



Screening the Sargasso Sea metagenome for data to investigate genome evolution in *Ostreococcus* (Prasinophyceae, Chlorophyta)

Gwenaël Piganeau^{a,b,*}, Hervé Moreau^{a,b}

^a *Université Pierre et Marie Curie-Paris6, Laboratoire Arago, BP44, 66651 Banyuls sur Mer Cedex, France*

^b *CNRS, UNMR7628, Avenue Fontaulé, BP44, 66651 Banyuls sur Mer Cedex, France*

Received 3 September 2007; received in revised form 7 September 2007; accepted 20 September 2007

Available online 3 October 2007

Abstract

The Sargasso Sea water shotgun sequencing unveiled an unprecedented glimpse of marine prokaryotic diversity and gene content. The sequence data was gathered from 0.8 μm filtered surface water extracts, and revealed picoeukaryotic (cell size $< 2 \mu\text{m}$) sequences alongside the prokaryotic data. We used the available genome sequence of the picoeukaryote *Ostreococcus tauri* (Prasinophyceae, Chlorophyta) as a benchmark for the eukaryotic sequence content of the Sargasso Sea metagenome. Sequence data from at least two new *Ostreococcus* strains were identified and analyzed, and showed a bias towards higher coverage of the AT-rich organellar genomes. The *Ostreococcus* nuclear sequence data retrieved from the Sargasso metagenome is divided onto 731 scaffolds of average size 3917 bp, and covers 23% of the complete nuclear genome and 14% of the total number of protein coding genes in *O. tauri*. We used this environmental *Ostreococcus* sequence data to estimate the level of constraint on intronic and intergenic sequences in this compact genome.

© 2007 Elsevier B.V. All rights reserved.

Keywords: Metagenome; Picoeukaryote; Molecular evolution; Intron; Base composition

1. Introduction

With genome sequencing becoming more and more affordable, the development of environmental shotgun sequencing of environmental microbial communities is providing a challenging amount of sequence data to the scientific community (Streit and Schmitz, 2004; Schloss and Handelsman, 2005). All these sequence data enable the diversity of the microbial world and the metabolic pathways within an environment to be investigated, a previously unthinkable achievement when using traditional approaches (Johnston et al., 2005; Tringe et al., 2005). The most ambitious marine metagenomics project is the global ocean survey (GOS) aiming to sequence picoplankton (smaller than 0.8 μm) diversity in many locations all over the oceans of the

planet (Rusch et al., 2007). The pilot project of this study was published 3 years ago (Venter et al., 2004) with samples from the Sargasso sea.

However, the analysis of these data has been on prokaryotes so far. Eukaryotic diversity represented and the sequence quality has not yet been fully exploited, since most of the sequenced organisms are unknown and there are no clear benchmarks for comparisons (Moreira and Lopez-Garcia, 2002; Schloss and Handelsman, 2005). Picoplankton is defined as a fraction of unicellular organisms having a cell size ranging from 0.2 to 2 or 3 μm (Li, 1994). Picoplankton is constituted both by prokaryotic and eukaryotic cells which can be either heterotrophic or autotrophic. The ecology of picoplankton has been an intense field of investigation this last decade and it now appears to play major roles in biogeochemical cycles that occur in oceans, especially in oligotrophic areas (Campbell et al., 1994; Li, 1994; Rocop et al., 2002). The diversity of prokaryotes has been much better studied than that of eukaryotes, mostly by PCR 16S rRNA gene based approaches (Giovannoni et al., 1990; Suzuki et al.,

Abbreviation: SSD, Sargasso Sea Database; AT, Adenine and thymine; GC, Guanine and cytosine.

* Corresponding author. Tel.: +33 4 68 88 73 43; fax: +33 4 68 88 73 98.

E-mail address: gwenael.piganeau@obs-banyuls.fr (G. Piganeau).

1998) or more recently by random sequencing of filtrated sea water (Venter et al., 2004). For example, in samples collected from the Sargasso sea, filtrated through pore size of 0.8 μm and randomly sequenced, the dominating groups were Proteobacteria, Cyanobacteria and species in the CFB phylum (Cytophaga, Flavobacterium, and Bacteroides) (Venter et al., 2004; Rusch et al., 2007). Among photosynthetic bacteria, the two genera *Prochlorococcus* and *Synechococcus* were clearly dominant, as described in many other areas (Partensky et al., 1999; Rocap et al., 2002).

However, although picoeukaryotes are known to be a minor component of picoplankton in terms of cell number, these organisms, at least those which are photosynthetic, are known to play a major role in primary productivity in oligotrophic areas, where it represents up to 80% of the autotrophic biomass (Worden et al., 2004). It has been shown that they usually have a bigger cell volume than prokaryotes, that they are subject to a high grazing mortality and can have higher growth rates compared to cyanobacteria. It has finally been shown that they can be responsible for 75% of net carbon production in some coastal areas (Worden et al., 2004). Picoeukaryote diversity is much less studied than its prokaryote counterpart, although that is starting to change (Moreira and Lopez-Garcia, 2002; Vault et al., 2002; Guillou et al., 2004; Lovejoy et al., 2006; Worden 2006). It is mainly composed of species from phyla such as Haptophytes, Dinoflagellates and Prasinophytes, some phylogenetic groups inside these very broad phyla being still unknown from a cytological point of view (Lopez-Garcia et al., 2001; Moon-van der Staay et al., 2001). Some quantitative studies based on *in situ* hybridisation experiments showed that among these groups, Prasinophytes apparently dominate picoeukaryotes in different oceanic areas, and more precisely the genus *Micromonas* (Not et al., 2004). However, many other species are found ubiquitously, even if they usually represent a minority of cells. For example, *Ostreococcus* has been reported in the Sargasso Sea (Worden, 2006) and shown to be abundant in the Mediterranean sea (Marie et al., 2006) and coastal California (Countway and Caron, 2006).

In the environmental shotgun sequencing of the Sargasso Sea, the analysis was focused on prokaryote diversity and gene content. However, some very small eukaryotes can work their way through the filtration selection used (0.8 μm). This is indeed what has been found in Sargasso Sea samples where 34 18S rRNA sequences were identified in these samples but not analysed in detail (Table S5 in (Venter et al., 2004)). A subsequent analysis (Worden, 2006) reported that the SSD harboured 18S rDNA sequences from *Ostreococcus* and *Micromonas*, and that the *Ostreococcus* sequences were phylogenetically closest to the “deep clade” of this genus initially described in (Rodriguez et al., 2005). Among picoeukaryote species or genera which could pass through the filtration cut off used, *Ostreococcus* is a likely candidate (Venter et al., 2004), a picophytoplankton genus that belongs to Prasinophytes, a group of widespread green algae thought to have diverged very early from the ancestor of all chloroplast-containing green plants and algae. *Ostreococcus*, with a 0.8- μm diameter, presently defines the tiniest free living eukaryotic cell and the smallest currently

described genome for a photosynthetic eukaryotic organism (Courties et al., 1994; Chretiennot-Dinet et al., 1995; Derelle et al., 2002). Since the publication of the shotgun sequencing work on the Sargasso Sea, complete genome sequences of two *Ostreococcus* strains have been obtained (Derelle et al., 2006; Palenik et al., 2007).

In this paper, we investigated the amount and quality of *Ostreococcus* sequences present in the Sargasso sea database using the available genomic data from *Ostreococcus*. We used this environmental *Ostreococcus* genome data to investigate the proportion of constrained sites in introns and intergenic regions.

2. Materials and methods

2.1. Data

The Sargasso Sea sequence data (Venter et al., 2004) was retrieved from GenBank at ([http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=Nucleotide&cmd=Search&term=CH004737:CH236877\[PACC\]](http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=Nucleotide&cmd=Search&term=CH004737:CH236877[PACC])). This sequence data is the database of scaffolds not associated with any particular organism. It was obtained from sample 1–4, filtered through 0.8 μm (Table S1 in Venter et al., 2004). Hereafter, we will refer to this database as SSD. We searched for the presence of mitochondrial, chloroplastic and nuclear *Ostreococcus* sequences in the SSD. We used the genomes of *Ostreococcus*, one of the strain *O. tauri* isolated from a French coastal lagoon (Derelle et al., 2006), and the other which is a pacific strain (Palenik et al., 2007) to assign SSD sequences to the *Ostreococcus* group. The nuclear, mitochondrial and chloroplastic *Ostreococcus tauri* gene content was retrieved from the web site of the Bioinformatics and Evolutionary Genomics Lab in Ghent (http://bioinformatics.psb.ugent.be/genomes/Ostreococcus_tauri/).

The *Ostreococcus lucimarinus* gene content was retrieved from the Joint Genome Institute (http://shake.jgi-psf.org/Ost9901_3/Ost9901_3.download.ftp.html).

2.2. Retrieval of *Ostreococcus* SSD sequences

We blasted the 8166 predicted protein coding genes of *O. tauri* nuclear genome, 31 chloroplast protein coding sequences and 39 mitochondrial coding sequences on the SSD using blast (Altschul et al., 1990). We used different blast thresholds for organellar and nuclear genes to take into account the higher conservation observed in organellar genes as compared to nuclear genes (Table 1). These thresholds were very conservative as shown by the fraction of *O. tauri* and *O. lucimarinus* genes that would have been retrieved by applying the same thresholds (Table 1).

To assess the minimum number of *Ostreococcus* strains sequenced in the SSD, we searched for overlapping SSD scaffolds matching with high blast score to *O. tauri*'s genome. To assign a SSD sequence to *Ostreococcus* we then used the following procedures. We first extracted the alignments containing coding sequences of *Ostreococcus* (CDS and ribosomal RNA genes). We then estimated the sequence identity between the SSD sequence and one of the two *Ostreococcus*, $\text{Id}_{\text{SSD-O}}$, and

Table 1
Ostreococcus tauri genome coverage in the SSD inferred from the number of coding sequences with a putative homolog

	Nuclear	Chloroplast	Mitochondrion
Number of genes in analysis	8166	31	39
Average GC content	0.59	0.41	0.37
Blastn threshold minimum	$105 - 10^{-20}$	$170 - 10^{-60}$	$101 - 10^{-20}$
Score–maximum e value			
Identity with SSD sequences average (min–max)	85 (77–100)	92 (87–98)	85 (80–95)
Identity between the two <i>Ostreococcus</i> average (min–max)	84 (77–100)	91 (83–99)	84 (80–95)
Expected coverage ^a	44%	95%	71%
Observed coverage in SSD	14%	81%	51%
Fraction of the expected genome coverage sequenced	32%	85%	72%

^a Given the comparison between the two *Ostreococcus* genomes with blast threshold used (see Materials and methods).

between *O. tauri* and *O. lucimarinus*, Id_{OI-OI} . We then used following criteria to conclude that a SSD scaffold was an *Ostreococcus* sequence: the identity of the SSD scaffold and the closest *Ostreococcus* had to be greater or equal than the identity between the two *Ostreococcus* sequences: $Id_{SSD-O} \geq Id_{OI-OI}$.

2.3. Comparison of AT content in alignments

For each SSD scaffold, we computed the length; the number of gaps, the distance between gaps and the base composition using computer programs developed by G.P. (C language). Statistical analysis were performed with R software (<http://www.R-project.org>).

To compare the AT frequency between the SSD scaffolds and the AT frequency of corresponding *O. tauri* sequence, we derived the variance, V , of the average of the difference in AT frequency between the 2 sequences, M . Under the null hypothesis of no difference in AT composition, M follows a normal distribution of mean 0 and variance V . $V = \frac{1}{n^2} \sum_{i=1}^n \frac{f_i(1-f_i) + f_i'(1-f_i')}{k_i}$, with n the number of SSD scaffolds used, k_i the length of the alignment over which the AT frequencies of the SSD scaffold, f_i , and the corresponding *O. tauri* sequence, f_i' .

2.4. Estimation of the selective constraint on introns.

We retrieved the 10 SSD *Ostreococcus* scaffolds containing the largest number of genes in same order and orientation as in *O. tauri*: 7 genes (CH008175.1), 5 genes (CH008341.1, CH024020.1, CH022316.1, CH022586.1), 4 genes (CH009042.1, CH023426.1, CH007651.1, CH008492.1) and 3 genes (CH027561.1). We aligned them to *O. tauri* genomic sequences using LFASTA (Pearson and Lipman 1988). We then manually extracted exonic, intronic and intergenic sequence alignments using *O. tauri*'s gene annotation (Derelle et al., 2006). We extracted 5 complete intergenic regions and 11 complete intronic sequences, of which we kept 7 introns, after removal of introns without canonical GT–AG splice sites in both sequences.

Substitution rates on synonymous sites, dS , were estimated with PAML (Yang 1997) using the codeml program with codon substitution model F3x4, substitution rates in introns, dI , were estimated with PAML using the baseml program.

Assuming that dS gives the neutral mutation rate, the fraction of intronic sites that are under selective constraint, f , (that is that mutations occurring on these sites are selectively removed by natural selection), we can write dI as the product of dS by the fraction of sites evolving neutrally, $1-f$, that is $dI = (1-f) \times dS$. So that $f = 1 - dI/dS$. This is a conservative measure of f since selection on codon usage bias leads to an underestimation of dS and thus to an underestimation of f . Intergenic sequences appeared to be too divergent to get significant alignments (substitution rate estimates from intergenic alignments were always greater than 2).

3. Results and discussion

3.1. Analysis of the *Ostreococcus* SSD sequence data

We found 731 SSD scaffolds matching nuclear *Ostreococcus* sequences with blast scores above the threshold defined in Table 1, with an average length of 3917 bp (excluding gaps). The data contains many gaps, on average one gap every 2322 bp. Despite this high gap frequency, up to 41% of the scaffolds contain genes in synteny groups, that is up to 7 genes in same order and orientation as in *O. tauri*. These data mean that 23% of the complete genome and 14% of the predicted proteins of *O. tauri* have a match in the SSD. This sequence data is scattered over the 20 chromosomes of *O. tauri*, and represents 7% (chromosome 9) to 20% (chromosome 12) of predicted proteins. Mitochondrial and chloroplast of *O. tauri* genome coverage were significantly higher, with 81% and 51% of genes having a match in the SSD, respectively (Table 1). All these additional scaffolds represent unexpected additional data for comparative genomics in this genus. We would like to point out that the number of reads sequenced and assembled for the SSD is 6 to 25 times greater than the number of reads sequenced for any of the seven other Open Ocean locations sampled in the GOS database (Rusch et al., 2007). We could only retrieve a maximum of 10 *Ostreococcus* like sequences from any of these other metagenomes (with the blast thresholds as given in Table 1), so that the data we extracted from the SSD represent to date the largest *Ostreococcus* genome coverage we can expect from a metagenome.

From the alignment of SSD scaffolds with the mitochondrial and chloroplastic genomes, we could show that at least two distinct *Ostreococcus* strains were sequenced (Fig. 1). Indeed, SSD scaffolds CH024056 and CH009852 are both *Ostreococcus* chloroplast scaffolds using the conservative assignment criteria described in the method section (Table 2) and both share 97.6% identity over a 1088 bp overlapping region. SSD scaffolds CH026328 and CH159030 are also both *Ostreococcus* mitochondrial scaffolds (Table 2) and share 83% identity over a 524 bp overlapping region.

The presence of two different strains could not be confirmed with nuclear sequences because no overlapping scaffolds on nuclear genes could be found. For example, we found two

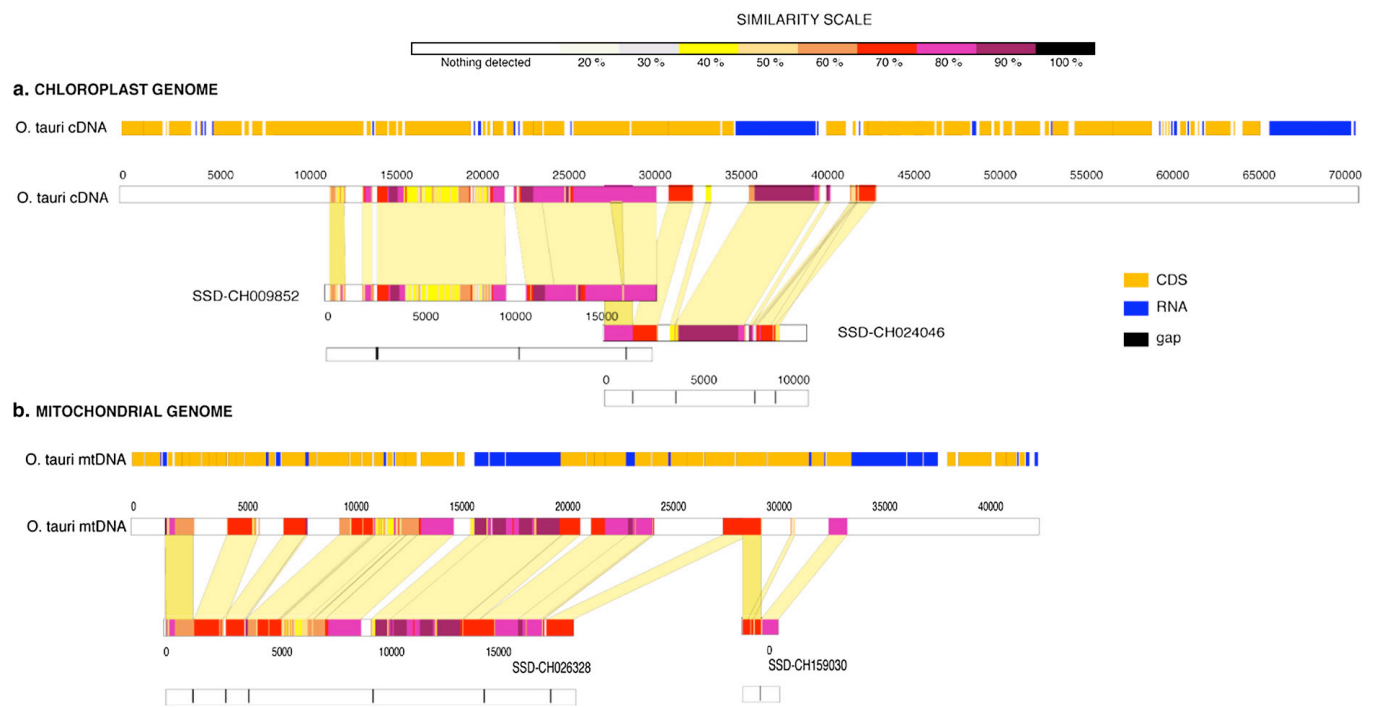


Fig. 1. Alignments of *O. tauri* chloroplast and mitochondrial genome with SSD scaffolds. Graph realized with Lalnview (Duret et al., 1996).

Table 2
Analysis of gene content and identity between overlapping SSD scaffolds and *Ostreococcus* sequences

SSD scaffold	bp ^a	Gene content ^b	Id _{OI-OI}	Id _{SSD-O}
CH026328	19,371	Mitochondrial <i>rps13, rpl5, rpl14, rps3, nad7, atp1, ymf39, atp8, cox1, rrl, cob, rps7, nad2</i>	0.92	0.90
CH159030	1673	Mitochondrial <i>cob, nad4</i>	0.90	1.00
CH024056	11,361	Chloroplast <i>rrl rRNA</i>	1.00	1.00
CH009852	18,895	Chloroplast <i>psbf, psbe, psbh, rps4, rps18, psbl, psac, petb</i>	0.93	0.94

^a Excluding gaps.

^b Most genes are not complete.

Ostreococcus like 18S sequences but they were non-overlapping and could thus belong to the same organism.

This is probably due to the lower nuclear sequence coverage of the SSD scaffolds as compared to the organellar genomes (Table 1). This higher coverage of organellar genomes as compared to the nuclear genome could be solely due to higher rates of evolution of nuclear genes, and thus a mere consequence of the homology based search we used. To correct for the higher rate of evolution of nuclear sequences, we compared the observed coverage of the genome by SSD sequences to the expected coverage, given by the fraction of the genome above the threshold blast scores used between the two available *Ostreococcus* strains. This comparison showed that there is a depletion in *Ostreococcus* nuclear sequences after correcting for different evolutionary rates: assuming an equal coverage of the nuclear and organellar genomes and given the blast thresholds used, we would expect 44% of the nuclear genome to be covered and 95 and 85% of the chloroplast and mitochondrial genomes to be covered (Table 1). The observed coverage represents 32% of the expected coverage, a fraction significantly smaller than those obtained for the organellar genomes: that is 85 and 72% (Chi square, $df=2$, p value $< 10^{-12}$). There are three possible explanations for this higher coverage bias towards organellar genomes. First, it is likely that organellar DNA is present in more copies than nuclear DNA as observed in the green alga *Chlamydomonas* (Bastia et al., 1971). Second, the nuclear genome may be more fragile compared to the small and circular genome of organelles. Third, it could be due to experimental bias between sequences having different G+C content (59%, 41% and 37% for nuclear, chloroplast and mitochondria, respectively). Consistent with the later, the SSD is extremely AT rich (Foerstner et al., 2005): the average GC content in the SSD is 38.6%, whereas the GC content of the *Ostreococcus* sequence data retrieved from the SSD, that aligns with *O. tauri*, is 58.4%. This is not significantly different from the GC content of *O. tauri* (58.2%) (Exact test of differences, p value=0.11). Foerstner and colleagues argued that environment shapes the nucleotide composition of prokaryotic genomes because the Sargasso Sea sequences have a higher AT content than the sequences from other environments, although the causes responsible of this compositional bias are not clear yet. Our analysis suggests that the base composition bias is opposite in *Ostreococcus* as compared to the bias observed in prokaryotes. Since there is no sequence

data for Prasinophytes from non-marine environment, it is not possible to draw conclusions about GC content specificity of marine organisms in this phylum. We show that there is no difference in base composition over the aligned regions between the SSD *Ostreococcus* strains and the Mediterranean *O. tauri*. However, it is nevertheless interesting to note that the GC content of the complete SSD *Ostreococcus* scaffolds is 54.7%, slightly lower than the average GC content of *O. tauri* (58%), even after correction of the contribution of each chromosome in the SSD (58.3%). This difference could be due to the insertion of AT richer elements in the SSD scaffolds.

3.2. Evolutionary genomics perspectives opened from metagenomes.

We show evidence that the SSD contains a random sample of at least two *Ostreococcus* genomes. Despite a high frequency of gaps and a short average scaffold length, it provides a unique opportunity to look at the evolution of non-coding regions in this group. Understanding the role of non-coding regions and identifying new regulatory regions of gene expression is a major challenge in evolutionary genomics (Halligan et al., 2004) and sequence availability of non-coding regions is scarce except for completely sequenced organisms. Recent investigations suggest that a large amount of non-coding sequences, at least as large as the coding regions, participate to the regulation of gene expression. This has been suggested in the fruit fly *Drosophila melanogaster* (Bergman and Kreitman, 2001), the nematodes *C. elegans* and *C. briggsae* (Shabalina, 1999), and in mice and human (Dermitzakis et al., 2002). *O. tauri* has a 12.6 Mb compact genome with an extremely high gene density (82% of coding sequences), that is in part due to extensive reduction of intergenic regions and short introns, which are present in only 39% of genes (Derelle et al., 2006). This raises the question of the structure of the regulatory sequences of gene expression that are localized in intergenic regions and introns. It is foreseen that intergenic regions and introns should contain a high proportion of these regulatory sequences in *Ostreococcus*. Because mutations accumulate much faster at non functional sites than in functional sites, the level of constraint of these regions should be high. This is what we observed for intron sequences, with an average level of constraint, f , of 70% ($n=7$). However, we were unable to estimate a positive f from intergenic regions from our sequence sample. This may be because intergenic functional elements have evolved faster than intronic ones. Since the average rate of synonymous substitution between the SSD *Ostreococcus* and the *O. tauri* sequence is close to saturation (average $dS=1.3$) (each neutral site has changed more than once on average since the divergence of the species), only highly constrained sites will be conserved between these distantly related species. The recent publication of a second *Ostreococcus* genome (Palenik et al., 2007) enabled us to confirm this finding on a larger dataset. We retrieved 42 conserved introns from orthologous genes between *O. tauri* and *O. lucimarinus*. The average f estimated from this additional data equals 0.64, whereas intergenic sequences were too divergent to be aligned. This is consistent a higher rate of evolution of intergenic versus intronic sequences in *Ostreococcus*.

4. Conclusions

We screened the SSD for the presence of *Ostreococcus* sequences, the smallest picoeukaryote known so far. We found evidence for the presence of at least two *Ostreococcus* strains, while previous work, based on 18S rRNA genes, could only discern one type. The chloroplast and mitochondrial genome coverage is significantly higher than the nuclear genome coverage, what might be due to striking differences in GC content or a higher copy number of organellar genomes as compared to the nuclear genome. This SSD *Ostreococcus* data is highly fragmented, but still sums up to approximately 23% of the complete genome and 14% of the total *O. tauri* gene content.

We used these data to investigate the level of constraint on non-coding regions and found that introns have a high proportion of constrained sites (70%) in this tightly compacted genome.

Acknowledgements

We are indebted to the *Ostreococcus* genome sequencing consortium: Yves Van de Peer's group in Ghent, especially Stephane Rombauts for its work on gene annotation in *O. tauri*, the Joint Genome Institute and Brian Palenik. We would also like to thank Evelyne Derelle, Nigel Grimsley, Séverine Jancek and Sebastien Gourbiere for insightful comments and discussions. This work was greatly improved by suggestions of three anonymous referees. The work presented here was conducted within the framework of the "Marine Genomics Europe" European Network of Excellence (2004–2008) (GOCE-CT-2004-505403).

References

- Altschul, S.F., Gish, W., Miller, W., Myers, E.W., Lipman, D.J., 1990. Basic local alignment search tool. *J. Mol. Biol.* 215, 403–410.
- Bastia, D., Chiang, K.S., Swift, H., Siersma, P., 1971. Heterogeneity, complexity, and repetition of the chloroplast DNA of *Chlamydomonas reinhardtii*. *Proc. Natl. Acad. Sci. U. S. A.* 68, 1157–1161.
- Bergman, C.M., Kreitman, M., 2001. Analysis of conserved noncoding DNA in *Drosophila* reveals similar constraints in intergenic and intronic sequences. *Genome Res.* 11, 1335–1345.
- Campbell, L., Nolla, H.A., Vaulot, D., 1994. The importance of *Prochlorococcus* to community structure in the central North Pacific Ocean. *Chlorococcus* sp. (Prochlorophyta) strains isolated from the North Atlantic and the Mediterranean Sea. *Plant. Limnol. Oceanogr.* 39, 954–961.
- Chretiennot-Dinet, M.J., Courties, C., Vaquer, A., Neveux, J., Claustre, H., Lautier, J., Machado, M.C., 1995. A new marine Picoeukaryote—*Ostreococcus tauri* gen. et sp. nov. (Chlorophyta, Prasinophyceae). *Phycologia* 34, 285–292.
- Countway, P., Caron, D.A., 2006. Abundance and distribution of *Ostreococcus* sp. in the San Pedro Channel, California, as revealed by quantitative PCR. *Applied Environ. Microbiol.* 72, 2496–2506.
- Courties, C., Vaquer, A., Troussellier, M., Lautier, J., Chretiennot-Dinet, M.J., Neveux, J., Machado, C., Claustre, H., 1994. Smallest eukaryotic organism. *Nature* 370, 255.
- Derelle, E., et al., 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., et al., 2006. Genome analysis of the smallest free-living eukaryote *Ostreococcus tauri* unveils many unique features. *Proc. Natl. Acad. Sci. U. S. A.* 103, 11647–11652.
- Dermitzakis, E., et al., 2002. Antonarakis. Numerous potentially functional but non-genic conserved sequences on human chromosome 21. *Nature* 420, 482–578.
- Duret, L., Gasteiger, E., Perriere, G., 1996. LALNVIEW: a graphical viewer for pairwise sequence alignments. *Comput. Appl. Biosci.* 12, 507–510.
- Foerster, K.U., von Mering, C., Hooper, S.D., Bork, P., 2005. Environments shape the nucleotide composition of genomes. *EMBO Rep.* 6, 1208–1213.
- Giovannoni, S.J., Britschgi, T.B., Moyer, C.L., Field, K.G., 1990. Genetic diversity in Sargasso Sea bacterioplankton. *Nature* 345, 60–63.
- Guillou, L., Eikrem, W., Chretiennot-Dinet, M.J., 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.
- Halligan, D.L., Eyre-Walker, A., Andolfatto, P., Keightley, P.D., 2004. Patterns of evolutionary constraints in intronic and intergenic DNA of *Drosophila*. *Genome Res.* 14, 273–279.
- Johnston, A.W., Li, Y., Ogilvie, L., 2005. Metagenomic marine nitrogen fixation-feast or famine? *Trends Microbiol.* 13, 416–420.
- Li, W., 1994. Primary production of prochlorophytes, cyanobacteria, and eukaryotic ultraphytoplankton: measurements from flow cytometric sorting. *Limnol. Oceanogr.* 39, 169–175.
- Lopez-Garcia, P., Lopez-Lopez, A., Moreira, D., Rodriguez-Valera, F., 2001. Diversity of free-living prokaryotes from a deep-sea site at the Antarctic Polar Front. *FEMS Microbiol. Ecol.* 36, 193–202.
- Lovejoy, C., Massana, R., Pedros-Alio, C., 2006. Diversity and distribution of marine microbial Eukaryotes in the Arctic ocean and adjacent seas. *Appl. Environ. Microbiol.* 72, 3085–3095.
- Marie, D., Zhu, F., Balagué, V., Ras, J., Vaulot, D., 2006. Eukaryotic picoplankton communities of the Mediterranean Sea in summer assessed by molecular approaches (DGGE, TTGE, QPCR). *FEMS Microbiol. Ecol.* 55, 403–415.
- Moon-van der Staay, S.Y., De Wachter, R., Vaulot, D., 2001. Oceanic 18S rDNA sequences from picoplankton reveal unsuspected eukaryotic diversity. *Nature* 409, 607–610.
- Moreira, D., Lopez-Garcia, P., 2002. The molecular ecology of microbial eukaryotes unveils a hidden world. *Trends Microbiol.* 10, 31–38.
- 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., et al., 2007. The tiny eukaryote *Ostreococcus* provides genomic insights into the paradox of plankton speciation. *Proc. Natl. Acad. Sci. U. S. A.* 104, 7705–7710.
- Partensky, F., Hess, W.R., Vaulot, D., 1999. *Prochlorococcus*, a marine photosynthetic prokaryote of global significance. *Microbiol. Mol. Biol. Rev.* 63, 106–127.
- Pearson, W.R., Lipman, D.J., 1988. Improved tools for biological sequence comparison. *Proc. Natl. Acad. Sci. U. S. A.* 85, 2444–2448.
- Rocap, G., Distel, D.L., Waterbury, J.B., Chisholm, S.W., 2002. Resolution of *Prochlorococcus* and *Synechococcus* ecotypes by using 16S–23S ribosomal DNA internal transcribed spacer sequences. *Appl. Environ. Microbiol.* 68, 1180–1191.
- 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.
- Rusch, D.B., et al., 2007. The Sorcerer II Global Ocean Sampling Expedition: Northwest Atlantic through Eastern Tropical Pacific. *PLoS Biology* e77, 3.
- Schloss, P.D., Handelsman, J., 2005. Metagenomics for studying unculturable microorganisms: cutting the Gordian knot. *Genome Biol.* 6, 229.
- Shabalina, S.A.a.K., A.S., 1999. Pattern of selective constraint in *C. elegans* and *C. briggsae* genomes. *Genet. Res.* 74, 23–30.
- Streit, W.R., Schmitz, R.A., 2004. Metagenomics—the key to the uncultured microbes. *Curr. Opin. Microbiol.* 7, 492–498.
- Suzuki, M., Rappe, M.S., Giovannoni, S.J., 1998. Kinetic bias in estimates of coastal picoplankton community structure obtained by measurements of small-subunit rRNA gene PCR amplicon length heterogeneity. *Appl. Environ. Microbiol.* 64, 4522–4529.

- Tringe, S.G., von Mering, C., Kobayashi, A., Salamov, A.A., Chen, K., Chang, H.W., Podar, M., Short, J.M., Mathur, E.J., Detter, J.C., Bork, P., Hugenholtz, P., Rubin, E.M., 2005. Comparative metagenomics of microbial communities. *Science* 308, 554–557.
- Vaulot, D., Romari, K., Not, F., 2002. Are autotrophs less diverse than heterotrophs in marine picoplankton? *Trends Microbiol.* 10, 266–267.
- Venter, J.C., et al., 2004. Environmental genome shotgun sequencing of the Sargasso Sea. *Science* 304, 66–74.
- Worden, A.Z., 2006. Picoeukaryote diversity in coastal waters of the Pacific Ocean. *Aquat. Microb. Ecol.* 43, 165–175.
- Worden, A.Z., Nolan, J.K., 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.