
Geranylgeranyl reductase	1	0.004	No	cyto	cyto
GTP/ATP/UDP binding	6	0.010 (0.005–0.017)[a]	No	n.s. (2), cysk (1), cyto (1), nucl (1), chlo (1)	cyto (3), n.s. (2), cysk (1)
Heat shock protein/GF14	2	0.015; 0.017	No	cyto (1), n.s. (1)	cyto (1), n.s. (1)
Helicase	4	0.012 (0.008–0.016)[a]	No	cyto (2), nucl (1), n.s. (1)	cyto (2), nucl (1), cysk (1), n.s. (1)
Histone	5	0.005 (0.001–0.011)[a]	No	nucl (5)	nucl (5)
Phosphatase	1	0.0139	No	cyto	cyto
Proteasome	3	0.011 (0.009–0.014)[a]	No	cyto (2), n.s. (1)	cyto (1), nucl (1), n.s. (1)
Ribosomal protein	4	0.011 (0.008–0.016)[a]	No	cyto (3), n.s. (1)	cyto (4)
RNA/DNA binding	7	0.013 (0.001–0.015)[a]	No	cyto (4), nucl (1), chlo (1), n.s. (1)	cyto (3), nucl (2), chlo (1), n.s. (1)
Thioredoxin	1	0.009	No	n.s.	n.s.
Translocase	1	0.017	Yes	n.s.	plas
Ubiquitin	4	0.011 (0.004–0.015)[a]	No	cyto (1), chlo (1), nucl (1), n.s. (1)	cyto (1), chlo (1), nucl (1), n.s. (1)
Unknown	3	0.010 (0.001–0.015)[a]	No	cyto (1), n.s. (2)	cyto (2), n.s. (1)

NOTE.—n.s., not significant; chlo, chloroplast; nucl, nucleus; cyto, cytoplasm; and plas, plasmic membrane.
[a] Average (minimum–maximum).

have low recombination rates and short chromosomes have high recombination rates (Jensen-Seaman et al. 2004). This is explained by the requirement for meiosis of at least one chiasma per chromosome and results in a higher chiasmata density and a longer map length per kilobase on shorter chromosomes. Evidence suggests that GC-biased mismatch repair exists in numerous organisms spanning six kingdoms (Birdsell 2002). A significant positive correlation between recombination and GC content is found in many of these organisms (Meunier and Duret 2004), suggesting that the processes influencing the evolution of GC content may be a general phenomenon. Nonrecombining regions of the genome and nonrecombining genomes would not be subject to this type of molecular drive (Birdsell 2002). Consistent with this scenario, GC content is negatively correlated with chromosome length in the yeast *S. cerevisiae* (Bradnam et al. 1999).

We observed a strong negative relationship between crude Chromosomal GC content and chromosome length in both genomes of *Ostreococcus*, when the outlier chromosomes are excluded (fig. 3, Ot: $R^2 = 0.64$; Ol: $R^2 = 0.54$). This result suggests that *Ostreococcus* species have experienced recombination since their divergence.

Discussion

Most marine picoeukaryotes are not yet cultivable in the laboratory (Moreira and Lopez-Garcia 2002) so that in silico approaches applied to whole or partial genome data (e.g., extracted from large scale metagenomes) provide the only clues about the lifestyle and evolution of these microorganisms. We investigated species-specific gene content and molecular rate of evolution by comparing two *Ostreococcus* ecotypes to unravel the genetic basis of their adaptation. Because the divergence between these two strains is very high, greater than that seen in the *Saccharomyces sensu stricto* genus (Liti and Louis 2005), many species-specific genes were identified. Despite a high proportion of species-specific genes with unknown functions, over 20% of these

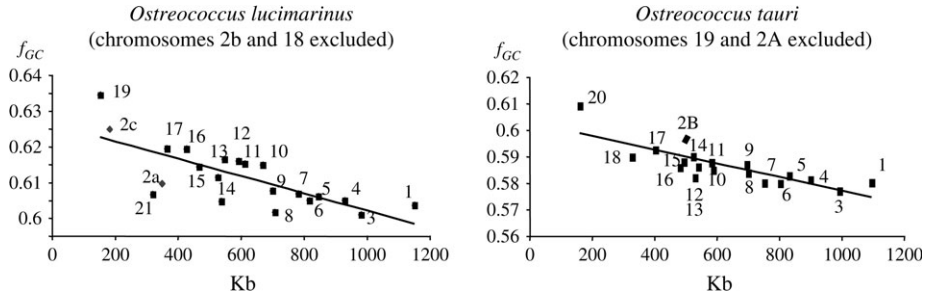

FIG. 3.—Chromosomal GC content, f_{GC}, versus chromosome size (kbp) in both genomes. Squares: chromosomes' numbers are given for both genomes, Diamonds: regions of outlier chromosomes 2 with high GC content. Linear regression lines are shown Olu ($R^2 = 0.54$, $P = 8 \times 10^{-5}$) and Ota ($R^2 = 0.64$, $P = 2 \times 10^{-5}$).

genes could be indexed in ontology databases. Functional analysis of these species-specific genes reveals various cellular functions, although it remains difficult to establish links between these specific functions and adaptation of the two *Ostreococcus* strains to their environment. Among these specific functions, the presence of iron metabolism genes found only in *O. tauri* (also see Palenik et al. 2007) potentially reflects a specific adaptation of this strain to the Thau lagoon where it was isolated. Indeed, this lagoon has higher overall nutrient concentration than the open sea (Pacific coast). Although the iron concentrations present in these locations are unknown, this could reflect a major metabolic difference between these two *Ostreococcus* strains, which show clear differences in their culture properties. These data justify a complete comparative physiological study of the iron metabolism of these two species.

Our observations thus highlight the importance of nutrient availability and interactions with viruses and grazers as major evolutionary forces in the evolution of phytoplanktonic species.

The high divergence observed between the two *Ostreococcus* genomes involved saturation at synonymous sites ($dS > 2$) for 79% of the 6,270 orthologs. As a consequence, the power of the dN/dS ratio test is too weak to detect positive Darwinian selection on amino acid composition from this genome comparison. However, the analysis of molecular evolution rates gave us insights into the genome dynamics of these species. First, we showed that the two outlier chromosomes, having 1) different base compositions, 2) most of the transposable elements, and 3) fewer orthologous genes, also have faster evolving genes. Thus, an increased mutation rate and/or a relaxed constraint on amino acid composition on these chromosomes could explain their high proportion of species-specific genes, without invoking massive horizontal gene transfer, as suggested previously (Palenik et al. 2007). Second, we showed that faster evolving genes contain more transmembrane proteins, as seen in yeast (Julenius and Pedersen 2006). This increased proportion of transmembrane proteins in faster evolving genes is likely to be the consequence of positive selection (extracellular proteins may interact with the environment and are thus potential targets for infecting pathogens). However, we cannot exclude the role of relaxed selection constraints on extracellular proteins (as a consequence of their fewer interactions with other proteins, for example).

Another striking observation of this genome analysis is the negative relationship between GC content and chromosome length. Because recombination rate decreases with chromosome length (Jensen-Seaman et al. 2004) and that GC content increases with recombination rate (Meunier and Duret 2004), probably via the process of biased gene conversion (Birdsell 2002), this suggests indirect evidence for recombination over a large evolutionary timescale in *Ostreococcus*. There is no experimental evidence yet that these haploid organisms are capable of sexual reproduction, but analysis of gene content suggests that some core meiotic genes are indeed present (Ramesh et al. 2005; Derelle et al. 2006). It has also been suggested that chromosome 2 is a sex chromosome (Derelle et al. 2006). Lower GC content and faster rates of evolution are two observations consistent with lack of recombination, as in the nonrecombining regions of the Y or W chromosomes, but these arguments are still too weak to provide definitive proof for sexual reproduction in *Ostreococcus*. Further, experimental analysis is required to assess whether meiosis is possible in these organisms.

Supplementary Material

Supplementary tables A1–A4 are available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

We are grateful to Gurvan Michel and Tristan Barbeyron for discussions on protein analysis software. We would also like to thank Nigel Grimsley, Pierre Rouzé, Stefan Rombauts, Klaas Vandepoele, and Yves van de Peer for stimulating discussions and comments. Igor Grigoriev (JGI DOE) and Brian Palenik (Scripps institution of oceanography, University of California, San Diego) are acknowledged for ongoing collaboration on *Ostreococcus* genome projects. The work presented here was conducted within the framework of the "Marine Genomics Europe" European Network of excellence (2004–2008) (GOGE-CT-505403).

Literature Cited

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. J Mol Biol. 215:403–410.

Armbrust VE, Berges JA, Bowler C, et al. (45 co-authors). 2004. The genome of the diatom *Thalassiosira pseudonana*: ecology, evolution, and metabolism. Science. 306:79–86.

Ashburner M, Ball CA, Blake JA, et al. (20 co-authors). 2000. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet. 25:25–29.

Bendtsen JD, Nielsen H, von Heijne G, Brunak S. 2004. Improved prediction of signal peptides: signalP 3.0. J Mol Biol. 340:783–795.

Birdsell JA. 2002. Integrating genomics, bioinformatics, and classical genetics to study the effects of recombination on genome evolution. Mol Biol Evol. 19:1181–1197.

Bradnam KR, Seoighe C, Sharp PM, Wolfe KH. 1999. G+C content variation along and among *Saccharomyces cerevisiae* chromosomes. Mol Biol Evol. 16:666–675.

Chretiennot-Dinet MJ, Courties C, Vaquer A, Neveux J, Claustres H, Lautier J, Machado MC. 1995. A new marine picoeukaryote: Ostreococcus tauri gen. et sp. nov. (Chlorophyta, Prasinophyceae). Phycologia. 34:285–292.

Countway P, Caron D. 2006. Abundance and distribution of *Ostreococcus* sp. in the San Pedro channel, California, as revealed by quantitative PCR. Appl Environ Microbiol. 72:2496–2506.

Courties C, Vaquer A, Troussellier M, Lautier J, Chretiennot-Dinet MJ, Neveux J, Machado C, Claustre H. 1994. Smallest eukaryotic organism. Nature. 370:255.

Derelle E, Ferraz C, Lagoda P, et al. (12 co-authors). 2002. DNA libraries for sequencing the genome of *Ostreococcus tauri* (Chlorophyta, Prasinophyceae): the smallest free-living eukaryotic cell. J Phycol. 38:1150–1156.

Derelle E, Ferraz C, Rombauts S, et al. (26 co-authors). 2006. Genome analysis of the smallest free-living eukaryote *Ostreococcus tauri* unveils many unique features. Proc Natl Acad Sci USA. 103:11647–11652.

Gene Ontology Consortium. 2001. Creating the gene ontology resource: design and implementation. Genome Res. 11: 1425–1433.

Guillou L, Eikrem W, Chretiennot-Dinet MJ, Le Gall F, Massana R, Romari K, Pedros-Alio C, Vaulot D. 2004. Diversity of picoplanktonic prasinophytes assessed by direct nuclear SSU rDNA sequencing of environmental samples and novel isolates retrieved from oceanic and coastal marine ecosystems. Protist. 155:193–214.

Jensen-Seaman MI, Furey TS, Payseur BA, Lu Y, Roskin KM, Chen CF, Thomas MA, Haussler D, Jacob HJ. 2004. Comparative recombination rates in the rat, mouse, and human genomes. Genome Res. 14:528–538.

Julenius K, Pedersen AG. 2006. Protein evolution is faster outside the cell. Mol Biol Evol. 23:2039–2048.

Kanehisa M, Goto S, Hattori M, Aoki-Kinoshita KF, Itoh M, Kawashima S, Katayama T, Araki M, Hirakawa M. 2006. From genomics to chemical genomics: new developments in KEGG. Nucleic Acids Res. 34:D354–D357.

Krogh A, Larsson B, von Heijne G, Sonnhammer EL. 2001. Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. J Mol Biol. 305:567–580.

Li W. 1994. Primary production of prochlorophytes, cyanobacteria, and eucaryotic ultraphytoplankton: measurements from flow cytometric sorting. Limnol Oceanogr. 39:169–175.

Liti G, Louis EJ. 2005. Yeast evolution and comparative genomics. Annu Rev Microbiol. 59:135–153.

Meunier J, Duret L. 2004. Recombination drives the evolution of GC-content in the human genome. Mol Biol Evol. 21: 984–990.

Moore LR, Rocap G, Chisholm SW. 1998. Physiology and molecular phylogeny of coexisting *Prochlorococcus* ecotypes. Nature. 393:464–467.

Moreira D, Lopez-Garcia P. 2002. The molecular ecology of microbial eukaryotes unveils a hidden world. Trends Microbiol. 10:31–38.

Nakai K, Horton P. 1999. PSORT: a program for detecting sorting signals in proteins and predicting their subcellular localization. Trends Biochem Sci. 24:34–36.

Nielsen H, Engelbrecht J, Brunak S, von Heijne G. 1997. Identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites. Protein Eng. 10:1–6.

Not F, Latasa M, Marie D, Cariou T, Vaulot D, Simon N. 2004. A single species, *Micromonas pusilla* (Prasinophyceae), dominates the eukaryotic picoplankton in the Western English Channel. Appl Environ Microbiol. 70:4064–4072.

Palenik B, Grimwood J, Aerts A, et al. (38 co-authors). 2007. The tiny eukaryote *Ostreococcus* provides genomic insights into the paradox of plankton speciation. Proc Natl Acad Sci USA. 104:7705–7710.

Piganeau G, Desdevises Y, Derelle E, Moreau H. 2008. Picoeukaryotic sequences in the Sargasso Sea metagenome. Genome Biol. 9:R5.

Piganeau G, Moreau H. 2007. Screening the Sargasso Sea metagenome for data to investigate genome evolution in *Ostreococcus* (Prasinophyceae, Chlorophyta). Gene. 406: 184–190.

Ramesh MA, Malik SB, Logsdon JM Jr. 2005. A phylogenomic inventory of meiotic genes; evidence for sex in Giardia and an early eukaryotic origin of meiosis. Curr Biol. 15:185–191.

Rodriguez F, Derelle E, Guillou L, Le Gall F, Vaulot D, Moreau H. 2005. Ecotype diversity in the marine picoeukaryote *Ostreococcus* (Chlorophyta, Prasinophyceae). Environ Microbiol. 7:853–859.

Sirko A, Brodzik R. 2000. Plant ureases: roles and regulation. Acta Biochimica Pol. 47:1189–1195.

Sonnhammer EL, von Heijne G, Krogh A. 1998. A hidden Markov model for predicting transmembrane helices in protein sequences. Proc Int Conf Intell Syst Mol Biol. 6: 175–182.

Strzepek R, Harrison P. 2004. Photosynthetic architecture differs in coastal and oceanic diatoms. Nature. 431:689–692.

Tatusov RL, Fedorova ND, Jackson JD, et al. (17 co-authors). 2003. The COG database: an updated version includes eukaryotes. BMC Bioinformatics. 4:41.

Thompson JD, Higgins DG, Gibson TJ. 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res. 22:4673–4680.

Vaulot D, Romari K, Not F. 2002. Are autotrophs less diverse than heterotrophs in marine picoplankton? Trends Microbiol. 10:266–267.

Witte CP, Rosso MG, Romeis T. 2005. Identification of three urease accessory proteins that are required for urease activation in *Arabidopsis*. Plant Physiol. 139:1155–1162.

Worden AZ, Nolan JK, Palenik B. 2004. Assessing the dynamics and ecology of marine picophytoplankton: the importance of the eukaryotic component. Limnol Oceanogr. 49:168–179.

Yang Z. 1997. PAML: a program package for phylogenetic analysis by maximum likelihood. Comput Appl Biosci. 13:555–556.

Yang Z, Bielawski JP. 2000. Statistical methods for detecting molecular adaptation. Trends Ecol Evol. 15:496–503.

Yang Z. 2006. On the varied pattern of evolution of 2 fungal genomes: a critique of Hughes and Friedman. Mol Biol Evol. 23:2279–2282.

Zhu F, Massana R, Not F, Marie D, Vaulot D. 2005. Mapping of picoeukaryotes in marine ecosystems with quantitative PCR of the 18S rRNA gene. FEMS Microbiol Ecol. 52:79–92.

Charles Delwiche, Associate Editor

Accepted July 21, 2008