Ghent University
Faculty of Sciences
Department of Molecular Genetics
VIB Department of Plant Systems Biology
Bioinformatics and Evolutionary Genomics

Barely visible but highly unique: the *Ostreococcus* genome unveils its secrets

Steven Robbens

Promotor: Prof. Dr. Yves Van de Peer

December 2007

Dissertation submitted in fulfillment of the requirements for the degree of Doctor (PhD) in Sciences, Biotechnology

Examination committee

Prof. Dr. Ann Depicker (chairwoman) - Faculty of Sciences, University Ghent

Prof. Dr. Yves Van de Peer (promotor) - Faculty of Sciences, University Ghent

Prof. Dr. Hervé Moreau - CNRS, France

Pierre Rouzé - Faculty of Sciences, University Ghent

Dr. Pieter De Bleser- Faculty of Sciences, University Ghent

Prof. Dr. Dirk Inzé - Faculty of Sciences, University Ghent

Dr. Lieven De Veylder - Faculty of Sciences, University Ghent

Dr. Martin Kuiper - Faculty of Sciences, University Ghent

Dr. Olivier De Clerck - Faculty of Sciences, University Ghent

Thanks to ...

Working in this lab has been really exciting for me, not only for the work I was able to perform, but even more in particular because of the people I worked with. So this page is meant for them…

As it always goes, I also would first like to thank my "boss" Yves. Not because this is the first person you should thank, but because he made my 5 years here really worth it. He gave me all the freedom, and if necessary the knowledge, I needed to finish this PhD. Most of all I appreciate you for not being a "bossy boss": paintball, karting, ice skating, playstation, … you joint all these activities. Good luck in the future with your group!!

Going a bit down in the official hierarchy, but that's really all it means, I bump into Pierre. As we both worked together a lot unveiling all the secrets of *Ostreococcus*, I was, and still am, amazed by your never-ending interest in science. Discovering one "special" gene could make you so enthusiastic, that we only could follow your stream of enthusiasm.

Going a bit away from Gent, I would like to thank Hervé and his team for introducing me to our tiny green friend. In the beginning, Banyuls seemed a remote place to go and to do science, but the beach, the sun, and more in particular the really nice people changed this view completely. Merci beaucoup!

Finally, I return to the people I spent almost 1800 days with, my colleagues. Following the time course of my PhD, many thanks to Klaas for introducing me to bioinformatics, Jan, Jeroen, Stefanie and Tineke for being real members and Lieven for being an ad interim member of the number one island of the bioinformatics room. Many thanks to all the other BEGers for the many laughs, lunches, scientific discussions and most of all the nice atmosphere you created here. Cool place to work!!

Working in this combined wet/dry lab also gave me to opportunity to leave my pc once and a while to interact with wet lab people, especially our Cuban Doctorandus (and his wife). Thanks for the nice talks in the lab and the many moves during one of the many parties.

Finally I would like to thank my parents, my brother and his wife, all my friends and of course Sofie for all the support during this nice period of my life. Thanks.

Table of contents

Chapter 1

Introduction

Introduction: algae (Graham and Wilcox, 1999)

When talking about algae, we all think of these green, slimy plants attached to the surface of a swimming pool; or of the brown seaweeds covering rocks and beaches across the shoreline of the sea. Asian people see algae as an important source of food as they harbour many essential vitamins and irons. Finally, especially women get in touch with algae when they follow a therapy using algae to clean and heal the skin of their body, because algae contain many minerals which feed our skin. So everybody has been in touch with these plants, but not many people are aware of the variety that is present among these algae.

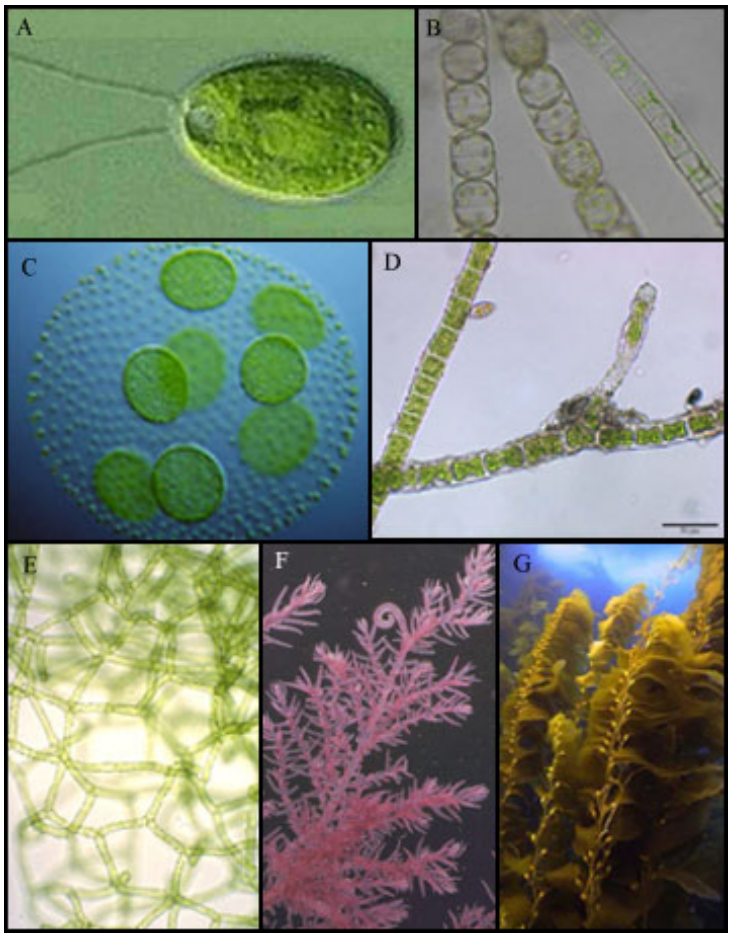


Figure 1. Morphological differences among algae: a) unicellular green alga containing two flagella; b) unbranched, single celled filament; c) spherical colony; d,e) branched filament; f) plant like structure; g) brown kelp

Algae can range in size from tiny single-celled organisms of only 1 micron in diameter to the 65 meter long giant brown kelps, forming underwater forests (fig. 1). In between these two extreme life forms lies a huge morphological diversity. Many algae occur as a unicellular organism, having a simple body plan, while other individual cells aggregate, forming a colony of non-specialized cells. These colonies can consist of a bunch of loosely packed cells, while others are formed in a highly organized fashion, thereby containing a fixed number and arrangement of cells throughout their life cycle. Some of these unicellular algae possess a flagellum which enables them to move around. Besides these unicellular algae, a common growth form among algae is the filament, where a chain of cells is formed after cell division, thereby sharing their cell wall. Filaments can form simple, unbranched structures, composed of a single series of cells; or they can grow into complex structures with many branches and containing multiple rows of cells next to each other. However, parenchymatous algae have the most complex body plan: they possess plant tissue (thallus) composed of undifferentiated cells, providing them with analogous structures as seen in vascular plants (roots, leaves and stems). Although these thalli are undifferentiated, visible differences and functions can be noticed. A nice example within the world of the algae is the kelp, which is a large brown seaweed which can form underwater forests. Their thallus can be divided into three parts: the holdfast, which serves as an anchor; a stipe which supports the blades; and the blades itself, providing them with a vascular leaf-like structure. The huge morphological differences found between algae can be linked to their presence in virtually every ecosystem in the biosphere. Depending on the environment (marine water, fresh water or even terrestrial), the algae have adapted themselves to survive in every circumstance. Algae capture light energy to convert inorganic substances into organic matter through oxygenic photosynthesis. Therefore, the green pigment chlorophyll a, which is the only pigment able to convert light energy into the high energy bonds of organic molecules, is universally present among algae. However, despite the presence of chlorophyll a, not all algae have a green colour. This is caused by the presence of accessory pigments, which are essential for the survival of photosynthetic organisms in both low- and high irradiance habitats: in low-irradiance habitats (the depth record for algae lies around 250 meters, where the light intensity is

only 0.0005% of the surface light) these accessory pigments help in the collection of light which is not completely absorbed by chlorophyll a; in high-irradiance places they help in protecting the algae against photodamage. Depending on the amount and composition of these pigments, eukaryotic algae can be roughly subdivided into different groups: the red algae (Rhodophyta), the brown algae (Phaeophyta) and the green algae (Ben Ali et al, 2001).



Rhodophyta: the 4,000-6,000 members of the red algae can occur as unicells, simple filaments, or complex filamentous aggregations. Besides chlorophyll a, they harbour accessory phycobilins and carotenoids (for protection against photodamage). Members of the light-harvesting phycobili-protein family are extremely efficient in harvesting blue and green lights in subtidal habitats. This explains the abundance of red algae in the greater depths of the oceans. Another unique characteristic among all red algae is the absence of a flagellum, making them immobile. The oldest fossil record was found in arctic Canada and is estimated to be around 750-1,250 million years old. Molecular measurements estimated the split of the red and green algae to have occurred 1,474 million years ago (Yoon et al, 2004)

Phaeophyta: over 1,500 different species are known, which range in structure from microscopic filaments to the giant kelps described above. Both marine and fresh water forms exist, both playing an important role in support of other species: they serve as food, and they create, due to their specific structures, habitats which are used by other species. They possess, besides chlorophyll a, also chlorophyll c and fucoxanthin, a carotenoid which gives them there specific brown colour.

Green algae: the green algae are the largest group of algae, consisting of more than 17,000 different species, living in a variety of habitats. They contain

both chlorophyll a and b but only few accessory pigments. The most famous green alga is *Chlamydomonas reinhardtii*, a fresh water alga which has become a model system for molecular studies. The distribution and classification of the green algae will be discussed in more detail below.

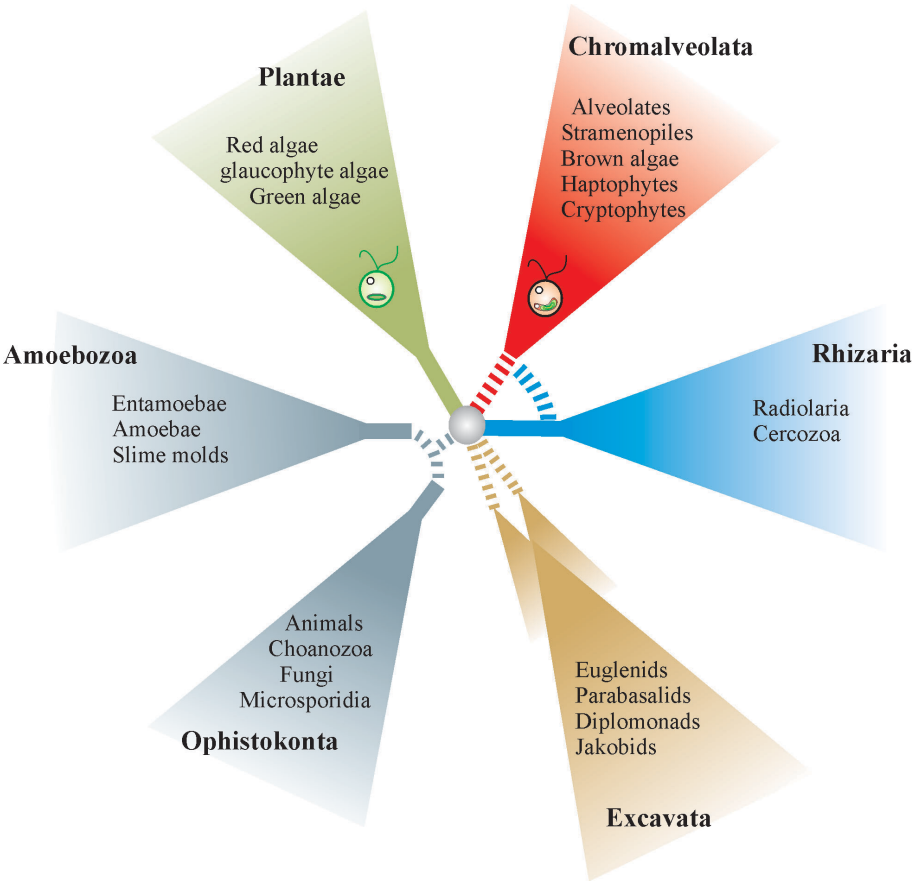


Figure 2. Schematic view of the eukaryotic tree of life showing the putative six supergroups. The broken lines denote uncertainty of branch positions in the tree.

Besides all being classified as algae, red, brown and green algae are located at different nodes within the tree of life. Today multigene phylogenetics. subdivides the eukaryotic tree of Life into 6 "supergroups" (fig. 2), defined as Ophistokonta, Amoebozoa, Plantae, Chromalveolata, Rhizaria and Excavata. Red and green algae are placed within the monophyletic group called Plantae, while brown algae are classified within the Chromalveolata. All members of the Plantae carry a photosynthetic plastid, derived after a free-living photosynthetic cyanobacterial capture and enslavement by a single celled protist. This primary

endosymbiotic event happened once in the common ancestor of all members of the Plantae. Soon after the split of red and green algae within the Plantae, a secondary endosymbiosis event took place, where a eukaryotic red alga, already containing a primary plastid, was engulfed by another host eukaryote, giving rise to the members of the Chromalveolata (containing the brown algae). Once these endosymbiotic events were well established and the new structural composition of the different organisms was inherited by their offspring, red, green and brown algae evolved each separately to the many different life forms we know today. Within this thesis, we only studied members of the green algae and only these will be discussed in more detail.

Green algae

The green algae have always drawn the attention of biologists thanks to the evolutionary relationship between the green algae and land plants. In this respect, the green algae are considered to be "primitive" plants that somewhere in time gave rise to the huge amount of land plants known to date. Most green algae are rather small, ranging from bacterial size to rarely more than one meter in their greatest dimension. Despite the small size, a wide variety of morphological types are present, going from unicellular non-flagellates, over motile colonies to multinucleate coenocytes (multinucleate cytoplasmic mass enclosed by a single cell wall).

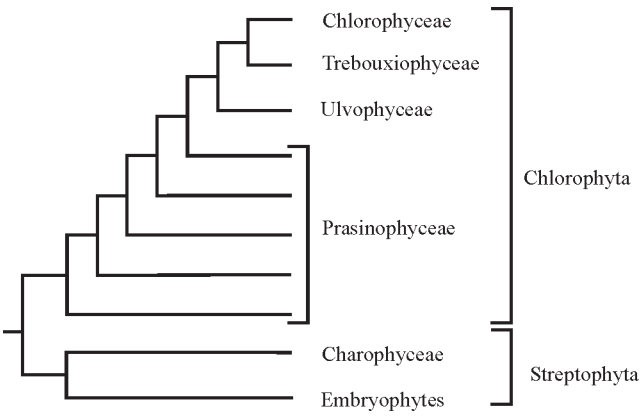


Figure 3. Phylogenetic classification of the green lineage

Classification of these algae was initially based on their growth habitat, as these small organisms could not be easily detected. With the development of new tools, a more in depth classification could be performed, based on their shape and morphology. However, as many vegetative features evolved numerous times, certain characters defining certain groups became unreliable. In this respect, the new data that became available through sequencing of different markers and genomes of different algae, led to a new area of classification, based on phylogeny. Today, the more data that arises from these studies, the more accurate one can produce images of how the world of the green algae really looks like. Molecular data indicates that the green algae originated around 1,500 million years ago (Yoon et al, 2001) and that they diverged from the land plants around 425-490 million years ago (Sanderson, 2003), thereby supporting the age suggested by the fossil record. The complete group of green algae have long been accepted to be the ancestors of land plants (McCourt et al, 1995), but with recent available molecular and morphological data, this doesn't hold true anymore (McCourt et al, 2004; Lewis and McCourt, 2004; and Kapraun 2007). The Chlorophyta have now been split up into two lineages: one known as the chlorophyte green algae (Chlorophyta), which includes the majority of what have been called green algae, and a second lineage entitled Charophyta, containing a smaller number of green taxa. Finally, a third phylogenetic green plant lineage, the Prasinophyceae, constitute a particularly interesting algal class, holding a basal position in the evolution of the extant green lineage (Fawley and Qin, 2000; and Guillou et al, 2004) As a result, today the green lineage can be subdivided into two monophyletic lineages: Streptophyta (land plants and the charophyte green algae) and Chlorophyta (the rest of the green algae) (fig.3) (Karol et al, 2001).

Chlorophyta: this clade, which consists of hundreds of genera and up to more than 10,000 species, contains three major groups: chlorophytes, trebouxiophytes and ulvophytes. Molecular work indicates the monophyly of this group of green algae, whereas the ulvophytes are a sister group of a branch containing both chlorophytes and trebouxiophytes (fig. 3). Ulvophyceans primarily occupy marine waters; Trebouxiophyceans are freshwater and terrestrial algae; while most Chlorophyceans can also be found back in fresh waters. Model

organisms like *Chlamydomonas reinhardtii* and *Volvox carteri* belong to this group of algae.

Charophyta: containing around 65 genera and a few thousand species, they represent the lineage that is ancestral to the land plants. Because of biochemical and structural characteristics unique to both land plants and charophyte green algae and being absent in chlorophyte green algae, both land plants and charophyte green algae were included in the same division, called Streptophyta (fig. 3). When including both the charophyte green algae and the embryophytes (land plants) in one lineage, the Charophyta can be considered as a monophyletic group.

Prasinophyta: members of the class Prasinophyceae have long been identified as being scaly green flagellates, but recent work has profoundly changed the taxonomical view of this class. Today, species with or without flagella and/or scales and even coccoids (Fawley and Qin, 2000) have been added to this class. This leads to a huge diversity in cell shape, organization of the flagellar apparatus and extremely complex assemblages of accessory pigments. These morphological heterogeneities combined with phylogenetic evidence based on SSU rDNA sequences led to the conclusion that the Prasinophyceae are a paraphyletic group.

Numerous prasinophyte algae are amongst the smallest eukaryotic picoplanktonic fraction, having a cell size of less than 3 µm (Stockner 1988; Zignone et al, 2002; and Derelle et al, 2006). Eukaryotic picoplankton plays worldwide an important role in microbial food webs and contribute significantly to microbial biomass and total primary productivity, both in coastal and oligotrophic waters (Johnson and Sieburth, 1979; Joint at al, 1986; and Li 1994). The abundance of the prasinophyte green algae as the major component of the Chlorophyta within the picoplanktonic fraction has been demonstrated by electron microscopy (Joint and Pipe, 1984; Silver et al, 1986), pigment analyses (Everitt et al, 1990; Peeken, 1997; and Rodríguez et al, 2003) and recently by direct gene sequencing (Díez et al, 2001; Moon-van der Staay et al, 2001; and Zeidner et al, 2003). The discovery, the widespread distribution and the large genetic diversity of picoeukaryotes like *Ostreococcus tauri* shows their huge biological importance.

Besides their environmental importance, members of the Prasinophyceae occupy a critical position at the base of the Chlorophyta and even at the base of the entire green lineage. This basal position combined with their primitive morphological make-up, makes them a highly interesting group for further evolutionary and comparative analyses. In this respect, the genome of the two members of the Prasinophyceae, *Ostreococcus tauri* and *Ostreococcus lucimarinus*, was sequenced (chapter 4 and 7).

Ostreococcus tauri

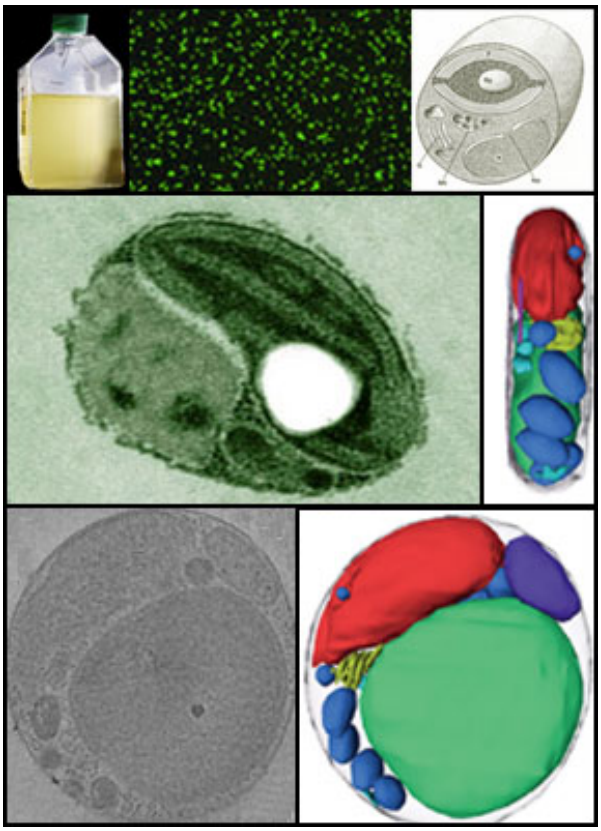


Figure 4. *Ostreococcus tauri* (pictures from Chrétiennot-Dinet et al, 1995; JGI website; and Henderson et al 2007)

The first scientific close encounter with an *Ostreococcus* species happened in the mid-nineties around the following geographical position: 43º24' north and 3º36' east, located in the south of France. Claude Courties and Marie-Josèphe Chrétiennot-Dinet discovered a new photosynthetic picoeukaryote in the marine

Mediterranean Thau lagoon in France, using flow cytometry (Courties et al, 1994 and Chrétiennot-Dinet et al, 1995). This tiny organism measured less than 1 μm, making it the smallest free living eukaryotic organism known to date and was, as a result, barely visible by light microscopy. A bimonthly count and a chlorophyll a biomass measurement indicated that this small organism, which they called *Ostreococcus tauri*, was the main component of the phytoplankton present in the lagoon. These pigment analyses also revealed, besides the presence of an a- and b-type chlorophyll, a c-like pigment indicating its affinity with members of the Prasinophyceae.

Morphologically, *O. tauri* showed a very simple and reduced cellular organisation (fig. 4). Analyses performed by transmission electron microscopy showed a relatively large nucleus with only one nuclear pore, one mitochondrion, a single chloroplast harbouring a starch granule, one Golgi body, and a highly reduced cytoplasm compartment (Chrétiennot-Dinet et al, 1995). A membrane surrounds the cells, but no cell wall can be observed. No flagellum and/or scales are present. Recently, Henderson and co-workers (2007) produced a three dimensional ultrastructure of *O. tauri*, using electron cryotomography. This intact reconstruction confirmed its simple body plan, consisting mainly of two large organelles, the chloroplast and the nucleus, together occupying around 64% of its volume (47% and 17% respectively) and the other organelles situated between them. The cell shape is flattened instead of spherical, having an average cell volume and surface area of 0.91±0.07 μm^3 and 8.3 μm^2, respectively. Besides these known features, some surprising characteristics were discovered. i) Only one (or sometimes two) microtubule(s) of 24±1 nm wide and 200-700 nm long was (were) observed in individual cells. This limited amount of microtubules could indicate that *O. tauri* segregates its twenty pairs of chromosomes one at a time. ii) The nuclear envelope seems to be open throughout the entire cell cycle, while higher eukaryotes have a partial breakdown of the nuclear envelope and in lower eukaryotes the nuclear envelope remains intact. iii) The endoplasmic reticulum of *O. tauri* has two unseen features: it is not always connected to the nuclear envelope and it lacks a ribosome-decorated part (rough ER). iv) Finally, the nucleus contains not only one, but up to three nuclear pore complexes, having a diameter of around 80 nm.

Since its discovery in the Thau lagoon in France, many groups have identified

Ostreococcus isolates at numerous sites around the world (fig. 5): near the French coastal site of the English Channel (Roscoff) (Romari and Vaulot, 2004; and Guillou et al, 2004), Long Island Bay in New York (O'Kelly et al, 2003), the Mediterranean Sea (Díez et al, 2001; and Zhu et al, 2005), Pacific Ocean coastal site in the Southern California Bight (Worden et al, 2004), the Arabian Sea (Brown et al, 2002), and the San Pedro Channel in California (Countway and Caron, 2006). At the same time, attention has focused on the tremendous diversity of picoeukaryotes (López-García et al 2001 and Moon-van der Staay et al, 2001), which holds true for *Ostreococcus* as well.

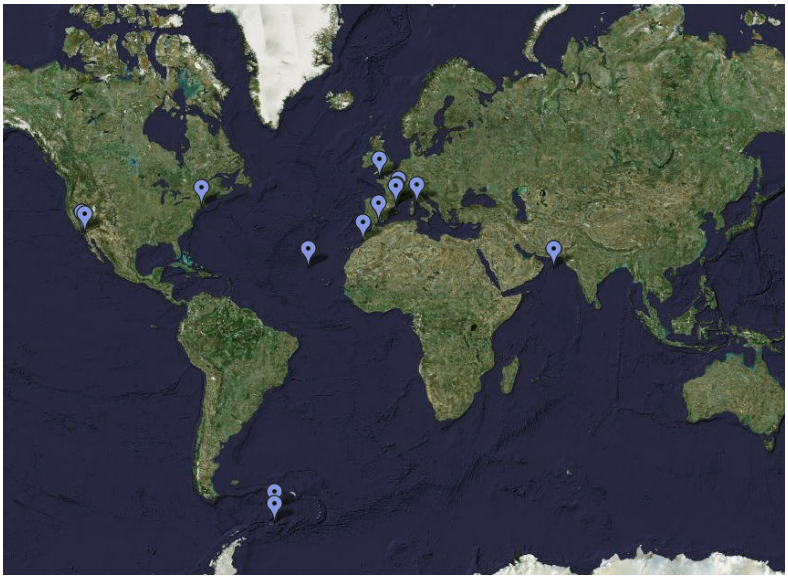


Figure 5. Widespread distribution of *Ostreococcus* (picture made with www.maps.google.be)

Phylogenetic analyses based on several 18S rDNA gene sequences showed that *O. tauri* belongs to the Mamiellales, a subgroup of the Prasinophyceae (Courties et al, 1998), closely related to *Micromonas*-, *Mantoniella*- and *Bathycoccus* species (Not et al, 2004). Rodríguez and co-workers (2005) isolated 12 *Ostreococcus* strains from different marine ecosystems, taken from different depths. Phylogenetic analyses supported previously known existence of four different *Ostreococcus* clades (Guillou et al, 2004). Interestingly, these four different clades could now be linked to the different depths, representing different life conditions of the samples, and not to their geographical origin.

Strains grouped into the different clades showed different karyotypes, hybridization patterns to certain probes, pigment composition, and growth behaviour under a range of light intensities. Within one clade, the same characteristics could be traced back. All this led to the conclusion that light and nutrient conditions, which are changing with depth, are the driving force behind their genetic divergence.
In this respect, besides the genome of *O. tauri* (chapter 4 and 5), the genome of another *Ostreococcus*, namely *Ostreococcus lucimarinus*, has been sequenced (chapter 7). *O. lucimarinus* strains were collected in the upper, illuminated water column of the ocean and, like *O. tauri*, its most striking feature is its minimal cellular organization: a naked, nearly 1-micron cell, lacking flagella, with a single chloroplast and mitochondrion. Comparative analysis of *Ostreococcus* species will help to understand niche differentiation (*O. tauri* was isolated from a coastal lagoon and can be considered light-polyvalent, while *O. lucimarinus* is a representative of high-light surface ocean) and evolution of genome size in eukaryotes.

Genome sequencing

Once the decision has been taken and the necessary money has been found to sequence a new genome, there is still a long way to go before any paper can be written or any conclusion can be drawn regarding the newly obtained data. After isolating the species of interest and extracting its DNA content, the genome needs to be sequenced. There are essentially two ways to sequence a complete genome (Green, 2001; and Meyers et al, 2004): the BAC-by-BAC method (also called hierarchical shotgun or map-based sequencing method), used to sequence yeast and *Caenorhabditis elegans* (Gardner et al, 1981; Smith et al, 1986; Burke et al, 1987; Shizuya et al, 1992; The yeast sequencing consortium, 1997; and The *C. elegans* sequencing consortium, 1998) and the whole genome shotgun sequencing method (Venter et al, 1996), used for both *Ostreococcus* species.

BAC-by-BAC: This method follows a ‘map first, sequence second’ progression: before actually sequencing the genome or smaller genomic region,

a physical map of the whole genome or genomic region is made. Therefore, each chromosome is randomly cut into large pieces (around 150,000 bp), each fragment is than inserted into a Bacterial Artificial Chromosome (BAC). A BAC, which is a man made piece of DNA, combined with external chromosomal DNA is then introduced into bacterial cells where they replicate each time the cell divides. The whole collection of BACs containing an entire genome is called a BAC library, with each BAC being a book that can be read and copied. Next, the fragments of the original chromosome are fingerprinted to give each piece a unique internal identification tag. These tags, which appear every 100,000 bp, will allow creating a map of each chromosome. Once the physical map is made, the actual sequencing can start. Each BAC is then broken at random places into fragments of around 1,500 bp, which are then placed into another vector called M13. Finally, the sequence of each of the 1,500 bp pieces is determined, creating millions of sequence reads, which will be computationally assembled based on sequence overlaps. For reproducing sequence drafts of high quality, a 8-10 fold sequence coverage (the average number of times a segment is present, also known as redundancy) is obtained.

Whole Genome Shotgun: The main difference with the method described above is that no physical map is created. Many copies of the genome are randomly broken into pieces of 2,000 and 10,000 bp long, which are then inserted into small, circular, bacterial molecules known as plasmids. The sequence of all 2,000 and 10,000 bp fragments are determined, thereby generating sequence reads from both ends of the clones. The sequence of both ends is very important for the reassembly of each chromosome. This last step is done by computer algorithms which try to find overlaps between the millions of created fragments: overlapping sequence reads are assembled into contigs, which are than organized into scaffolds, each consisting of a group of sequence contigs held together by read pairs. Scaffolds are than aligned to regenerate the initial genome. This entire procedure is repeated several times to increase the total coverage and to minimize the numbers of errors. The main problem lies in the presence of repetitive sequences which can cause gaps and misassemblies.

Genome annotation

Once the raw sequence data has been generated, bioinformatics comes into play. The genome sequence itself can already provide some useful information about the organisms, but one is of course interested in the number, localisation, structure and ultimately function of the different genes. Gene annotation, and more particular structural annotation, tries to answer the question "where are the genes" and "what do they look like", whereas people performing functional annotation try to get information on what their task is within the organism.

Structural annotation

The first step towards biological knowledge lies in identifying stretches of sequence, usually genomic DNA, that are biologically functional. These include protein-coding genes and other functional genetic elements such as RNA genes and regulatory regions. One can think of creating an algorithm which can be used for every genome, because a gene is a gene and once you know what it looks like, you can easily trace it back. This is unfortunately not true because although all eukaryotic genomes use the same DNA language, each organism has its own dialect. In this respect, for each new project, the known tools and algorithms need to be adjusted and remodelled, taking into account the organism's new dialect.

Essentially, two different types of information are used to locate genes in genomic sequences (Borodovsky et al, 1994; Fickett, 1996; Rouzé et al, 1999; Mathé et al, 2002; and Do and Choi, 2006). i) *Content sensors*. DNA regions can be classified into different categories (coding versus non-coding, exons versus introns). When obtaining this information using similarity based approaches, making use of sequences already available in protein and nucleic acid databases, one measures extrinsic *content sensors*. In general, different types of sequence data are used to provide information about the location of the gene: protein sequences, CoDing Sequences (CDS), transcripts (cDNA or EST) and genomic DNA. The weakness of similarity-based approaches is that or the databases do not contain a sufficiently similar sequence or when good similarity is found, the borders of the detected regions are not always precisely indicated, thereby

not providing enough signals to accurate identify the complete structure of the gene. As a result small genes and exons are quite often missed.
On the other hand, intrinsic properties extract information from the genome sequence itself: nucleotide composition (G+C content), codon composition, hexamer frequency, base occurrence periodicity, etc. Frequency tables and models are built that describe the specific features of the organisms as precise as possible, thereby limiting its use to one organism. Prior to creating these specific feature models, a training set of manual annotated genes needs to be built. The more different genes that are manually annotated, the more information will be gained for creating these models. As these models will then be used to annotate the complete genome and finally to test the outcome of the automatic annotation, the quality of the training set needs to be optimal: only experimental validated sequences should be included. As this experimental validation is often unfeasible, annotators with expertise in certain genes and/or gene families are recruited. In this respect, a set of *Ostreococcus tauri* cell cycle genes was created (chapter 2 and 3) and included in the training set, used for the whole genome annotation (chapter 4).
ii) *Signal sensors* are measures that try to detect the presence of functional sites specific to a gene such as splice sites, start and stop codons, promotors and terminators of transcription, poly A sites, ribosomal binding sites, cleavage sites and various transcription factor binding sites. The basic approach used to find a signal that could represent the presence of a functional site is to search for a match with a consensus sequence. A less strict but more sensitive approach is the use of position weight matrices, in which each position in the signal allows a match to any residue, but different costs are associated with matching each residue in each position. The score returned by a weight matrix sensor for a candidate site is the sum of the costs of the individual residue matches over that site. If the score exceeds a given threshold, the candidate site is predicted to be a true site. Today, more sophisticated types of signal finders are based on artificial intelligence such as Linear Support Vector Machines (Degroeve et al, 2004), Hidden Markov Models (HMM) and neural networks (Do and Choi, 2006).
For the complete genome annotation of both *Ostreococcus* genome sequences, a software called EuGène (Schiex et al, 2001) was used to detect the genes. The two main advantages of working with EuGène are that it combines both

extrinsic and intrinsic properties prior to prediction and that it has been designed to exploit information of external tools and provided resources (fig. 6). Extrinsic information, created by performing a blast (Altschul et al, 1990) of transcripts, expressed sequence tags (ESTs), protein homologs and related genomes to the genomic data will be fed to EuGène. Intrinsic data will be gained by training the different models, representing different characteristics of the gene (intron versus exon, coding versus intergenic, start site, etc.). For splice site prediction, a new software tool, SpliceMachine (Degroeve et al, 2004), was developed in our team and used as an external information source (plugin). After masking transposable elements, EuGène will use all the provided information to predict the structure and localization of the different genes. Once the prediction is finished, the accuracy should be tested by comparing the prediction with an independent training set (one which was not used to train the models). This whole procedure should be repeated until a proper structural annotation is provided.

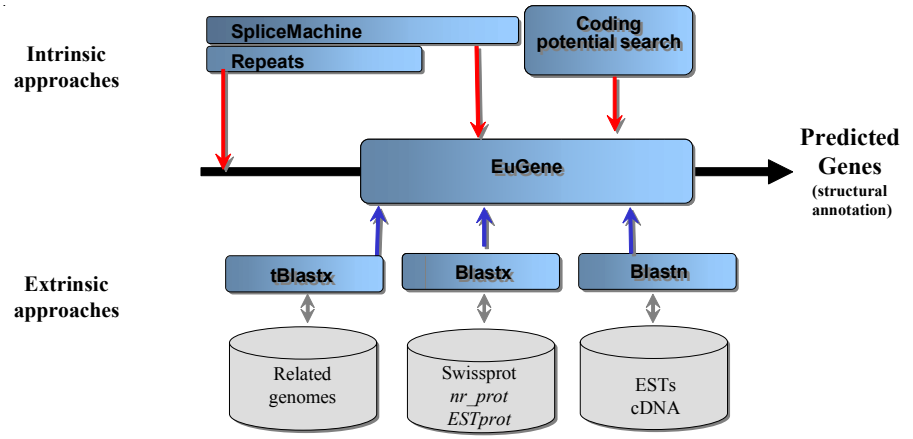


Figure 6. Eugene gene prediction platform

Functional annotation / comparative genomics

Once the different parameters are optimized and EuGène was able to deliver a proper structural annotation, some kind of order is created in the chaos of A, C, G and T's. The genes are now predicted, but their function is still unknown. The most reliable way to pinpoint a function to a certain gene is by doing experimental

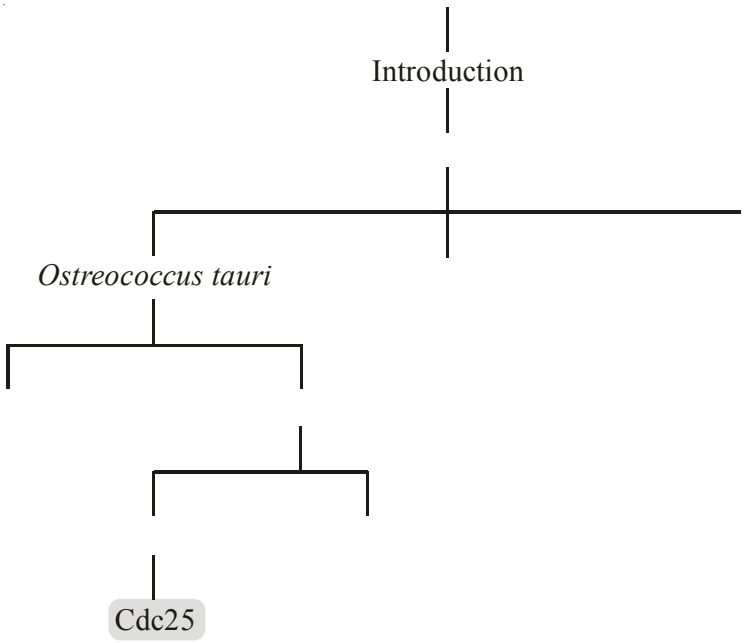
work on that gene. As most researchers are interested in the genes involved in a particular pathway, they can perform some case studies and investigate their gene/pathway of interest. By doing a homology search with their gene(s) of interest against the newly obtained gene set of the target species, one can get already an idea whether the genes are present or not. In this respect, chapter 6 describes the function and distribution of three genes involved in the regulation of the Calvin cycle in *O. tauri*. However, as it is impossible to perform a functional assay for every single gene, a complete fully automated functional prediction is needed. A function can be assigned to a protein based on domain occurrences, available in different databases. InterPro is a collection of various databases, representing protein families, domains and functional sites, in which identifiable features found in known proteins can be applied to unknown protein sequences (Apweiler et al, 2001 and Mulder et al, 2007). The Gene Ontology project (GO) provides a controlled vocabulary to describe gene and gene product attributes in any organism (Ashburner et al, 2000 and Harris et al, 2004). Go consist of three parts: molecular function, biological process and cellular component. The ontologies are structured as directed acyclic graphs, which are similar to hierarchies but differ in that a child, or more specialized, term can have many parents, or less specialized, terms. By scanning these databases, the predicted genes can be labelled with a putative function, based on similarity with the available data.

Next Chapters

The following two chapters will describe the manual annotation of the core cell cycle genes in *O. tauri*, which were used as a part of the training set for the complete genome annotation. Chapter 2 describes the first green lineage Cdc25, while chapter 3 provides a complete overview of the core cell cycle genes present in *O. tauri*. Once the different training sets were created, EuGène was used for the complete genome annotation, which is described in Chapter 4. As a result of the sequencing of its nuclear genome, both the genomes of the chloroplast and mitochondrion were also sequenced. Chapter 5 describes the manual annotation of both organelle genomes and sheds light on the comparison of *Ostreococcus* with other green algae. Once the structure and localisation of

the genes were determined, comparative genomics could provide some insight into the function of certain genes. In this respect, Chapter 6 described the regulation of the Calvin Cycle in *O. tauri* compared to other members of the green lineage. With the availability of a second *Ostreococcus* genome and the knowledge gained by annotating *O. tauri*, our team was involved in the annotation of the *O. lucimarinus* genome and the comparison of both species (chapter 7). Finally, chapter 8 describes a rather strange detection of an in human obesity-linked gene, FTO, in *Ostreococcus* and some other marine organisms.

Chapter 2



Cell Cycle 3(4), 513-518, 2004*

The First Green Lineage cdc25 Dual-Specificity Phosphatase

Basheer Khadaroo[1], Steven Robbens[2,3], Conchita Ferraz[4], Evelyne Derelle[1], Sophie Eychenié[4], Richard Cooke[5], Gérard Peaucellier[1], Michel Delseny[5], Jacques Demaille[4], Yves Van de Peer[2,3], André Picard[1] and Hervé Moreau[1*]

[1] Université Paris VI; Laboratoire Arago; Modèles en Biologie Cellulaire et Evolutive; Banyuls sur Mer, France
[2] Department of Plant Systems Biology, VIB, B-9052, Ghent, Belgium
[3] Department of Molecular Genetics, Ghent University, B-9052, Ghent, Belgium
[4] Institut de Génétique Humaine; Montpellier, France
[5] Université Perpignan; Génome et Développement des Plantes; Perpignan, France

[*] Correspondence: h.moreau@obs-banyuls.fr

Key Words: Cdc25, alga, plant, cell division, cell cycle, regulation, control, G2/M, transition, MPF

Author contribution see appendix page 227

Abstract

The Cdc25 protein phosphatase is a key enzyme involved in the regulation of the G2/M transition in metazoans and yeast. However, no Cdc25 ortholog has so far been identified in plants, although functional studies have shown that an activating dephosphorylation of the CDK-cyclin complex regulates the G2/M transition. In this paper, the first green lineage Cdc25 ortholog is described in the unicellular alga *Ostreococcus tauri*. It encodes a protein which is able to rescue the yeast *S. pombe* cdc25-22 conditional mutant. Furthermore, microinjection of GST-tagged *O. tauri* Cdc25 specifically activates prophase-arrested starfish oocytes. In vitro histone H1 kinase assays and anti-phosphotyrosine Western Blotting confirmed the in vivo activating dephosphorylation of starfish CDK1-cyclinB by recombinant *O. tauri* Cdc25. We propose that there has been coevolution of the regulatory proteins involved in the control of M-phase entry in the metazoan, yeast and green lineages.

INTRODUCTION

Cell cycle regulation is highly conserved throughout evolution and is basically ensured by the kinase activity of the family of cyclin dependent kinases (CDKs) and their regulatory cyclin subunits (Stals and Inze, 2001 and Mironov et al, 1999). Regulatory mechanisms present in some lineages have somehow evolved differently in other species; for example, the Rb/E2F/DP pathway which is present in animals and plants is nevertheless absent in yeast (Rubin et al, 2000). Therefore, comparative studies of the cell cycle among models belonging to different lineages can provide crucial information in distinguishing between the core cell cycle common to all phyla and adaptations specific to a lineage, clade or organism (Vandepoele et al, 2002).

Cdc25 is the key enzyme involved in the regulation of the G2/M transition (Nilsson and Hoffmann, 2000). Its phosphatase activity activates the CDK-cyclin complex which is inhibited by the kinase activity of Wee1 (Berry and Gould, 1996). This regulation pathway is essential for M-phase entry and a deletion of both Wee1 and Mik1 or of Cdc25 in *Schizosaccharomyces pombe* is lethal (Lundgren et al, 1991). The high degree of conservation of the G2/M transition regulation pathway implies that coevolution of the CDK-cyclin complex, Wee1 and Cdc25 has occurred (Goh et al, 2000). This hypothesis predicts that orthologs of all the genes, which were present early during evolution, must be found in the different lineages whose G2/M transition is regulated by the well-conserved CDK-cyclin complexes (Pellegrini et al, 1999). As predicted, CDKs, cyclins and wee1 orthologs are found in all currently sequenced eukaryotic genomes, including metazoans, fungi and plants (Rubin et al, 2000 and Sorrell et al, 2002). Also, orthologs of Cdc25 have been isolated in both vertebrate and invertebrate animals and in fungi, but no Cdc25 ortholog has so far been identified in plants, even though the genomes of *Arabidopsis thaliana* and rice have been completely sequenced. (Smits and Medema, 2001; Russell and Nurse, 1986; Criqui and Genschik, 2002; and Dewitte and Murray, 2003)

Though higher plants lack the cdc25 gene, the inhibitory kinase wee1 gene is well conserved: overexpression of *A. thaliana* Wee1 in fission yeast causes cell cycle arrest—the cells grow but do not divide, thus providing evidence for

a functional Wee1 protein *in A. thaliana* (Sorrell et al, 2002 and Sun et al, 1999). In addition, various studies have shown that

1. The Thr14-Tyr15 CDK site which is dephosphorylated by Cdc25 in animals is conserved in plant CDKs,

2. the in vitro dephosphorylation by *S. pombe* Cdc25 is correlated with the activation of the CDK-cyclin complex in tobacco stem cells, and

3. the in vivo overexpression of *S. pombe* Cdc25 phosphatase in tobacco BY2 cells yields smaller cells reminiscent of the wee phenotype (Zhang et al, 1996 and McKibbin et al, 1998).

Interestingly, CDK activity is also inhibited by tyrosine-phosphorylation in the *Fucus spiralis* zygote (Corellou et al, 2001). These arguments suggest that the plant G2/M transition is switched on by an unknown activating dual-specificity phosphatase and switched off by the inhibitory kinase, Wee1; this unknown green lineage activating phosphatase being very weakly related or unrelated to the Cdc25 phosphatase family (Dewitte and Murray, 2003; and Kraft, 2003).

Ostreococcus tauri is a unicellular alga which diverged very early in the green lineage (Bhattacharya and Medlin, 2004). It is the smallest free-living eukaryotic cell described to date (diameter 1 µm) and has a minimal cellular organization with a nucleus, only one chloroplast and one mitochondrion, and a nude plasma membrane (Courties et al, 1994 and Chrétiennot-Dinet et al, 1995). Its 12.56 Mb* genome which is distributed among 20 chromosomes* is currently being fully sequenced (Derelle et al, 2006)*.

In this chapter, we describe a functional cdc25 gene in *O. tauri* which is the first cdc25 ortholog found in the green lineage. The *O. tauri* cdc25 gene codes for a protein having a dual specificity phosphatase activity which is able to rescue the *S. pombe* Cdc25 conditional mutant and which specifically activates the CDK1-cyclin B complex in prophase-arrested starfish oocytes in vitro and in vivo.

* Updated version after publication

EXPERIMENTAL PROCEDURES

Materials.

The *Ostreococcus tauri* culture used in this study is the strain OTTH0595 (Courties et al, 1998). Cultures were grown in Keller's medium (Sigma-Aldrich, Saint-Quentin Fallavier, France) dissolved in 0.2 µm filtered sea water (NaCl 38 $g.l^{-1}$). Growth conditions were a temperature of 18°C, a permanent irradiance of 60 µmol $quanta.m^{-2}.s^{-1}$ and mild agitation (Derelle et al, 2002). Cell population growth was followed by flow cytometric analysis. The starfishes were collected in the Banyuls bay during the breeding season (November to March for *Marthasterias glacialis*, December to May for *Astropecten aranciacus*) and kept in running sea water (Vee et al, 2001). Pieces of ovaries were taken by incising the starfish dorsal wall. The temperature sensitive *S. pombe h+ cdc25-22 leu1-32 ura4-218 ade6-M210* mutant strain and the pREP41 vector were kindly provided by Dr. G. Lenaers (Université Paul Sabatier, Toulouse, France) (Jimenez et al, 1990 and Forsburg, 1993).

Cloning the *Ostreococcus tauri* cdc25 Gene.

The complete open reading frame (ORF) of *O. tauri* Cdc25 gene was obtained by the combined use of the SMART RACE™ cDNA Amplification kit (Clontech, Palo Alto, CA) and the Universal GenomeWalker™ kit (Clontech).

Phylogenetic Analysis of *Ostreococcus tauri* cdc25 Gene.

Sequences were aligned with CLUSTALW and the alignments were manually improved using BIOEDIT (Thompson et al, 1994 and Hall, 1999). Three different methods were used for the construction of the phylogenetic trees. Maximum likelihood phylogenetic analyses (quartet puzzling, using 25,000 puzzling steps and a VT+γ substitution model) were performed using TREE-PUZZLE.(Muller and Vingron, 2000; Uzzel and Corbin, 1971 and Schmidt et al, 2002) Neighbor-joining trees were constructed using TREECON, based on Poisson corrected distances (Van de Peer and De Wachter, 1997). MrBayes was used for Bayesian inference of phylogenetic trees, using a JTT+γ substitution model (Huelsenbeck et al, 2001 and Jones et al, 1992). Only unambiguously aligned positions were taken into account and in the NJ approach, bootstrap analyses with 500 replicates

were performed to test the significance of the nodes.

Production of Recombinant Cdc25 Protein in Bacteria.

The ORF of cdc25 was PCR amplified using the Advantage2 Taq DNA polymerase (Clontech) with specific primers containing restriction sites for directional cloning by BamHI and XhoI (forward primer FpgexBam : 5'-gtggatccATGGAGGTGCGGGAGGCGAACAAGCGCGCG-3' and reverse primer RpgexXho 5'-tagattatactcgagTCATTCGTTGTCCATGTCGGCCAC-3'). The PCR program used was an initial denaturation step at 95°C for 1 min to hotstart the enzyme, followed by 35 cycles at 94°C for 40 sec, 60°C for 1 min and 68°C for 1.5 min, and a final extension at 68°C for 5 min. The PCR product was gel purified using the Nucleospin® Extract 2 in 1 (Macherey-Nagel, Düren, Germany), and then cloned into the pGEM®-T Easy Vector (Promega, Madison, WI) for sequencing and subcloning. By restriction digesting the recombinant vector with BamHI (Promega) and XhoI (Promega), the cdc25 gene was directionally subcloned into the BamHI-XhoI digested pGex4T1 vector (Amersham Biosciences, Piscataway, NJ) and electroporated in *E. coli* DH5-alpha strain. For production, the recombinant pGex4T1-Cdc25 vector was electroporated in *E. coli* BL21 cells. The bacteria were grown to an A_{600} of 0.5 at RT in LB medium containing 100 µg.mL^{-1} ampicillin. Recombinant protein expression was induced by addition of Isopropyl-beta-D-thiogalactoside (IPTG) to the cells for 6 hours at a final concentration of 0.2 mM. The GST-fused protein was affinity purified on glutathione sepharose using a GStrap FF column with an Äktaprime system (Amersham Biosciences, Orsay, France).

In vitro Phosphatase Assays.

One arm of *Marthasterias glacialis* starfish was sectioned and the gonads were separated from their follicle cells by artificial calcium-free sea water treatment (Vee et al, 2001). Prophase-arrested oocytes were rinsed in normal sea water, aliquoted in IPNP medium (50 mmol.l^{-1} Tris pH 7.5, 150 mmol.l^{-1} NaCl, 50 mmol.l^{-1} NaF, 10 mmol.l^{-1} Na pyrophosphate, 1 mmol.l^{-1} Na3VO4, 10 mmol.l^{-1} Phenylphosphate, 0.1 mg.ml^{-1} soybean Trypsin Inhibitor, 0.1 % (V/V) Triton X100), were flash frozen in liquid nitrogen and stored at -80°C for further use. Aliquots of oocytes were thawed and centrifuged at 15,000 x g for 10 minutes

at 4°C. The extract was incubated at 4°C for 1h with sepharose beads grafted with $p13^{suc1}$ which specifically bind the CDK-cyclin complex. The beads were rinsed three times with TBS/T buffer and then washed with 1X PBS. They were further incubated for different time-lapses with recombinant *O. tauri* GST::Cdc25 at RT and washed three times with TBS/T.

a. Kinase Activity. Histone H1 kinase activity was measured using 5 µl of kinase buffer (8 mM Hepes, 10 mM $MgCl_2$, 0.1 mM ATP, 10 $\mu g.\mu l^{-1}$ histone H1, 1 % (V/V) gamma ^{32}P ATP). The reaction was stopped by adding 20 µl of 1X Laemmli sample buffer and boiled for 3 minutes. After autoradiography of the SDS-PAGE, the histone H1 bands were excised and counted by liquid scintillation (Picard and Peaucellier, 1998).

b. Anti-phosphotyrosine (Anti-Ptyr)Western Blotting. Immunoblotting using the anti-Ptyr antibody (Picard et al, 1996) was revealed by alkaline phosphatase detection method (Picard and Peaucellier, 1998).

Starfish Oocyte Microinjection.

Prophase-arrested *Astropecten aranciacus* oocytes were separated from their surrounding follicle cells by calcium-free sea water treatment, and kept in natural filtered sea water until microinjection following the Hiramoto method (Hiramoto, 1974).

Rescue of the *S. pombe cdc25-22* Conditional Mutant. Cloning into the NdeI-BamHI digested pREP41 yeast expression vector was carried out as for the pGEX4T1 construction, except that:

a. pREP41 specific primers (forward primer: FprepNde : 5'-gtggattcatATGGAGGTGCGGGAGGCGAA-CAAGCGCGCG-3' and reverse primer RrepBam : 5'-actcgaggatccTCATTCGTTGTCCATGTCGGCC-ACTTCGTC-3') were used, and

b. the NdeI (Promega) and BamHI (Promega) restriction enzymes were used (Maundrell, 1990).

The recombinant pREP41 vector was then electroporated into *E. coli* DH5-alpha cells. The resulting construct which contains the *O. tauri* cdc25 gene under control of thiamine-repressible nmt promoter, was transformed into the *S. pombe h+ cdc25-22 leu1-32 ura4-218 ade6-M210* strain by electroporation and selected on leucine- and thiamine-deficient minimal medium (Prentice and

Kingston, 1992). Positive clones growing on leucine-deficient medium at the permissive temperature (30°C) were tested at the restrictive temperature (37°C). Expression of the *O. tauri* Cdc25 was repressed by the addition of 25 µM thiamine.

RESULTS

An Ortholog of cdc25 is Present in *Ostreococcus tauri*.

During the sequencing of the complete genome of *O. tauri*, a clone showing significant similarity to the C-terminal part of the cdc25 gene family was identified. The complete gene was obtained by using a RACE-PCR approach and was later confirmed by the contig assembly of the genome sequencing project. This yielded a complete open reading frame (ORF) of 1,188 bp without any intron (accession number AY330645) encoding a 395 residue protein. *O. tauri* cdc25 ORF has a GC content of 62% which is much higher than the GC content range of 40-45% found in animal and yeast cdc25 orthologs. Blast search analysis of the *O. tauri* Cdc25 protein sequence revealed a typical Cdc25 signature encompassing the 180 amino acids of the C-terminal domain containing the conserved active site, a rhodanese fold with the CE[Y-F]SXXR motif and the aspartate residue of the essential DCR acceptor motif (fig. 1) (Altschul et al, 1997 and Bordo, 2002). No domain homology could be found in the databases for the N-terminal part of the protein, as is usually observed in animal and yeast Cdc25 N-terminal moieties. *O. tauri* cdc25 hosts a consensus 14-3-3 binding site, 2 destruction box (RXXL) motifs but no KEN motif (Tzivion and Avruch, 2002; Zur et al, 2002 and Pfleger and Kirschner, 2000). Finally, 15-serine, 8-threonine and 6-tyrosine potent phosphorylation sites are also counted, among which one putative MAP kinase phosphorylation site on threonine-183 and more importantly 4 putative CDK A phosphorylation sites (Blom et al, 1999).

The expression of *O. tauri* cdc25 was analysed by northern blotting of *O. tauri* total RNA isolated from non synchronized cultures. A band corresponding to approximately 1.6 kb was obtained (data not shown), indicating that the *O. tauri* cdc25 gene is expressed. This size is compatible with the estimated size

of the gene obtained from the complete genome sequencing project. This expression has also been confirmed by specific PCR amplification of a fragment of *O. tauri* cdc25 using a cDNA library and lastly from total RNA by the Titan one step RT-PCR kit (data not shown).

Figure 1. Alignment of Cdc25 amino acid sequences. Comparison of the C-terminal end of the *Ostreococcus tauri* (Ot) Cdc25 protein sequence with other Cdc25s from different organisms (*M. musculus* [Mm], *D. melanogaster* [Dm], *S. pombe* [Sp], *P. carinii* [Pc], *S. cerevisiae* [Sc] and *C. elegans* [Ce]). Identical amino acids are boxed, gaps introduced during the alignment process are indicated with dashes. A consensus sequence was made showing the conserved residues among the Cdc25-family. The two bars indicate the CE[Y-F]SXXR motif and the conserved aspartate residue upstream from the (H)CX5R motif, the typical signature for the Dual-Specificity Phosphatases (DSP) (Shi et al, 1998), to which Cdc25 belongs. The number of amino acid residues encompassing the regions beyond the conserved domains are given in parentheses.

A True cdc25 Dual-specificity Phosphatase in *Ostreococcus tauri.*

Alignments of the superfamily of protein phosphatases dataset, including the divergent family of low molecular weight phosphatases, showed unambiguously that the *O. tauri* sequence belongs to the cdc25 dual-specificity phosphatase subfamily (fig. 2). Within this cdc25 subfamily, *O. tauri* branched off first, being clearly different from metazoan and yeast sequences (fig. 2). Since the systematic and careful searches for orthologs of cdc25 genes in plants by using animal or yeast sequences have been unsuccessful, the availability of *O. tauri* cdc25 sequence originating from the green lineage prompted us to look again for any homology with its conserved C-terminal part. Again, no clear ortholog could be found in the two completely sequenced streptophyta genomes *Arabidopsis thaliana* and rice. Furthermore, no ortholog was found in the very large unpublished EST dataset of the moss *Physcomitrella patens* (Ralf Reski, Freiburg University, Germany, personal communication). Interestingly, a putative Cdc25-like phosphatase sequence containing the HCX_5R motif but a divergent DVR acceptor motif was identified in *Chlamydomonas reinhardtii* (ChlamyDB§ Accession No. 833010A05.y1; http://www.biology.duke.edu/

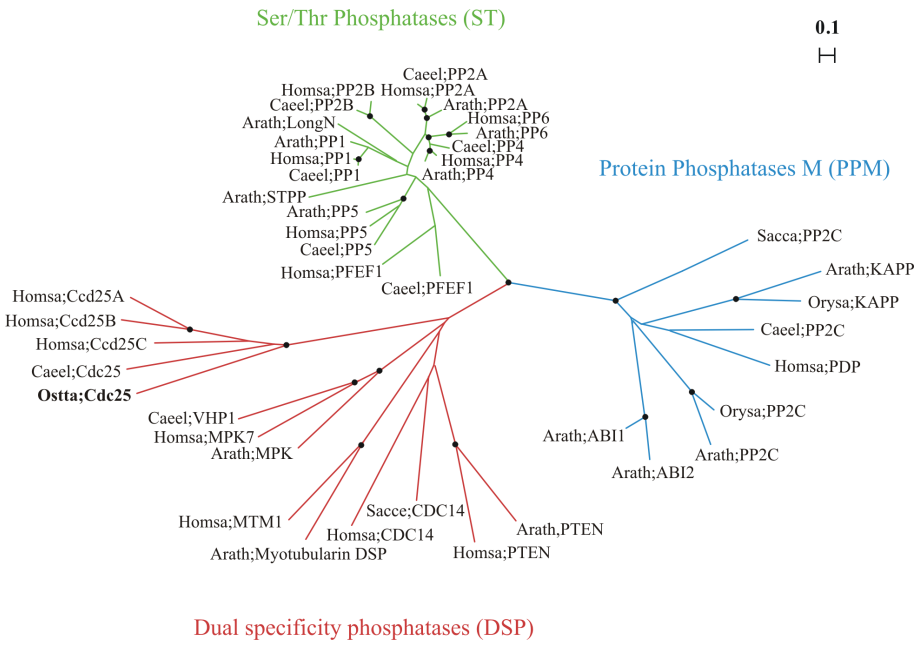


Figure 2. Phylogenetic analysis of Cdc25 genes. Classification of the *Ostreococcus tauri* Cdc25 within the phosphatase superfamily. For each family, representative protein sequences were taken from different organisms (*Saccharomyces cerevisiae* [Sacce], *Homo sapiens* [Homsa], *Arabidopsis thaliana* [Arath], *Caenorhabditis elegans* [Caeel], *Ostreococcus tauri* [Ostta], *Oryza sativa* [Orysa]), in order to perform phylogenetic analyses. Because of their short alignable region compared with the other members of the phosphatase superfamily, the Low Molecular Weight Phosphatases (low-Mr) were not included in the analysis. All tree construction methods (pairwise distance, maximum likelihood and Bayesian inference) gave the same topology. Here, the maximum likelihood tree constructed with TREE-PUZZLE is shown. Dots on the branches indicate a Quartet Puzzling Support Value higher than 70%.

chlamy_ genome). This new cdc25-like candidate sequence is highly divergent and only functional characterization will ensure that it is an ortholog of the Cdc25 phosphatase. These results indicate that a true cdc25 gene was present at the base of the green lineage as an ortholog is found in *O. tauri* and a putative one in *C. reinhardtii*. However, no ortholog has been identified in higher plants.

Ostreococcus tauri Cdc25 Rescues *S. pombe* cdc25-22 Temperature Sensitive Mutants.

Since there are no genetic systems yet available in *O. tauri*, the rescuing of a *Schizosaccharomyces pombe cdc25-22* conditional mutant by recombinant *O. tauri* Cdc25 was tested. This strain which grows normally at the permissive temperature of 30°C arrests at the G2/M transition at the restrictive temperature of 37°C due to inactivation of the thermo-labile endogenous cdc25. The complete ORF of the *O. tauri* gene was subcloned in the *S. pombe* expression vector pREP41 and transformed in the *S. pombe cdc25-22* mutant. After selection on leucine-deficient minimal medium plates, only colonies transformed by the pREP41 containing *O. tauri* cdc25 gene grew at restrictive temperature (37°C), whereas the colonies transfected with the vector alone (control) grew very poorly (fig. 3A and 3B). Furthermore, microscopic examination showed that the colonies transfected with *O. tauri* cdc25 have a similar morphology (small rod-shaped cells) to the wild type strains or to the colonies growing at permissive temperature (fig. 3D). Small and round cells corresponding to the wee phenotype are also observed. In contrast, the cells in the few colonies obtained at restrictive temperature and not rescued by *O. tauri* cdc25 appeared elongated with a phenotype typical of cdc25 cell cycle-arrested mutant cells (fig. 3C).

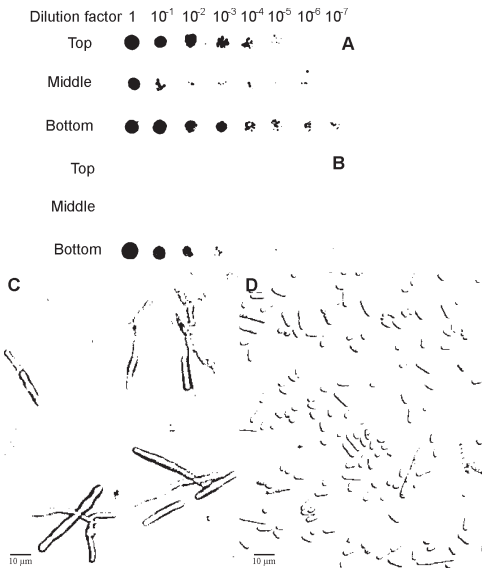


Figure 3. Rescue of *S. pombe cdc25-22* conditional mutant strain. Petri dishes containing 10-fold serial dilutions of (top) nontransformed (middle) pRep41 transformed and (bottom) *O. tauri* Cdc25 transformed *S. pombe cdc25-22* strains grown at 30°C (A) and 37°C (B) in EMM + 225 mg.l^{-1} leucine medium. Only *O. tauri* Cdc25 transformed cells grow under restrictive condition. (C and D) Photomicrographs of liquid cultured cells of recombinant *S. pombe cdc 25-22* at restrictive temperature: longer cells are observed for the pRep41 transformed cells (C) whereas *O. tauri* Cdc25 transformed cells (D) show small round cells reminiscent of the wee phenotype and rod-shaped cells corresponding to the phenotype of the cells growing normally at 30°C.

Ostreococcus tauri Cdc25 is an Active Phosphatase.

To investigate further the phosphatase activity of *O. tauri* cdc25 on the CDK-cyclin complex, starfish oocytes were microinjected with purified recombinant *O. tauri* GST::Cdc25 protein (Picard et al, 1987). These oocytes are arrested in the prophase of the first meiotic division (germinal vesicle {GV} stage) and contain an inactive Thr-14 Tyr-15 phosphorylated CDK (Borgne et al, 1999). Oocytes microinjected with recombinant *O. tauri* Cdc25 phosphatase underwent M-phase entry as shown by germinal vesicle breakdown (GVBD) (Fig. 4). We moreover checked that the GVBD oocytes contained active histone H1 kinase activity showing that the microinjected recombinant *O. tauri* GST::Cdc25 readily activated endogenous CDK-cyclin complex (data not shown). On the other hand, GVBD was not observed in control oocytes microinjected with buffer alone (fig. 4).

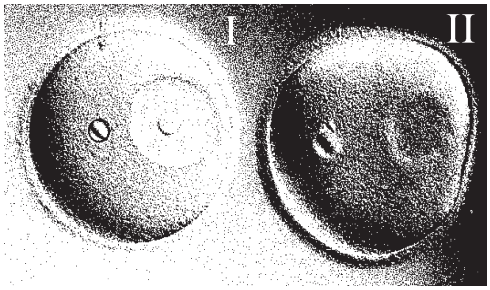


Figure 4. Microinjection of *O. tauri* Cdc25 in prophase arrested starfish oocyte. The recombinant *O. tauri* GST::Cdc25 was microinjected in starfish oocytes. (I) Control oocyte microinjected with buffer solution 1X PBS. The small sphere above the germinal vesicle is a mineral oil drop microinjected together with the solution. (II) Oocyte microinjected with *O. tauri* Cdc25 phosphatase: the oocyte matures into a Germinal Vesicle BreakDown (GVBD) state; the nuclear envelope fades and eventually disappears. The maturation of the oocyte observed shows that the CDK-cyclin complex has been activated by the *O. tauri* Cdc25

We also tested the in vitro activity of *O. tauri* Cdc25 phosphatase. For this, the histone H1 kinase activity of the CDK1-cyclin B complex was measured after addition of recombinant *O. tauri* GST::Cdc25 phosphatase to GV-stage inactive CDK which were affinity-purified from the oocyte protein crude extract by $p13^{suc1}$ grafted sepharose beads. This activity increased by a factor of 2.6 following the addition of recombinant *O. tauri* GST::Cdc25 (fig. 5A) for 1 hour as compared to the control and this was proportional to the time-lapse of *O. tauri* Cdc25 addition to the extracts. To ensure that the *O. tauri* Cdc25

phosphatase specifically activated the CDK1-cyclin B complex, a western-blot revealed by anti-phosphotyrosine antibody was carried out on the purified starfish CDK1-cyclin B complex and incubated with *O. tauri* Cdc25 for various times. The resulting blot clearly shows that the band corresponding to the tyrosine-phosphorylated form of CDK1 decreased significantly on addition of *O. tauri* Cdc25 (fig. 5B). More precisely, the tyrosine-phosphorylated band showed an intermediate band shift before a more than 5-fold decrease in intensity indicating a more rapid threonine than tyrosine dephosphorylation. This dephosphorylation was dose and time dependent and confirmed the activation of the CDK1-cyclin B complex by dual-specificity dephosphorylation. These results clearly show that the recombinant *O. tauri* Cdc25 protein is functional and has a dual specificity phosphatase activity assayed both in vitro and in vivo.

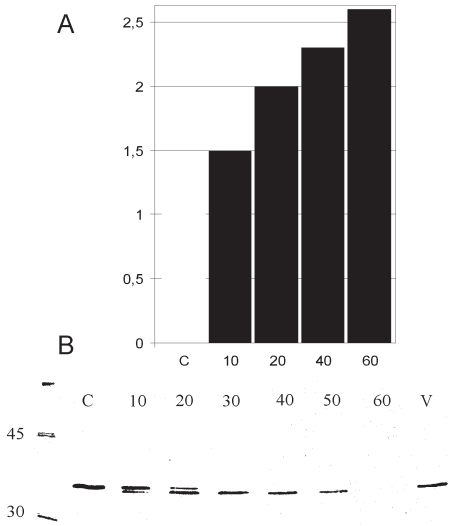


Figure 5. In vitro Phosphatase activity of *O. tauri* Cdc25 in starfish oocyte extracts. After purification of the starfish oocyte inactive CDK-cyclin complex by the use of sepharose beads grafted with $p13^{suc1}$, recombinant *O. tauri* GST::Cdc25 was incubated at RT for different time-lapses. (A) Shows the normalized radioactivity count of the resulting kinase assay of the CDK1cyclin B complex on its specific substrate histone H1. The kinase activity increases in *O. tauri* Cdc25 incubated samples as compared to the control which is incubated in PBS 1X buffer solution. Therefore in vitro *O. tauri* Cdc25 dephosphorylation activates the CDK1-cyclin B complex. (B) Western Blotting using anti-Ptyr antibody shows that there is a band shift followed by a uniform decrease in tyrosine phosphorylation of the starfish CDK1; *O. tauri* Cdc25 is a dual specificity phosphatase which dephosphorylates the Thr residue before the Tyr residue. These results confirm the in vitro dephosphorylating activity of *O. tauri* Cdc25. C: 60' PBS1X; 10: 10' otCdc25; 20: 20' otCdc25; 30: 30' otCdc25; 40: 40' otCdc25; 50: 50' otCdc25; 60: 60' otCdc25.

DISCUSSION

Once a cell is committed to division, the proper succession of the different phases is regulated by checkpoints which block the cell cycle progression until the previous phase is successfully completed (Smits and Medema, 2001). These checkpoints seem to be present in all eukaryotes and the increasing amount of available genomic data from organisms belonging to different phyla shows that most of the proteins involved in these regulations are conserved. In turn, this conservation led to the concept of an ancestral core cell division machinery, at least common to metazoans, fungi and plants (Dewitte and Murray, 2003). Despite this in silico global similarity, the plant cell cycle is nevertheless poorly functionally characterized. Little is known about the plant cell cycle in planta and only an extensive functional analysis will clearly discriminate the core cell cycle machinery from the important specific adaptations in the models. Among these putative specific adaptations was the persistent absence of a cdc25 ortholog in plants although this essential regulation pathway seems to be functionally conserved. This absence is puzzling because orthologs of other regulation partners of the G2/M transition such as CDKs, cyclins and Wee1 are present in plants. This is all the more surprising as most of the proteins known to interact with Cdc25 in metazoans and yeast have also been found in plants. Among these proteins are Pin1 and 14-3-3 and their presence indicates that the DNA damage and DNA replication checkpoints may also be evolutionarily conserved from the ancestral machinery.

The absence of Cdc25 in plants raises the evolutionary question of the presence or absence of this key enzyme at the base of the green lineage and its possible subsequent evolution as a weakly conserved sequence or its loss and substitution by a yet uncharacterised dual specificity phosphatase which mimics the activity of Cdc25. The ongoing sequencing project of the genome of the unicellular alga *Ostreococcus tauri* revealed for the first time the presence of a cdc25 gene at the base of the green lineage. We have shown that this gene encodes a functional Cdc25 protein. *O. tauri* Cdc25 phosphatase is active in the heterologous models starfish and yeast, thereby confirming its functional conservation across evolution, as can be observed with Cdc25 phosphatases from other organisms (Gustafson et al, 2001). By using the *O. tauri* cdc25

sequence, a putative candidate was found in *C. reinhardtii*, whose genome sequencing has recently been completed. This Cdc25 candidate contains the conserved HCX_5R and the essential aspartic acid of the DCR acceptor motif but does not have neither the consensus Cdc25 motif CE[Y-F]SXXR, nor the cysteine of the DCR motif which is necessary for reversing deleterious oxidation (Sohn and Rudolph, 2003). We, therefore, propose to annotate this *C. reinhardtii* sequence as a cdc25-like gene based on its partial sequence similarity to the *O. tauri* cdc25 gene. We further looked for orthologs of this new *C. reinhardtii* cdc25-like gene and found putative orthologs in the terrestrial plants *A. thaliana* and *O. sativa*. However, these cdc25-like sequences are highly divergent from the known cdc25 family and their identification as a putative Cdc25 phosphatase remains to be functionally confirmed. Further characterisation of such Cdc25 phosphatase orthologs in streptophyta will require a comparative approach including evolutionarily intermediate organisms such as the charophyta, but genomic data for these organisms are not yet available.

In conclusion, the conservation of the actors of the cell cycle regulation pathway suggests that the interacting partners must have coevolved so that any divergent change in one partner is adapted by the other (Goh et al, 2000). However, to confirm this hypothesis orthologs of all the genes present early during evolution must be found in the different lineages having conserved this regulatory pathway (Pellegrini et al, 1999). The absence of cdc25 in plants, in contrast to CDKs, cyclins and wee1, is contradictory to the hypothesis of coevolution and would imply that cdc25 phosphatase activity is a more recent evolutionary acquisition, occurring only in ophistokonts (metazoans and fungi) after their separation from the green lineage. But the formal identification of a functional Cdc25 ortholog at the base of the green lineage strengthens the hypothesis of the coevolution of the regulatory pathway of the M-phase entry and rules out the hypothesis of the cdc25 gene as being a recent evolutionary acquisition specific to the ophistokont lineage. It is now clear that Cdc25 phosphatase appeared earlier in evolution before the divergence between the ophistokont (metazoan and yeast) and the green lineages and was probably present in the ancestral eukaryotic cell.

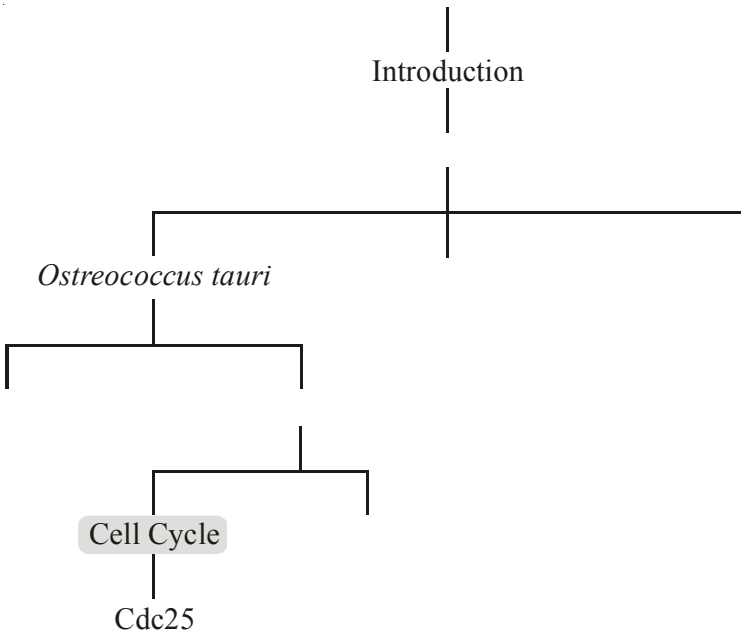
NOTE

Supplemental data can be found online at
http://www.landesbioscience.com/journals/cc/khadarooCC3-4-sup.pdf

ACKNOWLEDGEMENTS

We thank J.C. Lozano for technical assistance and discussions on RACE PCR, RT-PCR and cloning experiments, Sir P. Nurse, G. Lenaers and B. Ducommun for the *S. pombe* Cdc25-22 strain and the pRep vectors. We would also like to thank J. Raes for his help in phylogenetics, and S. Rombauts and S. De Bodt for technical advice. This work was supported by the Génopole Languedoc-Roussillon, the CNRS, the Pierre & Marie Curie pre-doctoral fellowship and the Ph.D. scholarship from the French Embassy in Mauritius.

Chapter 3



Mol. Biol. Evol. 22(3), 589-597, 2004*

Genome-Wide Analysis of Core Cell Cycle Genes in the Unicellular Green Alga *Ostreococcus tauri*

Steven Robbens[1,2*], Basheer Khadaroo[3*], Alain Camasses[3], Evelyne Derelle[3], Conchita Ferraz[4], Dirk Inzé[1,2], Yves Van de Peer[1,2] and Hervé Moreau[3,#]

[1] Department of Plant Systems Biology, VIB, B-9052, Ghent, Belgium
[2] Department of Molecular Genetics, Ghent University, B-9052, Ghent, Belgium
[3] Université Paris VI, Laboratoire Arago, Modèles en Biologie Cellulaire et Evolutive, Banyuls sur Mer, France
[4] Institut de Génétique Humaine, Montpellier, France

[*] Steven Robbens and Basheer Khadaroo have participated equally to this work
[#] Correspondence to: h.moreau@obs-banyuls.fr

Key Words: cell division cycle, cyclin-dependant kinase, cyclin, green alga, *Ostreococcus tauri*

* Modified version

Abstract

The cell cycle has been extensively studied in various organisms, and the recent access to an overwhelming amount of genomic data has given birth to a new integrated approach called comparative genomics. Comparing the cell cycle across species shows that its regulation is evolutionarily conserved; the best-known example is the pivotal role of cyclin dependent kinases in all the eukaryotic lineages hitherto investigated. Interestingly, the molecular network associated with the activity of the CDK-cyclin complexes is also evolutionarily conserved, thus, defining a core cell cycle set of genes together with lineage-specific adaptations. In this paper, we describe the core cell cycle genes of *Ostreococcus tauri*, the smallest free-living eukaryotic cell having a minimal cellular organization with a nucleus, a single chloroplast, and only one mitochondrion. This unicellular marine green alga, which has diverged at the base of the green lineage, shows the minimal yet complete set of core cell cycle genes described to date. It has only one homolog of CDKA, CDKB, CDKD, cyclin A, cyclin B, cyclin D*, cyclin H, Cks, Rb, E2F, DP, DEL, Cdc25, and Wee1*. We have also added the APC and SCF E3 ligases to the core cell cycle gene set. We discuss the potential of genome-wide analysis in the identification of divergent orthologs of cell cycle genes in different lineages by mining the genomes of evolutionarily important and strategic organisms.

* Updated version: 2 copies of CycD and Wee1/Myt1 present

INTRODUCTION

All living organisms undergo cell division, of which the regulation is highly conserved throughout evolution (Stals and Inzé, 2001). The eukaryotic cell cycle is regulated at multiple points, and cell division is ensured by cyclin-dependent kinase-cyclin (CDK-cyclin) complexes, heterodimers composed of a CDK subunit that binds a regulatory cyclin subunit. CDK-cyclin complexes are present in all eukaryotic lineages hitherto studied (Joubès et al, 2000). Their activity is, furthermore, controlled by evolutionarily conserved regulatory mechanisms: phosphorylation/dephosphorylation of the CDK subunit, binding of CDK inhibitors (CKI), cytoplasmic sequestration of the cyclin subunit, and specific ubiquitylation targeting of the cyclin subunit and CKI to proteasome-mediated proteolysis (Deshaies and Ferrell, 2001 and Obaya and Sedivy, 2002). Cell cycle control genes have been found in the different lineages investigated, including the animal, yeast, and plant lineages. Even though specific regulatory mechanisms are present in all lineages, some have evolved differently, such as the retinoblastoma (Rb/E2F/DP) pathway, which is present in animals and plants but absent in yeast (Rubin et al, 2000). Comparative studies of the cell cycle among model organisms belonging to different eukaryotic lineages can, thus, provide crucial information in distinguishing between the core cell cycle common to all phyla and lineage-specific adaptations. Most of the comparative analysis on the cell cycle regulation in ophistokonts (metazoans and fungi) has already yielded the identification of several evolutionarily conserved cell cycle control genes. Although many of these genes are also known in higher plants, their precise role is hard to grasp because of the high complexity of the plant model genomes; namely, the presence of multiple copies of key genes such as CDKs and cyclins. For example, genome-wide analysis shows that cell division control might involve nine CDKs (one of CDKA, four of CDKB, three of CDKD, and one of CDKF) and 30 cyclins (10 of cyclin A, nine of cyclin B, 10 of cyclin D, and one of cyclin H) in *Arabidopsis thaliana* (Vandepoele et al, 2002). The function of each copy is very difficult to investigate because their independent roles are blurred: silencing one copy does not necessarily yield the complete phenotype associated with the gene, as part or all of the function of the silenced copy can be rescued by the other copies. Thus, there is a need

for a simpler green lineage- specific model organism that can be used to unravel the cell cycle specificities of this phylum. Furthermore, studies in the major "classical"' model organisms are not sufficient to account for the common features and the particularities of each model. It is, for instance, difficult to determine whether the presence of only one CDK in yeast is a primeval feature inherited from the ancestral eukaryotic cell or a more recent simplification after the separation between the ophistokonts and the green lineage. These questions can only be answered by the study of new model organisms that occupy key phylogenetic positions. Undoubtedly the genome-wide analysis of their cellular functions, such as the core cell cycle genes, will help in the understanding of the complex green lineage-specific adaptations.

Ostreococcus tauri is a marine unicellular green alga of the Prasinophyceae clade that belongs to the Chlorophyta group of the Planta kingdom (Courties et al, 1998). Because Prasinophyceae have diverged early at the base of the Chlorophyta and consequently of the green lineage (Bhattacharya and Medlin, 1998), *O. tauri* holds a key phylogenetic position in the eukaryotic tree of life. It is, therefore, a potentially powerful model to differentiate between the processes that are common to all eukaryotes (i.e., inherited from the "ancestral eukaryotic cell") and specific adaptations that have occurred after the separation of the different lineages. *O. tauri* is the smallest free-living eukaryotic cell described to date (Courties et al, 1994), with a diameter of no more than 1 ìm. Furthermore, *O. tauri* has a minimal cellular organization, with a nucleus, a single chloroplast, and only one mitochondrion (ChrétiennotDinet et al, 1995). It has a nude plasma membrane without scales or flagella, a reduced cytoplasm (Chrétiennot-Dinet et al, 1995), and a small 12.5-MB to 13-Mb genome, which is currently being sequenced (Derelle et al, 2002). The high-throughput sequencing step of its complete genome is now finished (data not shown), and first analyses indicate that most of the genes have high similarities with genes belonging to the higher plant lineage and can, thus, be easily annotated by sequence similarity. Here, we compare the core cell cycle genes of *O. tauri* with those of *A. thaliana* and discuss new features of their evolution.

MATERIALS AND METHODS

Ostreococcus tauri Cultures.

The *O. tauri* culture used in this study is the strain OTTH0595 (Courties et al, 1998). Cultures were grown in Keller medium (Sigma), diluted in 0.2 µm filtered Banyuls bay-sampled sea water (NaCl 38 g/L), at a temperature of 18°C, with a permanent irradiance of 60 µmol quanta/m^2/s, and under mild agitation. Growth was followed by flow cytometer analysis.

Annotation of the *Ostreococcus tauri* Cell Cycle Genes.

All the genes were annotated based on their similarity with other cell cycle genes available in the public databases. These sequences were aligned using Blast (Altschul et al, 1990) against the *O. tauri* database, and the best hits were further manually annotated using Artemis (Rutherford et al, 2000). The mRNA expression of all the genes described in this study has been confirmed either by their presence among the ESTs sequenced from a cDNA library or from Northern blots or RT-PCR. All the sequences reported in this paper have been submitted to GenBank under the following accession numbers: AY675093 (CDKA), AY675094 (CDKB), AY675095 (CDKC), AY675096 (CDKD/CAK), AY675097 (CycA), AY675098 (CycB), AY675099 (CycD), AY675100 (CycH), AY330645 (Cdc25), AY675101 (Wee1), AY675102 (Rb), AY675103 (E2F), AY675104 (Del), AY675105 (Dp), AY675106 (Cks), AY675107 (Cdc20), AY675108 (CDH1/CCS52), AY675109 (Skp1), AY675110 (Apc1), AY675111 (Apc2), AY675112 (Apc5), AY675113 (Apc6/Cdc16), AY675114 (Apc10), AY675115 (Cdc26p), AY675116 (Apc7), AY675 117 (Apc8/Cdc23), AY675118 (Apc11), AY675119 (Apc4), and AY675120 (Apc3).

Phylogenetic Analysis.

Sequences were aligned with CLUSTALW (Thompson et al, 1994). The sequence alignments were manually improved using BIOEDIT (Hall 1999). TREECON (Van de Peer and De Wachter 1997) was used for constructing the neighbor-joining (Saitou and Nei 1987) trees based on Poisson-corrected distances, only taking into account unambiguously aligned positions. Bootstrap analysis with 500 replicates was performed to test the significance of the nodes.

RESULTS AND DISCUSSION

Ostreococcus tauri Genome Status.

The genome of the *O. tauri* strain OTH95 has been sequenced using the random sequencing method completed by an oriented walking strategy (data not shown). Approximately 120,000 reads corresponding to the extremities of 60,000 shotgun clones and 5,500 reads of the extremities of a BAC library have been assembled using the Phred-Phrap package. A total of 1,989 contigs longer than 2 kbp were obtained for an overall sevenfold depth of coverage. At this stage, specific oligonucleotides were designed at the extremities of the biggest contigs and used to specifically sequence the shotgun clones flanking these contigs. All the contigs obtained were physically located on chromosomes by using a direct hybridization approach on pulse field electrophoresis gels. A total of 5,441 nuclear protein-coding genes were identified using the EuGene gene prediction software version 1.64 (Schiex et al, 2001), which includes both intrinsic and extrinsic approaches for better performance. The mitochondria and chloroplast genomes have also been determined.

The nuclear genome size of *O. tauri* has been estimated at approximately 12.6 Mbp by pulsed-field electrophoresis (PFGE), whereas the total size obtained from the contigs is 12.4 Mbp. A total of 1,850 unique *O. tauri* expressed sequence tags (ESTs) have been sequenced, and around 99% of these sequences mapped with identity greater than 95% onto the genome by using Blast. This is further evidence of the completeness of the genome sequence. and this genome draft has then been used for the complete analysis of the *O. tauri* core cell cycle genes.

CDK-Cyclin Complexes.

CDKs are universally conserved cell cycle regulators. Six classes of CDKs have been described in the plant model *A. thaliana* (Vandepoele et al, 2002). CDKA has a PSTAIRE cyclin-binding motif and is the plant ortholog of the universal eukaryotic cell cycle regulator CDK1 (Dorée and Hunt, 2002). CDKB, whose expression is cell cycle regulated, belongs to a plant-specific CDK clade and plays a role at the G2/M-phase transition. CDKD and CDKF are CDK activating kinases (CAK), which activate the CDK by phosphorylating the threonine residue in the T-loop (Jeffrey et al, 1995). CDKC and CDKE are not

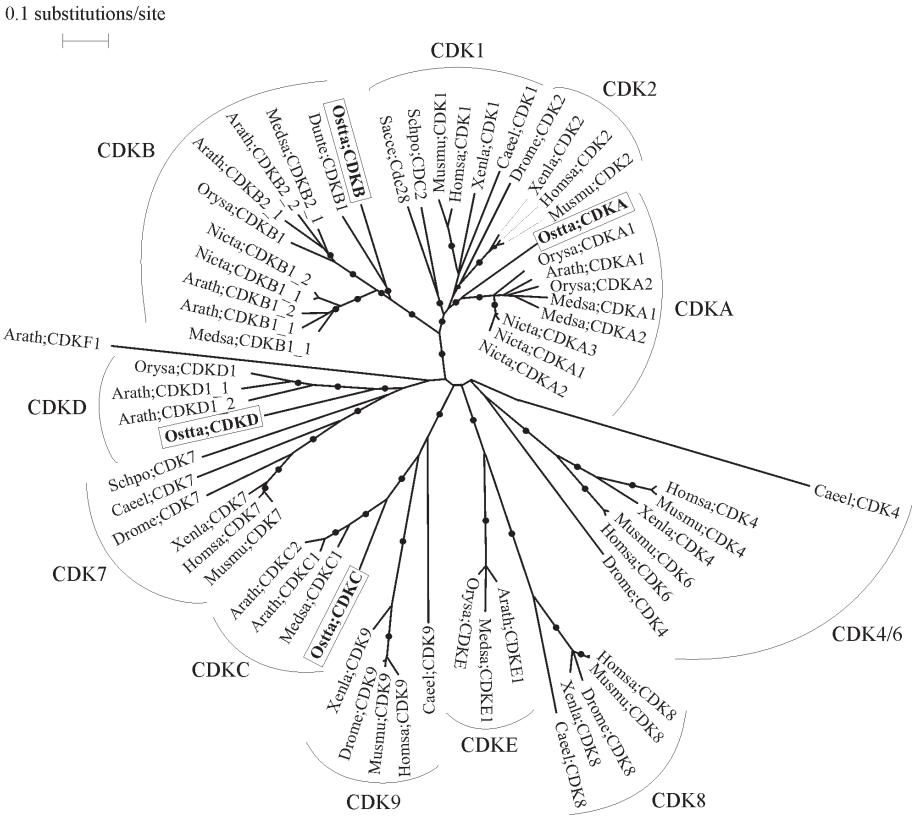


Figure 1. The CDK gene family. Unrooted neighbor-joining tree inferred from Poisson-corrected evolutionary distances for the CDK gene family involved in the cell cycle. The black dots indicate bootstrap values above 70 out of 500 samples. Arath: *Arabidopsis thaliana*; Medsa: *Medicago sativa*; Nicta: *Nicotiana tabacum*; Orysa: *Oryza sativa*; Dunte: *Dunaliella tertiolecta*; Sacce: *Saccharomyces cerevisiae*; Schpo: *Schizosaccharomyces pombe*; Musmu: *Mus musculus*; Homsa: *Homo sapiens*; Xenla: *Xenopus laevis*; Caeel: *Caenorhabditis elegans*; Drome: *Drosophila melanogaster*; Ostta: *Ostreococcus tauri*.

directly involved in the cell cycle control. According to this plant nomenclature (Joubès et al, 2000), four CDKs belonging to the A to D classes have been found in the genome of *O. tauri* (fig. 1 and table 1), but only three (CDKA, CDKB, and CDKD) are involved in cell division control.

Table 1. Comparison of CDK Genes of Metazoans, Yeasts, Plants, and *Ostreococcus tauri*

Genes	*Ostta*	*Aarath*	*Sacce*	*Homsa*	*Ostta*	*Aarath*	*Sacce*	*Homsa*
CDKA	1	1	1	3	PSTAIRE	PSTAIRE	PSTAIRE	PSTAIRE
CDKB	1	4	0	0	PSTALRE	P[S/P]T[A/T]LRE	--	--
CDKC	1	2	1	1	PITAIRE	PITAIRE	PITAIRE	PITALRE
CDKD	1	3	0	1	NFTAIRE	N[I/F/V]TALRE	--	NRTALRE
CDKF	--	1	1	1	--	YQSAFRE	PHNAKFE	PNQALRE

Arath: *Arabidopsis thaliana*; Sacce: *Saccharomyces cerevisiae*; Homsa: *Homo sapiens*; Ostta: *Ostreococcus tauri*.

O. tauri CDKA contains the canonical PSTAIRE motif that is the hallmark of the central cell cycle regulator whose orthologs are Cdc2 in *S. pombe*, Cdc28 in *S. cerevisiae*, CDK1 in vertebrates, and CDKA in plants. This intronless gene is well conserved when compared with the PSTAIRE CDKs of other organisms and shows 66% sequence identity with the CDKA of *A. thaliana*. Moreover, the latter has four copies of the plant-specific B-class CDKs, whereas *O. tauri* has only one copy of a B-class-like CDK. *O. tauri* CDKB is intronless and contains a novel PSTALRE motif that is midway between the PSTAIRE CDKA and the P[S/P]T[A/T]LRE CDKB motifs. However, its overall sequence similarity and phylogenetic position confirms that this *O. tauri* gene is orthologous to higher plant CDKBs and is clearly not a divergent CDKA (fig. 1). Both *O. tauri* CDKA and CDKB have diverged early in their respective clade, confirming the phylogenetic position of *O. tauri* at the base of the green lineage (Courties et al, 1998). Furthermore, *O. tauri* has only one copy of CDKD bearing an NFTAIRE motif, homologous to the CDK-activating kinase (CAK), which positively regulates the activity of CDKA by phosphorylation of the threonine-161 residue. It is orthologous to the three CDKDs of *A. thaliana* (fig. 1 and table 1).

Finally, only one PITAIRE motif *O. tauri* CDKC has been identified, as compared with the two CDKCs described in *A. thaliana* (fig. 1 and table 1). PITAIRE CDKs have been reported to phosphorylate the carboxyl terminal domain (CTD) of the RNA polymerase II, and they do not participate directly in the cell cycle control (Barroco et al, 2003). No E-class or F-class CDKs have been found in the genome of *O. tauri*.

Table 2. Comparison of Green Lineage Cyclin Genes

	Number of Copies		Protein Motifs	
Genes	*O. tauri*	*A. thaliana*	*O. tauri*	*A. thaliana*
CycA	1	10	LxCxE MIEVxEEY MRNILVDW HxKf	No LxCxE
CycB	1	9	MRAILIDW	Hx[R/K]F
CycD	2*	10	No LxCxE	LxCxE except D4 and D6
CycH	1	1	IVRHEAK	MRAFYEAK

* updated

Cyclins are the regulatory binding partners of the CDKs, which confer the timing and substrate specificity to the activity of the CDK-cyclin complexes (Futcher 1996). *O. tauri* has the minimum set of cyclins described to date in

any organism. More importantly, it has only one copy of each of the A-class, B-class, D-class*, and H-class cyclins (Renaudin et al, 1996), thus, presenting even fewer cyclin gene copies than yeasts because *S. cerevisiae* has three G1 cyclins (CLN) and six S/G2/M cyclins (CLB) (fig. 2 and table 2). The activities of the different cyclins in *S. cerevisiae* are redundant. In G1-phase, threshold Cln3 kinase activity is necessary for going through the START point, at which time it switches on the other two G1 cyclins, Cln1 and Cln2. The peak activity of these two cyclins induces the activity of the S-phase specific CDK-cyclin complexes by releasing the Clb5-associated and Clb6-associated kinases from the inhibitory CKI Sic1 (Schwob et al, 1994). The Clb5 and Clb6 kinase activities are necessary for progression from the G1 to the end of the S-phases. Finally, the cyclins Clb1-4 are turned on at the G2/M-phase transition, thus, leading the cell into M-phase (Mendenhall and Hodge 1998).

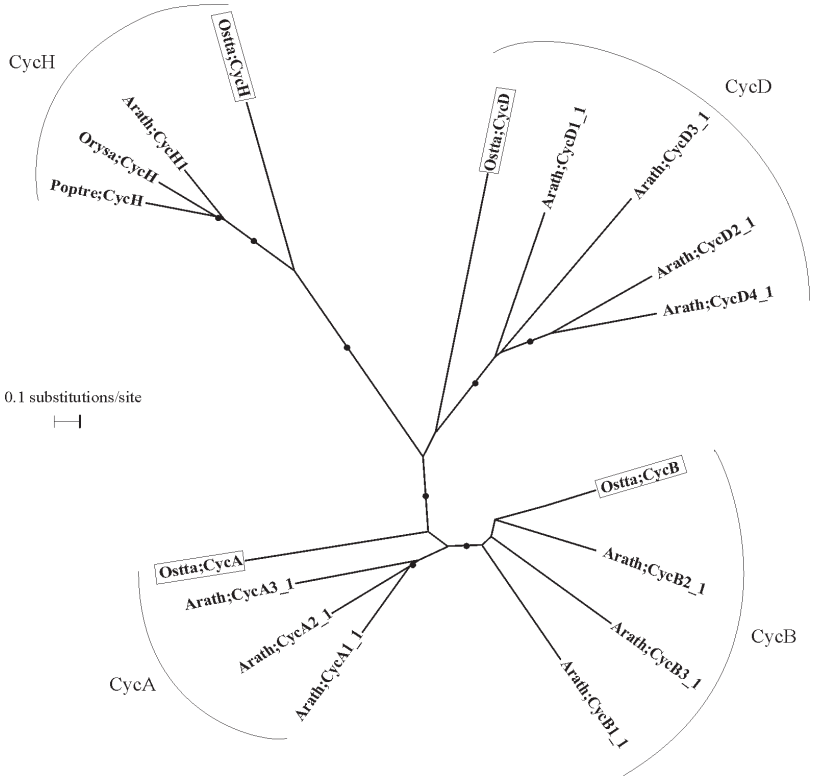


Figure 2. The cyclin gene family. Unrooted neighbor-joining tree inferred from Poisson-corrected evolutionary distances for the cyclin gene family involved in the cell cycle. The black dots indicate bootstrap values above 70 out of 500 samples. Arath: *Arabidopsis thaliana*; Orysa: *Oryza sativa*; Poptre: *Populus tremula*; Ostta: *Ostreococcus tauri*.

* Updated version: 2 copies of CycD present

In *O. tauri*, there is only one putative G1 cyclin, the cyclin D*, which contains one intron (fig 2. and table 2). Surprisingly, the LxCxE retinoblastoma (Rb) binding motif that is normally present on cyclin D of animals and higher plants is not found on the putative *O. tauri* cyclin D but has been identified in the C-terminal part of the *O. tauri* cyclin A. *O. tauri* cyclin A also contains the well-conserved cyclin box motif MRNILVDW and a MIEVAEEY cyclin A-specific motif similar to the *A. thaliana* MRx[I/V]L[I/V]DW and LVEVxEEY cyclin A motifs. This gene has one intron, and it shares the highest homology of 26% sequence identity with the *A. thaliana* cyclin A2 (accession number PIR: D96505). The putative M-phase cyclin B has two introns, and it contains the well-conserved cyclin box motif MRAILVDW and the cyclin B-specific HxKF motif. Hence, this in silico analysis suggests that only one cyclin would be sufficient for each specific phase of the cell cycle, whereas two or more cyclins are present in *S. cerevisiae*, and even more genes are present in the multicellular organisms. Last, the cyclin H, which is the regulatory subunit of the CDKD, does not have an intron, and its sequence analysis yields a cyclin domain similar to homologs of cyclin H from *A. thaliana* with no particular feature (fig 2. and table 2).

Therefore, *O. tauri* presents a minimal, yet complete, set of cell division control genes necessary to drive a eukaryotic cell through the complete division cycle: one of CDKA, one of CDKB, one of cyclin A, one of cyclin B, and one of cyclin D*. Furthermore, the presence in this organism of two CDKs (the universal regulator CDKA and the green lineage-specific CDKB) supports the hypothesis that having one CDK would be a primeval feature that has been conserved in yeast but not a more recent simplification specifically acquired in this lineage. Furthermore, the simplification to only one copy for each cyclin type, in contrast to the usual high copy number of these genes in the other organisms, makes *O. tauri* a potentially powerful model for functional plant core cell cycle studies.

CDK subunits (CKS) are proteins that bind to the CDK protein, and their function is important in the transcriptional activation of Cdc20, the activating protein of the APC complex (see below) (Morris et al, 2003). Only one putative Cks has been found in the genome of *O. tauri* (table 3). This gene has four introns, and the encoded protein has a well-conserved N-terminus sharing an

* Updated version: 2 copies of CycD present

overall 78% and 82% similarity with *A. thaliana* Cks1 and Cks2, respectively. CDK inhibitor (CKI) Kip-related proteins (KRPs) are inhibitors of CDK activities, and seven KRP genes have been found in *A. thaliana* (De Veylder et al, 2001). Despite many efforts, no such genes and no other related CDK inhibitors could be found in *O. tauri* by sequence similarity searches. Only a highly divergent sequence, sharing low sequence similarity to the KRP3 of *A. thaliana* has been found as a putative candidate (table 3). However, KRP genes are usually very divergent; for example, only few key amino acids are conserved between *A. thaliana* and animal inhibitors (De Veylder et al, 2001), and likewise for the CKI between the *S. cerevisiae* Sic1 and *S. pombe* Rum1 inhibitors (Sanchez-Diaz et al, 1998). Furthermore, no sequence conservation was found between the CKI from yeast and other lineages. This absence or very low sequence similarity means that the only possibility of identifying the *O. tauri* cell cycle inhibitors will be by using functional genetic and/or biochemical approaches.

Table 3. Comparison of Several Core Cell Cycle Genes

	Number of Copies			
Genes	*O. tauri*	*A. thaliana*	*S. cerevisiae*	*H. sapiens*
Cdc25	1	1#	1	3
Wee1/Myt1/Mik	2*	1	2	2
Rb/p107/p130	1	1	0	3
E2F	1	3	0	6
DEL	1	3	0	1
DP	1	2	0	2
Cks	1	2	1	2
CKI	1?	7	1	8

Very poorly conserved.
* updated

Retinoblastoma (Rb/E2F/DP) Pathway.

At the G1/S transition, the Rb pathway is conserved among the animal and plant kingdoms but is absent in yeasts. An Rb homolog has also been described in the unicellular green alga *Chlamydomonas reinhardtii* (Umen and Goodenough 2001). However, this alga has a peculiar cell division, and the question of whether this Rb pathway is characteristic for multicellular organisms has remained unclear (Cross and Roberts 2001). In algae, plants, and metazoans, the retinoblastoma pathway induces the expression of S-phase-specific genes. The Rb protein sequesters and inactivates the DNA-bound heterodimer transcription factor comprising the E2F protein and its dimerization partner (DP) protein (Weinberg 1995). It also recruits the chromatin remodeling

machinery for silencing the target genes (Shen 2002) and is responsible for maintenance of quiescence (Sage et al, 2003). At the G1/S transition, the CDK-cyclin complex phosphorylates Rb. Once phosphorylated, the Rb protein frees the E2F/DP complex, which is able to induce the transcription of its target genes. In contrast to vertebrates, which have three copies of pocket proteins

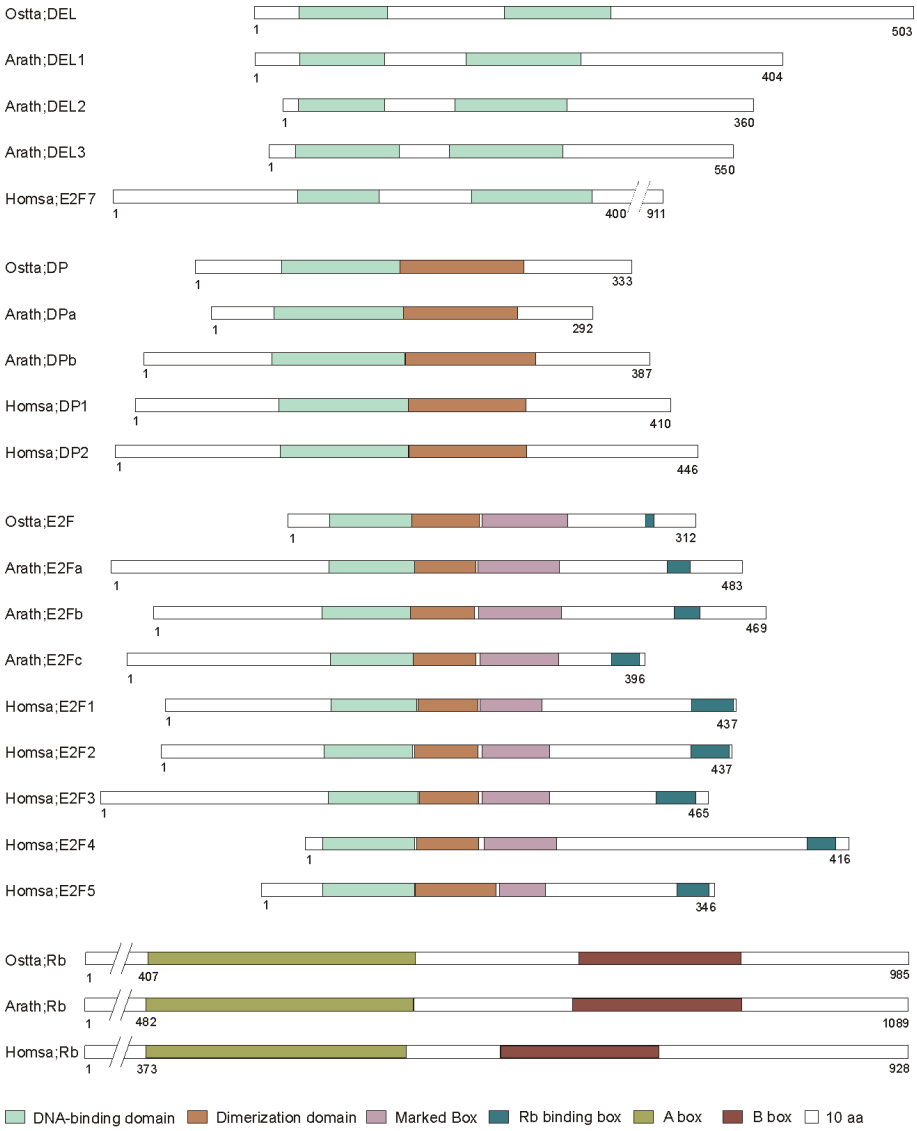


Figure 3. Schematic representation of the E2F family and Rb genes. Structural organization of the DEL, DP, E2F, and Rb proteins of *Ostreococcus tauri* (Ostta) compared with *Arabidopsis thaliana* (Arath) and *Homo sapiens* (Homsa). The DNA-binding, dimerization, Marked, and Rb- binding domains are indicated with colored boxes.

(Rb, p107, and p130), *O. tauri*, similar to *A. thaliana* and *C. reinhardtii*, has only one homolog of Rb gene (fig. 3 and table 3).

The E2F family of transcription factors comprise the subfamilies of activating E2Fs, inhibitory E2Fs, dimerization partner proteins (DPs), and DP-like and E2F-like proteins (DELs) (Shen 2002). The three E2F, two DP, and three DEL genes identified in *A. thaliana* have approximately 22% overall sequence similarity (Vandepoele et al, 2002). E2FA and E2FB have four binding domains, a DNA-binding, a DP-binding, an Rb-binding, and a transactivation-binging domain. When bound to DPA or DPB proteins, they are transcriptional activators that are repressed by Rb protein. E2FC lacks the transactivation-binding domain and is a homolog of animal inhibitory E2Fs E2F4-6. Also, the three *A. thaliana* DEL proteins, of which E2F7 is the recently described animal homolog (Di Stefano, Jensen, and Helin 2003), each have two DNA-binding domains but do not have either DP-binding, Rb-binding, or transactivation-binding domains. Hence, DEL proteins contribute to the class of E2F inhibitory subfamilies. Only one E2F, one DP, and one DEL gene have been identified in the genome of *O. tauri* by alignments of their sequences with their orthologs from *A. thaliana* (figs. 3 and 4). The phylogenetic analysis of the E2F family confirms the early divergence of *O. tauri* genes with respect to the higher plants (fig. 4). The *O. tauri* E2F is a homolog of activating E2Fs and has a conserved binding domain

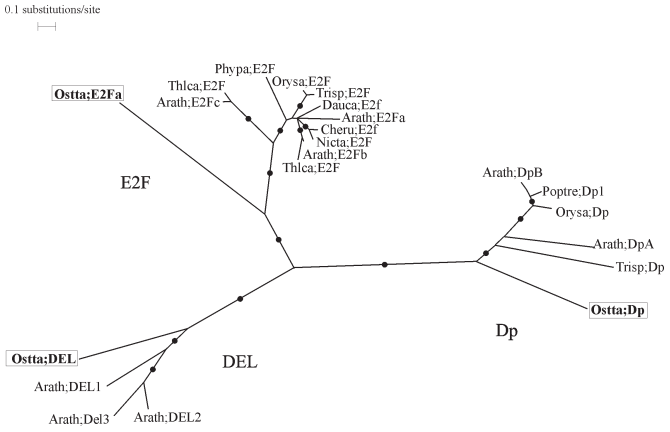


Figure 4. The E2F gene family. Unrooted neighbor-joining tree inferred from Poisson-corrected evolutionary distances for the E2F, Dp, and DEL families. The black dots indicate bootstrap values above 70. Arath: *Arabidopsis thaliana*; Nicta: *Nicotiana tabacum*; Orysa: *Oryza sativa*; Cheru: *Chenopodium rubrum*; Dauca: *Daucus carota*; Thlca: *Thlaspi caerulescens*; Trisp: *Triticum sp*.; Phypa: *Physcomitrella patens*; Poptre: *Populus tremula*; Ostta: *Ostreococcus tauri*.

to the DP-binding and Rb-binding domains and DNA-binding and transactivation-binding domains, whereas DEL has only two DNA-binding domains but no other domain (fig. 3).

Furthermore, the three *O. tauri* genes have diverged very early in each group, as observed for the CDK and cyclin genes. As for the CDKs and cyclins, *O. tauri* has the minimal but complete set of genes for the Rb pathway comprising one Rb gene, one E2F gene, one DP gene, and one DEL gene. Finding the Rb pathway in *O. tauri* confirms that its presence in *C. reinhardtii* is evolutionarily conserved, and the more parsimonious explanation is that Rb was present in the ancestral eukaryotic cell and has been subsequently lost in the yeast phylum. Recently, an unrelated gene called Whi5, which substitutes the role of Rb in yeast at the G1/S transition by inhibiting the transactivation activity of the SBF and MBF transcription factors, reinforces the hypothesis of the loss of the Rb gene family (Costanzo et al, 2004).

Cdc25/Wee1 Control of CDK-Cyclin Activity.

The kinase Wee1 and the phosphatase Cdc25 regulate the G2/M transition in metazoans and yeasts by posttranslational regulation of the CDK-cyclin complex. In plants, an ortholog of the wee1 gene has been found, but the cdc25 gene has never been identified either in the fully sequenced genomes of higher plants such as *A. thaliana* and rice or in that of the unicellular green alga *C. reinhardtii*. The absence of the M-phase inducer Cdc25 phosphatase in plants is puzzling because the other actors of this regulation pathway, namely the CDK-cyclin B complex and the Wee1 kinase, are evolutionarily conserved. Furthermore, the activating dephosphorylation at the M-phase entry is conserved in plants (Zhang, Letham, and John 1996; McKibbin, Halford, and Franci 1998). An intronless putative wee1* gene has been identified in the genome of *O. tauri* (table 3). It is similar to the Wee1 kinase of *A. thaliana*, maize, and *C. reinhardtii* and potentially inhibits the activity of the CDK-cyclin complex by its inhibitory phosphorylation. Surprisingly, the ortholog of the activating phosphatase cdc25 gene has been identified in *O. tauri* (table 3). It is the first time that cdc25 is described in the green lineage, and it shows that the cdc25 gene is present at the base of this lineage. The *O. tauri* cdc25 gene codes for a protein that is able to rescue the *S. pombe* cdc25-22 conditional mutant.

* Updated version: Wee1 and Myt1 present

Furthermore, microinjected *O. tauri* Cdc25 specifically activates starfish oocytes arrested in the prophase of the first meiotic division, thus, causing germinal vesicle breakdown. In vitro phosphatase assays, namely antiphosphotyrosine Western blotting and the histone H1 kinase assay confirmed the in vivo activity of *O. tauri* Cdc25 (Khadaroo et al, 2004) (chapter 2).
The presence of the first functional green-lineage Cdc25 dual-specificity phosphatase discovered in *O. tauri* indicates that this gene was probably present in the ancestor of the eukaryotic cell. In this respect, it should be noted that a putative Cdc25-like gene has also been identified in the complete genome of *C. reinhardtii*. Furthermore, because its activity is necessary in higher plants, the most parsimonious hypothesis is that the sequence of Cdc25 in higher plants has diverged so much that it can no longer be recognized by sequence homology analysis (Khadaroo et al, 2004). This has been very recently confirmed by the identification in *A. thaliana* of a poorly conserved Cdc25-related protein having a tyrosine-phosphatase activity stimulating the kinase activity of *A. thaliana* CDKs (Landrieu et al, 2004).

Ubiquitin Ligases APC and SCF.

Both the anaphase promoting complex (APC) and Skp1/Cdc53/F box protein (SCF) complex, which are the key enzymes for tagging proteins in the ubiquitin pathway, are the E3 ligases responsible for the specificity of protein degradation by the 26S proteasome (Irniger 2002; Cope and Deshaies 2003). The two evolutionarily conserved functions of the APC is the cell cycle-specific targeting of securin and the mitotic cyclin B for proteasome-mediated destruction. The APC is composed of at least 13 protein subunits initially identified in yeast and animals (Schwickart et al, 2004). All vertebrate APC subunits have their homologs in plants (Capron, Okresz, and Genschik 2003; Tarayre et al, 2004). Interestingly, although many genes (and notably cell cycle genes [see above]) are present as multiple copies in *A. thaliana*, its APC genes are present as single copies, except for APC3, which is represented by two slightly different genes (table 4, modified from Capron, Okresz, and Genschik [2003]). In *O. tauri*, putative orthologs of most of these genes have also been found as single copies, including only one copy of APC3. They show very high similarity scores with the *A. thaliana* APC genes. Furthermore, putative orthologs of the two

activators (Cdc20 and Cdh1) regulating the activity of the APC complex have also been found in the *O. tauri* genome. In contrast with the many putative orthologs of Cdc20 and Cdh1 in *A. thaliana* (Capron, Okresz, and Genschik 2003), there seems to be only one Cdc20 and one Cdh1 gene in *O. tauri* (table 4). Thus, *O. tauri* has a complete set of APC genes, with a minimal number of activators.

The Skp1/Cullin/F (SCF) box protein is the other evolutionarily conserved E3 ligase that is responsible for cell cycle control; namely, targeting the CKI for proteasome-mediated proteolysis, which is essential for the cell to progress in late G1-phase (Deshaies and Ferrell 2001). The annotation of SCF genes of *O. tauri* has shown only one Skp1 gene and four cullin-domain proteins, of which two are putative Cdc53 genes and one is the Skp2 putative gene. Skp2 has been identified by using the *A. thaliana* Skp2 gene (Del Pozo, Boniotti, and Gutierez 2002), which contained both an F-box and a leucine-rich domain. These putative genes need to be functionally assayed to confirm this annotation.

Once more, *O. tauri* presents a conserved set of APC and SCF genes with a lower copy number of genes than in *A. thaliana*.

Table 4. Evolutionary Conservation of the Subunits of the APC E3 ligase

	Number of Copies				
APC subunits	*O. tauri*	*A. thaliana*	*S. cerevisiae*	*H. sapiens*	Protein Motifs
APC1	1	1	1	1	Rpn 1 and 2 proteasome repeats
APC2	1	1	1	1	Cullin domain
APC3	1	2	1	1	TPR repeats
APC4	1	1	1	1	
APC5	1	1	1	1	
APC6/Cdc16	1	1	1	1	TPR repeats
APC7	1	1	Unclear	1	TPR repeats
APC8/Cdc23	1	1	1	1	TPR repeats
APC9	Unclear	Unclear	1	Unclear	
APC10/Doc1p	1	1	1	1	Doc domains
APC11	1	1	1	1	Ring-H2 domain
Mnd2p	Unclear	1	1	1	
Swm1p	1	1	1	1	
Cdc26p	1	1	1	1	
Activators APC					
Cdc20	1	5	1	1	
CDH1/Ccs52	1	3	1	2	

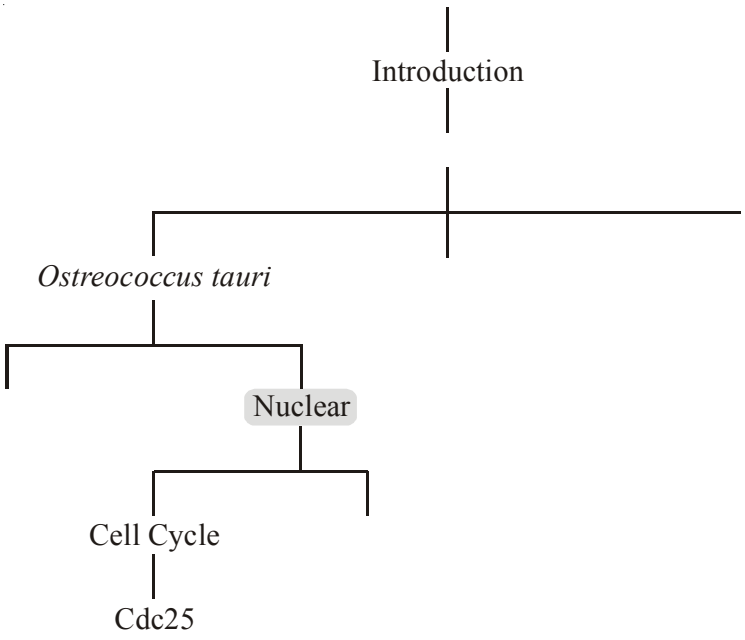
CONCLUSION

The data above reveal that the cell cycle control in the unicellular marine green alga *Ostreococcus tauri* is the simplest described to date (and one of the most complete across the different lineages). It has the minimum set of cyclins for driving the cell cycle and has indeed conserved the Rb pathway, which has been lost in the yeast phylum. It has also retained the plant-specific B-class CDK, and it presents the first green-lineage Cdc25 phosphatase, which has only been identified as a very poorly related gene in higher plants. *O. tauri* displays the minimum, yet complete, set of core cell cycle, and its functional analysis will undoubtedly yield valuable information providing a clear picture not blurred by the activity of other functionally redundant members of the gene family.

ACKNOWLEDGEMENTS

We thank Dr. W. Zacchariae for precious help in the annotation of E3 ligases and K. Vandepoele and Dr. P. Rouzé for their technical help. This work has been supported by the Génopole Languedoc-Roussillon, the CNRS, the Marie Curie predoctoral fellowship number HPMT-CT-2000-00211 for S.R. and the Ph.D. scholarship from the French Embassy in Mauritius for B.K.

Chapter 4


Introduction
Ostreococcus tauri
Nuclear
Cell Cycle
Cdc25

PNAS 103(31), 11647-11652, 2006*

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

Evelyne Derelle[1], Conchita Ferraz[2], Stephane Rombauts[3,4], Pierre Rouzé[3,4,5], Alexandra Z. Worden[6], Steven Robbens[3,4], Frédéric Partensky[7], Sven Degroeve[3,4,8], Sophie Echeynié[2], Richard Cooke[9], Yvan Saeys[3,4], Jan Wuyts[3,4], Kamel Jabbari[10], Chris Bowler[11], Olivier Panaud[9], Benoît Piégu[9], Steven G. Ball[11], Jean-Philippe Ral[11], François-Yves Bouget[1], Gwenael Piganeau[1], Bernard De Baets[8], André Picard[1,9], Michel Delseny[9], Jacques Demaille[2], Yves Van de Peer[3,4] and Hervé Moreau[1]

[1] Observatoire Océanologique, Laboratoire Arago, France
[2] Institut de Génétique Humaine, France
[3] Department of Plant Systems Biology, VIB, B-9052, Ghent, Belgium
[4] Department of Molecular Genetics, Ghent University, B-9052, Ghent, Belgium
[5] Végétaux Marins et Biomolécules, France
[6] Rosenstiel School of Marine and Atmospheric Science, USA
[7] Centre National de la Recherche Scientifique, France
[8] Department of Applied Mathematics, Belgium
[9] Génome et Développement des Plantes, France
[10] Département de Biologie, France
[11] Laboratoire de Chimie Biologique, France

[#] Correspondence to: yves.vandepeer@psb.ugent.be

Key Words: genome heterogeneity, genome sequence, green alga, gene prediction, Prasinophyceae

Author contribution see appendix page 227

* Modified version

Abstract

The green lineage is reportedly 1,500 million years old, evolving shortly after the endosymbiosis event that gave rise to early photosynthetic eukaryotes. In this study, we unveil the complete genome sequence of an ancient member of this lineage, the unicellular green alga *Ostreococcus tauri* (Prasinophyceae). This cosmopolitan marine primary producer is the world's smallest free-living eukaryote known to date. Features likely reflecting optimization of environmentally relevant pathways, including resource acquisition, unusual photosynthesis apparatus, and genes potentially involved in C_4 photosynthesis, were observed, as was downsizing of many gene families. Overall, the 12.56-Mb nuclear genome has an extremely high gene density, in part because of extensive reduction of intergenic regions and other forms of compaction such as gene fusion. However, the genome is structurally complex. It exhibits previously unobserved levels of heterogeneity for a eukaryote. Two chromosomes differ structurally from the other eighteen. Both have a significantly biased G+C content, and, remarkably, they contain the majority of transposable elements. Many chromosome 2 genes also have unique codon usage and splicing, but phylogenetic analysis and composition do not support alien gene origin. In contrast, most chromosome 19 genes show no similarity to green lineage genes and a large number of them are specialized in cell surface processes. Taken together, the complete genome sequence, unusual features, and downsized gene families, make *O. tauri* an ideal model system for research on eukaryotic genome evolution, including chromosome specialization and green lineage ancestry.

INTRODUCTION

The smallest free-living eukaryote known so far is *Ostreococcus tauri* (Courties et al, 1994). This tiny unicellular green alga belongs to the Prasinophyceae, one of the most ancient groups (Courties et al, 1998) within the lineage giving rise to the green plants currently dominating terrestrial photosynthesis (the green lineage) (Baldauf, 2003 and Yoon et al, 2004). Consequently, since its discovery, there has been great interest in *O. tauri*, which, because of its apparent overall simplicity, a naked, nonflagellated cell possessing a single mitochondrion and chloroplast, in addition to its small size and ease in culturing, renders it an excellent model organism (Chrétiennot-Dinet et al, 1995). Furthermore, it has been hypothesized, based on its small cellular and genome sizes (Courties et al, 1998 and Derelle et al, 2002), that it may reveal the "bare limits" of life as a free-living photosynthetic eukaryote, presumably having disposed of redundancies and presenting a simple organization and very little noncoding sequence.

Since its identification in 1994, *Ostreococcus* has been recognized as a common member of the natural marine phytoplankton assemblage. It is cosmopolitan in distribution, having been found from coastal to oligotrophic waters, including the English Channel, the Mediterranean and Sargasso Seas, and the North Atlantic, Indian, and Pacific Oceans (Diéz et al, 2001; Guillou et al, 2004; Worden et al, 2004; Zhu et al, 2005 Countway and Caron, 2006; and Worden, 2006). Eukaryotes within the picosize fraction (2- to 3 μm diameter) have been shown to contribute significantly to marine primary production (Worden et al, 2004 and Li, 1994). *Ostreococcus* itself is notable for its rapid growth rates and potential grazer susceptibility (Worden, 2004 and Fouilland et al, 2004). Furthermore, dramatic blooms of this organism have been recorded off the coasts of Long Island (O'Kelly et al, 2003) and California (Countway and Caron, 2006). At the same time, attention has focused on the tremendous diversity of picoeukaryotes (López-García et al, 2001 and Moon-van der Staay et al, 2001), which holds true for *Ostreococcus* as well. Recently, *Ostreococcus* strains isolated from surface waters were shown to represent genetically and physiologically distinct ecotypes, with light-regulated growth optima different from those isolated from the deep chlorophyll maximum (Rodríguez et al, 2005). These findings are similar to the niche adaptations documented in different ecotypes of the abundant

marine cyanobacteria Prochlorococcus (Moore et al, 1998 and Rocap et al, 2003).

Overall, marine picophytoplankton play a significant role in primary productivity and food webs, especially in oligotrophic environments where they account for up to 90% of the autotrophic biomass (Worden et al, 2004; Li, 1994; Campbell et al, 1994; and Rocap et al, 2002). Several recent studies have undertaken a genome sequencing approach to understand the ocean ecology of phytoplankton. To date, these studies have focused on the bacterial component of the plankton, particularly on the picocyanobacteria *Prochlorococcus* (Rocap et al, 2003) and *Synechococcus* (Palenik et al, 2003), for which 9 complete genome sequences are already publicly available and 13 others on the way. Much less is known about eukaryotic phytoplankton, because only one, the diatom *Thalassiosira pseudonana*, has a complete genome sequenced (Armbrust et al, 2004). Picoeukaryotes are especially interesting in the context of marine primary production, given the combination of their broad environmental distribution and the fact that their surface area to volume ratio, a critical factor in resource acquisition and success in oligotrophic environments (Raven and Kübler, 2002), is similar to that of prokaryotic counterparts generally considered superior in uptake and transport of nutrients.

In this article, we describe the complete genome sequence of *O. tauri* OTH95, a strain isolated in the Thau lagoon (France) in which this species makes recurrent, quasimonospecific blooms in summer (Courties et al, 1994). This genome is particularly significant in that it represents a complete genome sequence of a member of the Prasinophyceae, which diverged at the base of the green lineage (Courties et al, 1998). It is also the complete genome sequence of a picoeukaryote thought to be of ecological importance to primary production. Analysis of the *O. tauri* genome and comparison with other genomes available to date, including algal, plant, and fungal genomes, allowed delineation of both specific gene features and identification of unique aspects of this genome.

RESULTS AND DISCUSSIONS

Global Genome Structure.

Whole genome shotgun sequencing and an oriented walking strategy were used to sequence the genome of *O. tauri* strain OTH95 (tables 2 and 3, which are published as supporting information on the PNAS web site). A genome size of 12.56 Mb distributed in 20 superscaffolds corresponding to 20 chromosomes was determined by means of sequence assembly (fig. 1; and figs. 4 and 5, which are published as supporting information on the PNAS web site), fully consistent with pulsed-field gel electrophoresis results indicating a total size of 12.5 to 13 Mb (fig. 4 and Supporting Text, which are published as supporting information on the PNAS web site). This genome size is similar to that of the yeasts *Saccharomyces cerevisiae* and *Schizosaccharomyces pombe*, despite their larger cell size, but smaller than any other oxyphototrophic eukaryote known so far, including the red alga *Cyanidioschyzon merolae* (Matsuzaki et al, 2004) (fig. 2 and table 1). The G+C content of *O. tauri* is more akin to that of *C. merolae* than to that of plants, fungi, or even *T. pseudonana* (table 1). As shown in Fig. 2 and Table 1, 8,166 protein-coding genes were predicted in the nuclear genome, making *O. tauri* the most gene dense free-living eukaryote known to date. Only the chromosomes of the nucleomorphs within chlorachniophyte and cryptophyte algae are more gene-dense bodies (Gilson, 2001), which are internally contained and not capable of independent propagation.

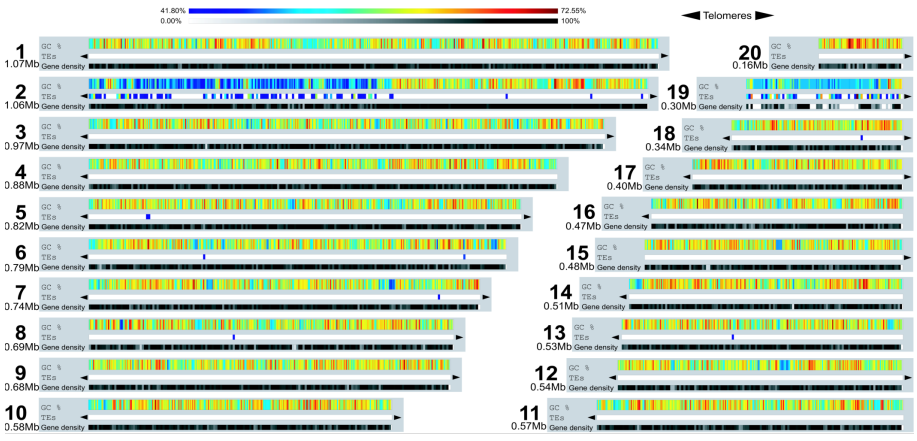


Figure 1. General characteristics of the 20 *O. tauri* chromosomes. TEs, transposon frequency. Size is indicated to the left of each chromosome (Mb). Colored bars indicate the percentage G+C content (upper bar) and of transposons (lower bar).

We found that 6,265 genes are supported by homology with known genes in public databases (e-value 10^{-5}), of which the majority (46%) were most similar to plant orthologs (fig. 3). Very few repeated sequences have been found in this genome, except for a long internal duplication of 146,028 kb on chromosome 19. Because the duplicated sequence is 99% identical, it is probably of recent origin.

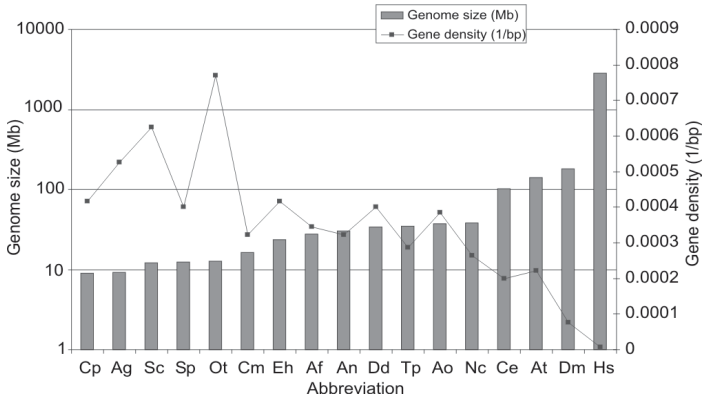


Figure 2. Genome size and gene density for various eukaryote genomes. Cp, *Cryptosporidium parvum*; Ag, *Ashbya gossypii*, Sp, *Schizosaccharomyces pombe*; Sc, *Saccharomyces cerevisiae*; Ot, *Ostreococcus tauri*; Cm, *Cyanydioschyzon merolae*; Eh, *Entamoeba hemolytica*; Af, *Aspergillus fumigatu*;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

Genome Heterogeneity.

In view of what is currently known about eukaryotic nuclear genomes, one of the most striking features of the *O. tauri* genome is its heterogeneity, a feature which is not only unusual but also perplexing from an evolutionary perspective. Two chromosomes (2 and 19) are different from the other 18, in terms of organization for chromosome 2 and function for chromosome 19 (fig. 1; and fig. 6, which is published as supporting information on the PNAS web site). Both of these chromosomes have lower G+C content than the 59% G+C of the other 18 chromosomes (fig. 1). Chromosome 2 is composed primarily of two blocks, one with a G+C content similar to that of the other chromosomes and the other with a markedly lower G+C content (52%). The average G+C content of the entire chromosome 2 amounts to 55%. Likewise, the G+C content of chromosome 19 (54%) is similar to the atypical region of chromosome 2. Taken together, these two aberrant chromosomes contain 77% of the 417 transposable

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

Feature	*Ostta*	*Thaps*	*Cyame*	*Arath*	*Ashgo*	*Sacce*	*Schpo*	*Crypa*
Size, Mbp	12.56	34.50	16.52	140.12	9.20	12.07	12.46	9.10
No. of chromosomes	20	24	20	5	7	16	3	8
G+C content, %	58.0 (59.0*)	47	55	36	52	38.3	36	30
Gene number	8,166	11,242	5,331	26,207	4,718	6,563	4,824	9,807
Gene density, kb per gene	1.3	3.5	3.1	4.5	1.9	1.6	2.5	2.4
Mean gene size, bp per gene†	1,257	992	1,552	2,232	N.A.	N.A.	1,426	1,795
Mean inter-ORF distance	197	N.A.	1,543	2,213	341	N.A.	952	566
Genes with introns, %	39	N.A.	0.5	79	5	5	43	5
Mean length of introns, bp	103 (187*)	N.A.	248	164	N.A.	N.A.	81	N.A.
Coding sequences, %	81.6	N.A.	44.9	33	79.5	N.A.	57.5	75.3
No. of ribosomal RNA units	4	N.A.	3	700-800	50	100-150	200-400	5

N.A., not available.
* Data that exclude chromosomes 2 and 19.
† Data that exclude introns.

elements (TEs), or relics thereof, which are identified in the genome (57% in chromosome 2 and 20% in chromosome 19) (fig. 1 and table 4, which is published as supporting information on the PNAS web site). Other chromosomes therefore contain very few or no TEs. TEs have a G+C content similar to the rest of the genome and cannot explain the global lower G+C content observed in these two chromosomal regions. Moreover, almost all of the known TE types can be found in the *O. tauri* genome: fifteen class I TE families [i.e., 3 TY1 Copia-like LTR-retrotransposons and 12 terminal-repeat retrotransposons in miniature (TRIMs)], nine transposon families, [4 Mariner-like elements, 2 P instability factors (PIFs), 1 homology and transposition (hAT), 1 foldback, and 1 unclassified (Feschotte and Wessler, 2002)], and three miniature inverted repeat transposable element (MITE) families were identified. Only long interspersed nuclear elements (LINEs), short interspersed nuclear elements (SINEs), and helitrons were not detected. In the case of *O. tauri*, the distribution bias could have two origins: either the species originated from an allopolyploidization event between two donor parents with a genome contrasting for their TE content or there is a strong insertion bias for the TEs on both chromosomes 2 and 19. For most of the TE families, several partial copies or relics can be found throughout the 20 chromosomes, indicating their ancient origin in the genome, therefore not supporting the first hypothesis. Nevertheless, further analyses are needed to conclude on this matter.

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

multiple small introns with specific features (fig. 4 a and b). These two differences make gene modeling more complicated for this region, although at least 61 predicted peptides were supported by ESTs (see table 6, which is published as supporting information on the PNAS web site). Chromosome 2 small introns differ in many respects from the other introns, such as their size (40-65 bp), composition (they are AT rich and richer by 10% than the neighboring exons), and splice sites and branch points that are less conserved than for other introns (fig. 4b).

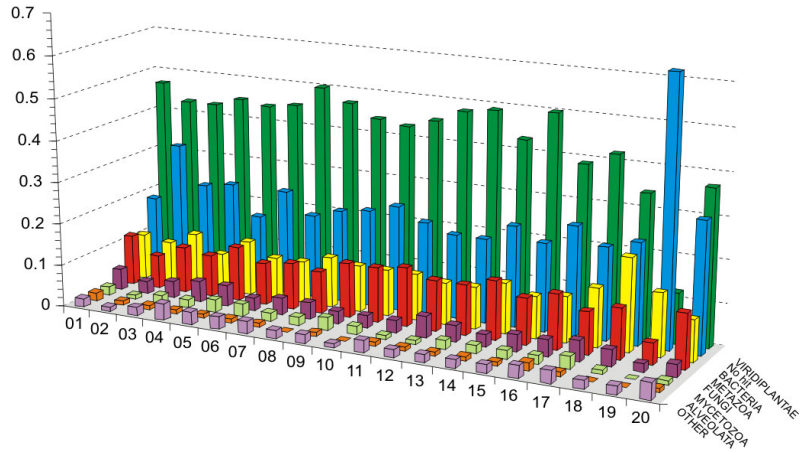


Figure 3. Taxon distribution of best hits for genes from each of the *O. tauri* chromosomes. Green, viridiplantae; blue, no hit; yellow, bacteria; red, metazoa; pink, fungi; gray, mycetozoa; orange, alveolates; purple, others. Annotation of genes on the low G+C part of chromosome 2 is difficult, and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

Interestingly, phylogenetic analysis (see Materials and Methods) shows that 43% of the genes on this chromosome, including the small intron-containing genes, have green lineage ancestry (fig. 3). Of those, 44% cluster specifically (with bootstrap values 70%) with genes of *Chlamydomonas reinhardtii* (data not shown but available on request). Together with the fact that the genes encoded in this region are essential housekeeping genes not duplicated elsewhere in the genome, this observation argues against an alien (horizontal transfer) origin for the low G+C region of chromosome 2. Thus, the origin of the chromosome 2 peculiarities remains elusive. One possibility is that it represents a sexual chromosome. It has been shown before that such chromosomes possess distinctive features for avoiding recombination and are characterized by an

unusual richness in transposable elements (Fraser and Heitman, 2004). Meiosis has not been observed in culture, and no equivalent of a mating-type locus has been found akin to that in *C. reinhardtii*. Nevertheless, the presence of most of the core meiotic genes homologous to those identified in other organisms found in *O. tauri* (table 7, which is published as supporting information on the PNAS web site) is at least a strong indication that *O. tauri* may be a sexual organism (Ramesh et al, 2005). Indeed other marine algae known to undergo sexual reproduction commonly suppress this capability in culture (Chepurnov et al, 2004).

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

Genome Compaction.

A second remarkable feature of the *O. tauri* genome is the intense degree of genome compaction, which appears to be the result of several processes. Shortening of intergenic regions is clearly a major factor. The average intergenic size is only 196 bp, which is shorter than that of other eukaryotes having a similar genome size (table 1). Two other important factors are gene fusion, for which several cases are observed (fig. 8, which is published as supporting information on the PNAS web site), and reduction of the size of gene families. For example, the gene complement involved in cell division control is one of the most complete across eukaryotes, although there is only one copy of each gene (Robbens et al, 2005). Although this type of reduction is often the case in *O. tauri*, there are some exceptions. For example, the full set of partially redundant

enzymes required for polysaccharide metabolism in land plants is present. Here, the maintenance of 27 genes, including multicopy genes, related to synthesis and breakdown of only two types of chemical linkages in the chloroplast, seems excessive for building the semicrystalline starch granule of *O. tauri* (Ral et al, 2004). Indeed, apicomplexa parasites or even red algae require only 10 genes to build and degrade simple polymers in their cytoplasm (Coppin et al, 2005). *O. tauri* appears to be quite similar to other unicellular organisms in terms of numbers of transcription factors, with no further reduction than what has commonly been reported. Approximately 2.5-3.8% of predicted proteins of unicellular organisms fall within the category for transcription factors (table 9, which is

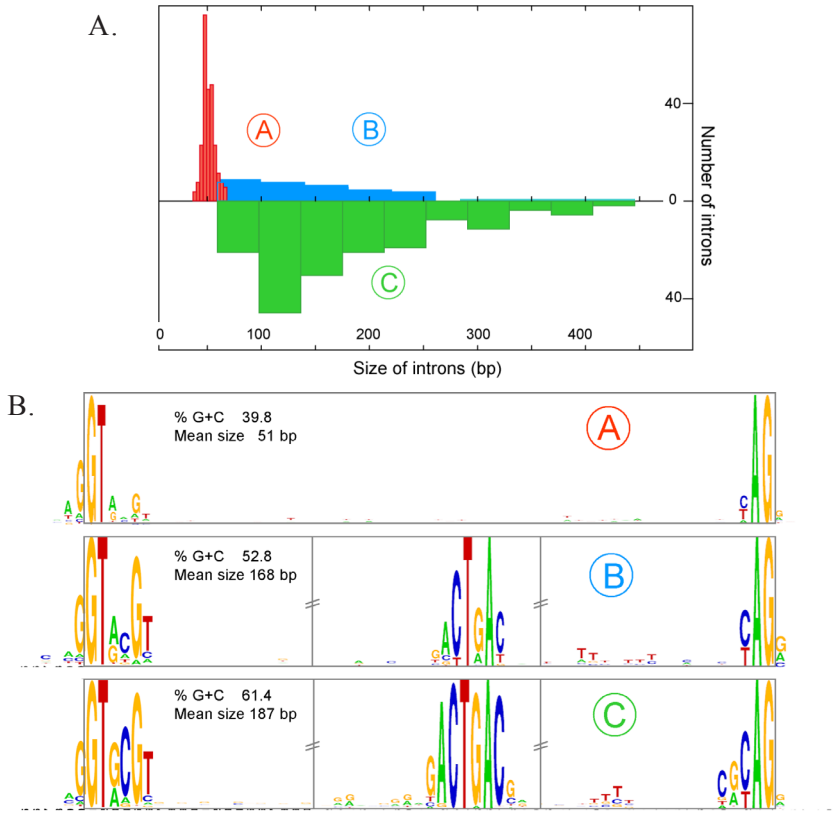


Figure 4. Intron heterogeneity of the Ostreococcus tauri genome. Comparison of chromosome 2 intron structure with that of other chromosomes. (a) Size distribution (bp) of documented small-type (A, red), "normal"-type (B, blue) introns in chromosome 2, and documented introns from other chromosomes (C, green). (b) Intron composition and splicing motifs (donor, acceptor, and branch site) of the three intron types. Different font sizes indicate the probability of a particular nucleotide at the respective motif position. The %G+C and mean size for each of the intron types are also indicated.

published as supporting information on the PNAS web site). This finding is in contrast to multicellular organisms, for which 12-15% of the predicted proteins generally fall within the transcription factor category. With respect to pigment biosynthesis and photosynthesis, many genes involved in these pathways are found in multiple copies in other photosynthetic eukaryotes. In *O. tauri*, they also form multigene families, but the copy number is generally lower (e.g., table 10, which is published as supporting information on the PNAS web site). As expected, *O. tauri* maintains all essential enzymes for carbon fixation, and, based on available data for other algae and land plants, homologs are generally present at half the copy number (Six et al, 2005). Double sets of several carbon metabolism-related genes, including phosphoglycerate kinase, ribulose-bisphosphate carboxylase, and triosephosphate isomerase, can be found in the *O. tauri* genome. Based on both best hit and subsequent phylogenetic analyses these "doubles" each appear to have different origins (bacterial versus eukaryotic).

O. tauri Metabolic Pathways.

O. tauri displays some other characteristics unusual for land plants and algae. For instance, the typical genes encoding the major light-harvestinG+Complex proteins associated with photosystem II (LHCII) are lacking. Instead, paralogs encoding prasinophyte-specific chlorophyll-binding proteins are present, making a special antenna as previously observed in *Mantoniella squamata* (Six et al, 2005). Interestingly, *O. tauri* also possesses a small set of five lhcA genes, encoding an LHCI antenna. Combined with the absence of major LHCII protein-encoding genes, this finding supports the hypothesis that the LHCI antenna type is more ancestral than is LHCII (Six et al, 2005). Unique features are also seen in the carbon assimilation machinery. Only one carbonic anhydrase (CA), most similar to bacterial β-CA, was identified. No carbonconcentrating mechanism (CCM) genes (Giordano et al, 2005) comparable with those of *C. reinhardtii* or common to organisms that actively or passively enhance inorganic carbon influx were found. However, genes putatively encoding all of the enzymes required for C_4 photosynthesis were identified. Whereas C_4 photosynthesis has yet to be unequivocally shown in unicellular organisms (Armbrust et al, 2004; Raven and Kübler, 2002; Reinfelder et al, 2004; and Sage, 2004), C_4 in the

absence of Kranz anatomy is now well documented, especially in *Hydrilla verticillata*, a facultative C_4 aquatic monocot (Rao et al, 2002). Unlike *T. pseudonana*, which appears to lack plastid-localized NADP-dependent malic enzymes (NADP-ME), *O. tauri* has two NADP-ME orthologs most similar to *H. verticillata* (Bowes et al, 2002) with at least one apparently targeted to the chloroplast based on ChloroP and TargetP predictions. *O. tauri* also has phosphoenolpyruvate (PEP) carboxylase, NADP malate dehydrogenase, and pyruvate-orthophosphate dikinase, with predicted chloroplast targeting transit peptides in the latter two. C_4 photosynthesis is thought to have evolved multiple times from C_3 ancestors. Although timing is uncertain, it is currently thought to have first evolved 24-35 million years ago in relation to environmental pressures (e.g., declining atmospheric CO_2) (Giordano et al, 2005 and Sage 2004). Interestingly, only one member of the Chlorophyta, the macroalga *Udotea*, has been shown to perform C_4 photosynthesis. *Udotea* utilizes PEP carboxykinase (PEPCK) (NADP-ME being absent) (Quesada et al, 1993), a C_4 photosynthesis form variant to that suggested here, although not yet confirmed experimentally, for *O. tauri*. Despite its energetic cost, if *O. tauri* is capable of C_4 photosynthesis, it could constitute a critical ecological advantage in the CO_2-limiting conditions of phytoplankton blooms, especially in circumstances where competitors have lower CCM efficiencies (or no CCM at all).

Resource acquisition is critical to survival in the frequently limiting marine environment, and here *O. tauri* seems to have developed competitive strategies currently thought uncommon amongst eukaryotic algae. Nitrogen is typically a major limiting nutrient of marine phytoplankton growth. *O. tauri* is known to grow on nitrate, ammonium, and urea (Worden et al, 2004), and complete sets of genes allowing transport and assimilation of these substrates have been identified (fig. 5; and table 11 which is published as supporting information on the PNAS web site). Interestingly, four genes encoding ammonium transporters were identified, two being green lineage-related and the other two prokaryote-like. Eukaryotic algae are generally considered ineffective competitors for ammonium; however, the high number of ammonium transporters in *O. tauri* (unlike e.g., *T. pseudonana*) indicates it may be a strong competitor for this resource. All other genes related to nitrogen acquisition and assimilation are found in a single copy, including those for nitrate, again in contrast to *T.*

pseudonana. It is notable that eight of the genes involved in nitrate uptake and assimilation are found next to each other on chromosome 10 (fig. 5A), as well as four genes for urea assimilation genes on chromosome 15 (fig. 5B). A comparable clustering of nitrate assimilation genes was also observed in *C. reinhardtii* (Quesada et al, 1993) but grouping fewer genes. This organization is reminiscent of prokaryotes, especially cyanobacteria (Rocap et al, 2003), and indicates a possible selective pressure for optimization of nitrate and urea uptake and assimilation, although experimental evidence for the regulation of expression of these genes is currently lacking. The nitrite reductase (NIR) apoenzyme has a unique structure, with two additional redox domains at the C terminus of canonical ferredoxin-NIR, rubredoxin and cytochrome b5 reductase. This structure should allow this enzyme to use NAD(P)H directly as reducing substrate, which may also contribute to optimization of the pathway. Within this cluster, Snt encodes a protein with weak similarity to sulfate transporters. Nonetheless, its specific position in the cluster suggests that Snt probably encodes a molybdate transporter, a gene predicted to exist but so far unidentified in any species. Taken together with the possibility that *O. tauri* may be capable of C_4 photosynthesis and the relatively high surface area to volume ratio of this tiny phytoplankter, these various ways to optimize nitrogen assimilation could yield a major competitive advantage over other unicellular phytoplankton. This adaptation would be particularly important to its relative success under environmental scenarios, such as intense bloom conditions, where limitation of multiple resources can be encountered.

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

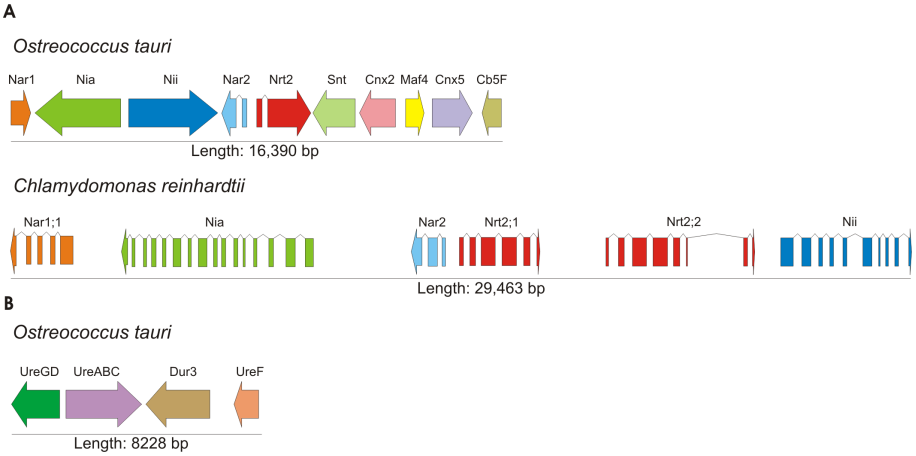


Figure 5. (A) Comparison of nitrate assimilation clusters in Ostreococcus tauri (chromosome 10) and Chlamydomonas reinhardtii (V3.Scaffold 30). Nar, plastid nitrite transporter; Nia, nitrate reductase apoenzyme. [The functional NIA protein reduces nitrate to nitrite using NAD(P)H through the contribution of three redox cofactors: FAD, Heme, and MoCo (molybdenum cofactor).] Nii, plastid-targeted nitrite reductase apoenzyme. (The functional NII reduces nitrite to ammonium using ferredoxin through of a siroheme-iron sulfur cofactor.) In Ostreococcus, NII comprises two additional redox domains, rubredoxin-like and its corresponding reductase, suggested to allow reduction of nitrite directly from NAD(P)H); Nar2, nitrate high-affinity transporter accessory protein; Ntr2, nitrate high-affinity transporter; Snt, putative molybdate transporter; Cnx2, molybdenum cofactor biosynthesis protein (CNX2 performs the first step of MoCo synthesis together with CNX3, forming molybdopterin precursor Z); Maf4, Maf4-related hypothetical protein (although Maf4 is a MADS-box protein in higher plants, OtMaf4 has not the features of a MADs-box protein); Cnx5, molybdenum cofactor biosynthesis protein (molybdopterin synthase sulfurylase); Cb5f, Cytochrome b5 reductase (closer to nitrate reductase FAD/heme reductase domain than to stand alone cytochrome b5 reductase). (B) Urea assimilation cluster in O. tauri (chromosome 15). UreABC, Ni-dependant urease apoenzyme. The A, B, and C subunits are encoded by three separate genes in bacteria but form a single gene here, as in higher plants. UreGD, urease accessory proteins G and D are fused together, whereas, in other organisms, they are encoded by two separate genes; UreF, urease accessory protein F, which forms a complex with G and D and with apourease to allow nickel insertion, resulting in activation of urease; Dur3, urea high-affinity symporter.

In conclusion, the genome structure of *O. tauri* generally follows predictions of compaction and streamlining that might be driven by its specific lifestyle and ecology. However, the heterogeneity we reveal here concerning two chromosomes raises the challenge of elucidating its origin, which could either be a reminiscence of this alga's ancient nature or on the contrary more recent adaptations to its environmental niche. It also raises the question of whether this type of heterogeneity is in fact not unique to *O. tauri*, but rather a common

feature of some eukaryotes, given that current understanding of eukaryotic genomes relies on a genome database so far dominated by "higher organisms''. Understanding features specific to success in the marine environment as well as of evolutionary processes within the green lineage relies on new hypotheses and further experimentation for which this complete genome sequence provides a powerful resource. The exceptional features unveiled in the genome of this ubiquitous, ancient, autonomous unicell high-light the fundamental level at which we might reconsider current paradigms.

MATERIALS AND METHODS

BAC Library.

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

Shotgun Libraries.

Purified DNA was broken by sonication, and, after filling ends, DNA fragments ranging from 1 to 5 kb were separated in an agarose gel. Blunt-end fragments were inserted into pBluescript II KS (Stratagene) digested with EcoRV and dephosphorylated. About 60,000 clones were isolated from four independent *O. tauri* shotgun libraries. Plasmid DNA from recombinant *Escherichia coli* strains was extracted according to the TempliPhi method (GE Healthcare), and

inserts were sequenced on both strands by using universal forward and reverse M13 primers and the ET DYEnamic terminator kit (GE Healthcare). Sequences were obtained with MegaBace 1,000 automated sequencers (GE Healthcare). Data were analyzed, and contigs were assembled by using Phred-Phrap (Ewing et al, 1998) and Consed software packages. Gaps were filled through primer-directed sequencing by using custom made primers.

cDNA Library.

Two cDNA libraries were generated from cultures grown under different conditions to improve the representation of the expressed sequences. Exponentially growing cells sampled at various stages of the cell cycle of cultures synchronized by light/dark cycles were mixed with a stationary stage culture. Poly(A) mRNAs from the different cultures were isolated and then mixed together. One cDNA library was created in the λ ZAP vector (Stratagene) and the second in the Gateway system according to the manufacturer's instructions (Invitrogen). The average insert size analyzed on agarose gels was 1.5 kb for both libraries. Sequences were obtained by using the forward primer, and single reads were assembled in contigs by using Phred-Phrap (Ewing et al, 1998).

Genome Annotation.

The genomic sequence of *O. tauri* was annotated by using the EuGène (Schiex et al, 2001) gene finding system with SpliceMachine (Degroeve et al, 2005) signal sensor components trained specifically on *O. tauri* datasets. A set of 152 GT donor and 152 AG acceptor sites was constructed to optimize the SpliceMachine context representations and to train the splice site sensors that were used to recognize *O. tauri* splice sites. We found GT donor sites to be highly conserved, which resulted in a highly accurate donor site signal sensor. For acceptor sites, the AG consensus pattern was less conserved, whereas the branch point motif was again highly conserved. SpliceMachine was able to extract this branch point pattern and to use it to recognize AG acceptor sets, again resulting in a highly accurate acceptor site sensor. The content sensor used by EuGène to recognize coding sequences is an interpolated Markov model that was computed from 145 *O. tauri* ORFs and 167 intron sequences (used as background). Training EuGène requires the estimation of scaling parameters

from known *O. tauri* genes within their genomic context. As such, 17 genomic *O. tauri* sequences that each contained abutting genes were constructed and used to train EuGène. Peptides for two deviant chromosomes, numbers 2 and 19, were modeled by using EuGène and SpliceMachine trained specifically on low GC chromosome 2 special genes. A set of 253 GT donor and 253 AG acceptor sites was constructed to optimize the SpliceMachine context representations and to train the sensors used to recognize the splice sites on these two deviant chromosomes. In contrast to the splice sites of the normal *O. tauri* genes, these GT-AG splice sites were less conserved, resulting in less accurate splice site sensors. However, splice site recognition accuracy was boosted by incorporating intron length constraints (introns in these genes are shorter than in so-called normal genes, with lengths typically between 40 and 60 bp, compared with 170-190 bp for the 18 other chromosomes) at the level of gene recognition. The interpolated Markov model used by EuGène to recognize the special coding sequences was computed from 43 *O. tauri* ORFs and 209 intron sequences (used as background). Ten genes within their genomic context were used to optimize the scaling parameters within EuGène. The data sources used to complement the ab initio part of EuGène were composed of *O. tauri* expressed sequence tags (ESTs), proteins, and genomic sequences. ESTs sequenced over the course of the project were aligned on the genome and used as the most reliable source of extrinsic information. For BlastX, the Swissprot protein dataset (v. 42), *C. merolae* proteins (Matsuzaki et al, 2004), publicly available *C. reinhardtii* proteins, and predicted proteins from Sargasso Sea environmental sequences (Venter et al, 2004) were used in a decreasing order of priority to avoid error propagation, because the latter dataset is the least reliable. The functional annotation resulted from the synthesis of InterPro and Gene Ontogeny (GO) assignments based on domain occurrences in the predicted proteins by using the InterPro scripts, BlastP against the clusters of eukaryotic orthologous groups (KOG) database, and a top four of BlastP hits (e-value $<10^{-5}$) against the nonredundant UniProt database. Throughout this process, genes and pathways of particular importance were curated manually by specialists and integrated into the genome annotation. The resulting database is publicly available at http://bioinformatics.psb.ugent.be/genomes/*Ostreococcus_tauri*/ in a format that includes browse and query options.

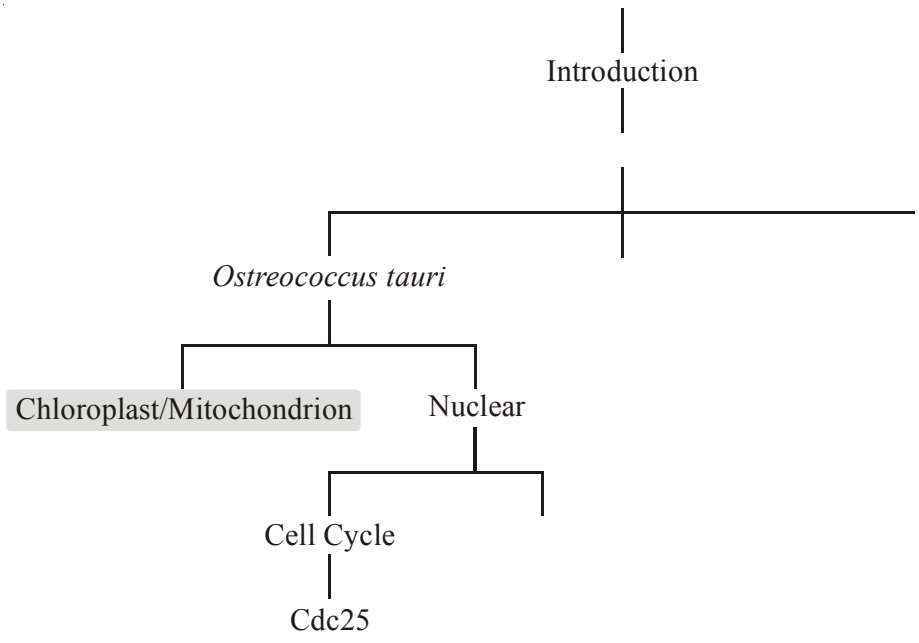
Phylogenetic Analyses.

Homologous genes of *O. tauri* were searched for in public databases by using BlastP. All top hits were retrieved (up to a significant rise in e-value), and the amino acid sequences were aligned by using CLUSTALW. Alignment columns containing gaps were removed when a gap was present in >10% of the sequences. To reduce the chance of including misaligned amino acids, all positions in the alignment left or right from the gap were also removed until a column in the sequence alignment was found where the residues were conserved in all genes included in our analyses. Column conservation was determined as follows: For every pair of residues in the column, the BLOSUM62 value was retrieved. If at least half of the pairs had a BLOSUM62 value = 0, the column was considered as conserved. Neighbor-joining trees were constructed by using TREECON (Van de Peer and De Wachter, 1997), based on Poisson- and Kimura-corrected distances. Bootstrap analyses with 500 replicates were performed to test the significance of the nodes. Genes were only ascribed to a certain taxon if supported at a bootstrap level >70%.

ACKNOWLEDGEMENTS

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

Chapter 5


Introduction
Ostreococcus tauri
Chloroplast/Mitochondrion
Nuclear
Cell Cycle
Cdc25

Mol. Biol. Evol. 24(4), 956-968, 2007*

The Complete Chloroplast and Mitochondrial DNA Sequence of *Ostreococcus tauri*: Organelle Genomes of the Smallest Eukaryote Are Examples of Compaction

Steven Robbens[1,2], Evelyne Derelle[3], Conchita Ferraz[4], Jan Wuyts[1,2], Hervé Moreau[3] and Yves Van de Peer[1,2,#]

[1] Department of Plant Systems Biology, VIB, B-9052, Ghent, Belgium
[2] Department of Molecular Genetics, Ghent University, B-9052, Ghent, Belgium
[3] UMR 7628 CNRS, Université Paris VI, Laboratoire Arago, Banyuls sur Mer, France
[4] Institut de Génétique Humaine, UPR CNRS 1142, Montpellier, France

[#] Correspondence to: yves.vandepeer@psb.ugent.be

Key Words: chloroplast genome, mitochondrial genome, Chlorophyta, *Ostreococcus tauri*

* Modified version

Abstract

The complete nucleotide sequence of the mt (mitochondrial) and cp (chloroplast) genomes of the unicellular green alga *Ostreococcus tauri* has been determined. The mt genome assembles as a circle of 44,237 bp and contains 65 genes. With an overall average length of only 42 bp for the intergenic regions, this is the most gene-dense mt genome of all Chlorophyta. Furthermore, it is characterized by a unique segmental duplication, encompassing 22 genes and covering 44% of the genome. Such a duplication has not been observed before in green algae, although it is also present in the mt genomes of higher plants. The quadripartite cp genome forms a circle of 71,666 bp, containing 86 genes divided over a larger and a smaller single-copy region, separated by 2 inverted repeat sequences. Based on genome size and number of genes, the *Ostreococcus* cp genome is the smallest known among the green algae. Phylogenetic analyses based on a concatenated alignment of cp, mt, and nuclear genes confirm the position of *O. tauri* within the Prasinophyceae, an early branch of the Chlorophyta.

INTRODUCTION

The so-called green lineage (Viridiplantae) is divided into 2 major divisions, namely, Streptophyta and Chlorophyta. Streptophyta contain all known land plants and their immediate ancestors, a group of algae known as "charophyte green algae" (e.g., *Chaetosphaeridium globosum*), whereas Chlorophyta contain the other green algae (e.g., *Chlamydomonas reinhardtii*) that form a monophyletic assemblage and are a sister group to the Streptophyta (Graham and Wilcox, 2000). So far, only 25 (27*) complete mt (mitochondrial) genomes have been sequenced for representatives of the green lineage, 17 (18*) from Streptophyta and 8 (9*) from Chlorophyta. Regarding plastid genomes, 68 (96*) genome sequences are available in public databases, of which 60 (86*) are from Streptophyta and 8 (10*) from Chlorophyta.

The mt genomes of chlorophytes are usually small (25-90 kb), whereas in general a bigger genome size is observed for the streptophytes (from 68 kb for *Chara vulgaris* to around 400 kb for higher plants). The great majority of these genomes are circular, except for some species of Chlamydomonales that have a linear genome (Vahrenholz et al, 1993). The increase of the genome size observed within Streptophyta does not necessarily reflect an increase in coding capacity. Indeed, the transfer of mt genes to the nucleus over evolutionary time (Brennicke et al, 1993), the enlargement and incorporation of new sequences within the mt intergenic spacers, the loss of genes, the increase of intron size, and the resulting decrease of the coding density are all characteristic for the mt genomes of higher land plants. In angiosperms, the most striking feature is the presence of a multipartite genome structure, which results in high-frequency recombination via repeated sequences in the genome (Fauron et al, 1995), altering the genome copy number, which can result in different phenotypes (Kanazawa et al, 1994 and Janska et al, 1998).

All cp (chloroplast) genomes that have been described for land plants have a very conserved genome size, usually around 150 kb covering about 70-80 genes. In contrast, the cp genomes of green algae, although having a rather similar genome size between 150 and 200 kb, show a tremendous variation in gene content, due to massive gene loss, genome erosion, and gene transfer to the nucleus (Grzebyk and Schofield, 2003). All cp genomes described so far are

* udated version (September 2007)

circular. Previous studies have shown that, although in green algae (e.g., *C. reinhardtii*) more genes have been transferred to the nucleus compared with land plants (e.g.,tobacco), the rate of gene flow has subsequently slowed down dramatically and the transfer of DNA from cp to the nucleus is now very rare (Lister et al, 2003). However, until very recently (Derelle et al, 2006 and this study), there was no chlorophyte that had both its nuclear, cp, and mt genome published, and it therefore remained difficult to quantify precisely the extent of gene transfer from the organelles to the nucleus.

Ostreococcus tauri is a unicellular green alga that was discovered in the Mediterranean Thau lagoon (France) in 1994. With a size less than 1 μm, comparable to that of a bacterium, it is the smallest eukaryotic organism currently described (Courties et al, 1994). Its cellular organization is rather simple with a relatively large nucleus with 1 up to 3 nuclear pore(s), a single chloroplast 1 mitochondrion, 1 Golgi body, and a highly reduced cytoplasm compartment (Chrétiennot-Dinet et al, 1995). A membrane surrounds the cells, but no cell wall can be observed. Apart from this simple cellular structure, the *O. tauri* nuclear genome is small (12.56 Mb) and is fragmented into 20 chromosomes (Derelle et al, 2006). Phylogenetically, *O. tauri* belongs to the Prasinophyceae, an early branch of the Chlorophyta (Courties et al, 1998). The presence of only 1 chloroplast and 1 mitochondrion and its basal position in the green lineage makes this alga interesting for studying the structure and evolution of both genomes, whereas comparison with other members of the green lineage sheds light on the evolution of organelle genomes.

MATERIALS AND METHODS

Sequencing.

For the sequencing of the nuclear genome, cellular DNA was used for the preparation of the shotgun libraries (Derelle et al, 2006). Consequently, mt and cp sequences were also obtained and identified by their high similarity with genes of other green algae or green plants. Purified DNA was broken by sonication, and after filling ends, DNA fragments ranging from 1 kb to 5 kb were separated in an agarose gel. Blunt-end fragments were inserted into

pBluescript II KS (Stratagene, The Netherlands), digested with *Eco*RV, and dephosphorylated. Plasmid DNA from recombinant *Escherichia coli* strains was extracted according to the TempliPhi method (Amersham, GE Healthcare, France), and inserts were sequenced on both strands using universal forward and reverse M13 primers and the ET DYEnamic terminator kit (Amersham). Sequences were obtained with MegaBace 1,000 automated sequencers (Amersham). Data were analyzed and contigs were assembled using Phred-Phrap (Ewing et al, 1998) and Consed software packages (http://bozeman.mbt.washington.edu/ consed/consed.html). Gaps were filled through primerdirected sequencing using custom made primers.

Gene Prediction and Annotation.

All genes were annotated based on their similarity with cp and mt genes that were available in public databases and if necessary manually corrected using Artemis (Rutherford et al, 2000). Homologous relationships between publicly available genes and the *O. tauri* genes were identified through Blast (Altschul et al, 1990). Also small and large ribosomal subunit RNA genes were identified by Blast. Alignment and secondary structure annotation was done using the DCSE alignment editor (De Rijk and De Wachter, 1993). The secondary structure drawings were made using RNAVIZ (De Rijk et al, 2003). tRNA genes were identified by TRNASCAN-SE (Lowe and Eddy, 1997) using the option "search for organellar tRNAs (-O)". The 5S rRNA gene of the cp genome was identified using the CMSEARCH program from the INFERNAL package (Eddy 2002) with the 5S rRNA covariance model (RF00001) from the RFAM database (Griffiths-Jones et al, 2005).

Sequence Analyses.

Pairwise comparison of gene permutations by inversions between different mt and cp genomes was obtained using the GRIMM web server (Tesler, 2002). The data sets used contained, respectively, 54 conserved mt and 82 conserved cp genes. As this tool cannot deal with duplicated genes, genes located in the inverted repeats (IRs) were counted only once.

Duplicated sequences within both genomes were identified using DOTTER (Sonnhammer and Durbin, 1995). For both genomes (but including only one

of the IR sequences), short repeated sequences were identified with REPUTER 3.1 (Kurtz et al, 2001), using the -p (palindromic), -f (forward), -l (minimum length), and -allmax parameters; and MUMMER 3.0 (Kurtz et al, 2004), using the -l (minimum length) and -b (forward and reverse complement matches) options. PIPMAKER (Schwartz et al, 2000) was used to visualize the location of the repeated sequences.

Phylogenetic Analysis.

Homologous genes of *O. tauri* cp and mt genes were searched for in the public databases (GenBank/EMBL/ DDBJ) (Benson et al, 2002; Stoesser et al, 2002; and Tateno et al, 2002) using BLASTP (Altschul et al, 1997). Protein sequences were aligned with CLUSTALW (Thompson et al, 1994). Two different data sets were built:

1. Forty-seven cp protein sequences (*atpA, atpB, atpE, atpF, atpH, clpP, petB, petG, psaA, psaB, psaC, psaJ, psbA, psbB, psbC, psbD, psbE, psbF, psbH, psbI, psbJ, psbK, psbL, psbN, psbT, psbZ, rbcL, rpl14, rpl16, rpl2, rpl20, rpl36, rpoA, rpoB, rpoC1, rpoC2, rps11, rps12, rps14, rps18, rps19, rps3, rps4, rps7, rps8, ycf3,* and *ycf4*) from 14 different organisms (*Chlorella vulgaris* [Wakasugi et al, 1997] [AB001684], *Nephroselmis olivacea* [Turmel et al, 1999a] [AF137379], *Pseudendoclonium akinetum* [Pombert et al, 2005] [AY835431], *Stigeoclonium helveticum* [Bélanger et al, 2006] [DQ630521], *Scenedesmus obliquus* [de Cambiaire et al, 2006] [DQ396875], *Oltmannsiellopsis viridis* [Pombert et al, 2006a] [DQ291132], *C. reinhardtii* [Maul et al, 2002] [BK000554], *Mesostigma viride* [Lemieux et al, 2000] [AF166114], *C. globosum* [Turmel et al, 2002a] [AF494278], *Marchantia polymorpha* [Ohyama et al, 1986] [M68929], *Nicotiana tabacum* [Shinozaki et al, 1986] [Z00044], *Pinus thunbergii* [Wakasugi et al, 1994] [D17510], *Cyanophora paradoxa* [Stirewalt et al, 1995] [U30821], and *O. tauri*) were independently aligned and concatenated into a data set of 9,553 amino acids.

2. A nuclear gene (small subunit [SSU] rRNA), 1 mt gene (*nad5*), and 2 cp genes (*rbcL* and *atpB*), encompassing 44 organisms, were combined into a data set of 5,053 nucleotides (based on Karol et al, 2001)

PHYML 2.4.4 (Guindon and Gascuel, 2003) was used to compute maximum likelihood trees, using the cpREV45 model for cp sequences and the Hasegawa,

Kishino and Yano (1985) model for the combined nucleic acid data set. Pairwise distance trees were obtained using TREECON (Van de Peer and De Wachter, 1994), based on Poisson (Zuckerkandl and Pauling, 1965 and Dickerson, 1971) and Kimura (1983) corrected distances for the protein alignment and Jukes and Cantor (1969) corrections for nucleic acid sequences. PHYLIP (the Phylogeny Inference Package; Felsenstein, 1989) was used for 1) computing pairwise distance trees using the Dayhoff PAM matrix (1979) for protein alignment and Jukes and Cantor (1969) for nucleic acid sequences and 2) obtaining maximum parsimony trees for both data sets. For each method, bootstrap analyses with 500 replicates were performed to test the significance of the nodes. Finally, MrBayes (500.000 generations and 4 chains) was used for Bayesian inference of phylogenetic trees (Huelsenbeck et al, 2001), using a JTT+γ substitution model (Jones et al, 1992).

After manual improvement of the alignments using BIOEDIT (Hall, 1999), only unambiguously aligned positions were taken into account for tree construction. TREEVIEW was used to visualize the trees (Page, 1996).

RESULTS AND DISCUSSION

Phylogenetic Analyses.

Previous phylogenetic analyses based on the 18S rDNA sequence of different Chlorophyta suggested that *O. tauri* belongs to the Prasinophyceae, an early diverging group within the green plant lineage (Courties et al, 1998). Now, with the availability of the cp, mt, and even nuclear (Derelle et al, 2006) genomes of *O. tauri*, a more extensive phylogenetic analysis could be performed. To this end, we prepared 2 different data sets, that is, 1 consisting of concatenated cp genes and 1 consisting of a mix of concatenated cp, mt, and nuclear genes (see Materials and Methods). Using these different data sets and different methods of phylogenetic tree construction, *O. tauri* always clustered with other members of the Chlorophyta, clearly confirming its Chlorophycean heritage (fig. 1; see appendix fig. A1). Furthermore, the different classes within the Chlorophyta, namely, Chlorophyceae, Trebouxiophyceae, Ulvophyceae, and Prasinophyceae formed monophyletic groups, well supported by bootstrap analyses. *O. tauri* and *N. olivacea* were always grouped together within an

early diverging group referred to as the Prasinophyceae, thereby confirming the previous analyses done by Courties (1998).

In our phylogenetic analyses, we have also included the unicellular freshwater

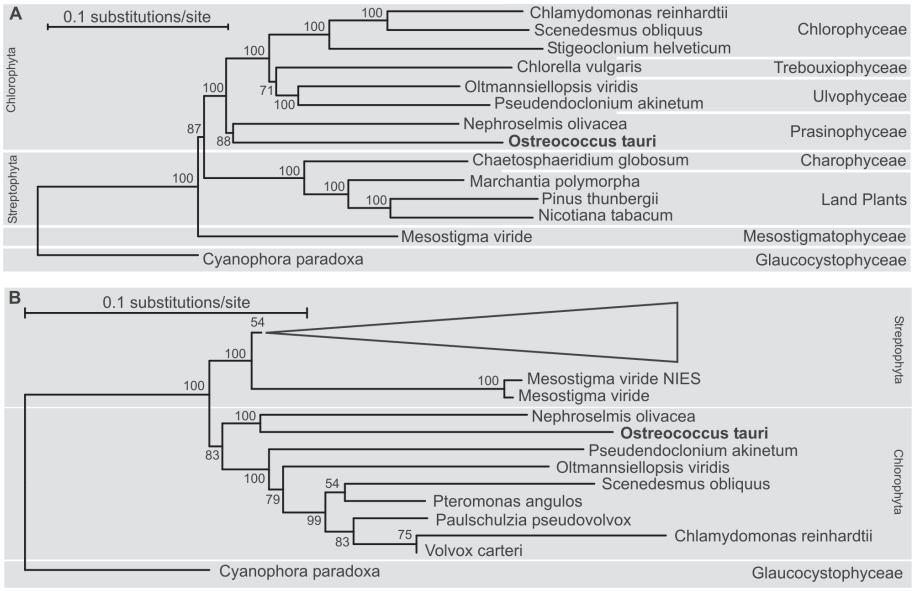


Figure 1. Phylogenetic position of *Ostreococcus tauri*. (A) Tree was made based on 47 concatenated sequences of cp genes, using the pairwise distance method. As each method gave the same topology, the lowest bootstrap values are shown at the branches of the nodes. (B) Tree inferred by pairwise distance methods of a combined data set of a mt, nuclear, and 2 cp genes, based on the Karol et al, (2001) data set. All members of the Streptophyta except *Mesostigma viride* are replaced by a triangle. The complete tree can be found in appendix fig. A1 . Lowest bootstrap values, using different methods, are shown at the nodes. For both trees, the branch lengths are drawn to scale.

alga, *M. viride*, whose phylogenetic position is still being discussed (previously referred to as the "enigma of Mesostigma" [McCourt et al, 2004]). After being classified as an primitive chlorophyte (Mattox and Stewart, 1984; Grzebyk and Schofield, 2003; and Nozaki et al, 2003), a charophyte (Melkonian, 1989; Bhattacharya et al, 1998; Karol et al, 2001; and Martin et al, 2002), or as a species branching off prior to the divergence of the Streptophyta and Chlorophyta (Lemieux et al, 2000 and Turmel et al, 2002b), Petersen et al, (2006) provided unequivocal support for its Streptophycean affiliation, based on the presence of a land plant-specific *gapB* gene and the absence of this gene in the different orders of chlorophyte green algae. However, our tree based on concatenated cp genes (fig. 1a) clearly support *M. viride* branching off before

the divergence of Chlorophyta and Streptophyta. On the other hand, trees bases on a combination of concatenated plastid, mt, and nuclear genes did group *M. viride* with the other streptophytes (fig. 1b, see appendix fig. A1). In addition, we recently showed (Robbens et al, 2007) that 2 *Ostreococcus* strains (*O. tauri* and *Ostreococcus lucimarinus*) also contain the *gapB* gene (DQ649078 and DQ649079), making this gene no longer land plant specific as postulated by Petersen et al, (2006). As a matter of fact, all this adds some more mystery to the phylogenetic position of *M. viride*.

Structure and Gene Content of the mt Genome.

The *O. tauri* mt genome assembles as a circle of 44,237 bp (fig. 2), with an overall GC content of 38%. This size is similar to the mt genome of another early branching chlorophyte *N. olivacea* (45,223 bp) (Turmel et al, 1999b). However, in contrast to the *N. olivacea* genome, the *O. tauri* mt sequence contains a duplicated region, containing 22 genes and covering 44% (19,542 bp) of the genome (see further). Sixty-five genes (unique open reading frames [ORFs] were not taken into account, and duplicated genes were counted only once) are encoded on both strands, encompassing 93% of the genome, which makes the mt genome of *O. tauri* the most gene dense among the Chlorophyta. For comparison, both *M. viride* (Turmel et al, 2002b) and *N. olivacea* also have sixty-five genes, but only covering 87% and 81% of their genome, respectively (table 1). Among the 65 genes, 36 are protein-encoding genes, 26 are transfer RNAs, and 3 are rRNAs (table 2). Two predicted proteins (*orf129* and *orf153*) coding for 129 and 153 amino acids, respectively, did not show any clear similarity to other known genes. The compactness of the *O. tauri* mt genome is further illustrated by the shortness of the intergenic regions, ranging from 1 to 475 bp, with an average of 42 bp. Only 5 intergenic regions exceed 100 bp, and these are all located in the duplicated region. In addition, there are 3 cases of overlapping genes (trnR1-*rnpB*, *rps14*-*rpl5*, and *orf153*-trnH). Lastly, in contrast to other members of the green lineage, neither group I nor group II type introns are present in any of the genes.

All 26 tRNAs fold into the conventional cloverleaf secondary structure and are able to decode all codons. The small subunit rRNA (SSU rRNA, rns in fig. 2) gene is fragmented into 2 parts, but retains its ability to fold into the normal

secondary structure model (see appendix fig. A2). The fragmentation site is located near the hairpin loop of helix 29 (indicated by gray area) of the secondary structure model (Wuyts et al, 2004), and the location of the 2 fragments has been rearranged in the genome such that both fragments are located on the forward strand but their order is reversed. However, this fragmentation site does not correspond to one of the several fragmentation sites that have been previously identified in the small subunit rRNA genes of chlorophyte mt genomes (Nedelcu et al, 2000) (other known fragmentation sites are indicated by gray areas on appendix fig. A2). The SSU rRNA, LSU rRNA (large subunit rRNA, rnl in fig.2), and 5S rRNA gene (rrn5 in fig. 2) are all located in the duplicated region. Like other known members of the green lineage, the *O. tauri* mitochondrion uses the standard genetic code, and all the 61 codons are used. As in *M. viride*, there is a strong bias in favor of codons that end in A or U. The 3 types of stop codons are present with UAA being the one most often used (81.4%) (see appendix table A3).

Table 1. General features of mt-genomes of different green organisms

Feature	Ot	No	Cr	Mv	Cg	Mp	At
Size (bp)	44,237	45,223	15,758	42,424	56,574	186,609	366,924
GC (%)	38.0	32.8	45.2	32.2	34.4	42.2	44.8
Gene number[a]	65	65	13	65	67	69	50
Gene density (%)	93.0	80.6	83.1	86.6	76.3	65.0	45.5
Intron number	-	4 (I)	-	4 (I), 3 (II)	9 (I), 2 (II)	7 (I), 25 (II)	23 (II)
Repeated seq (bp)[b]	9,771	-	500	-	-	-	11,372
Map	circular	circular	linear	circular	circular	circular	circular

Ot: Ostreococcus tauri, No: Nephroselmis olivacea, Cr: Chlamydomonas reinhardtii, Mv: Mesostigma viride, Cg: Chaetosphaeridium globosum, Mp: Marchantia polymorpha, At: Arabidopsis thaliana

- not present

[a] Unique ORFs were not taken into account and duplicated genes were counted only once

[b] minimum 500 bp long and 100% identical

The most striking feature of the *O. tauri* mt genome is the presence of a large duplicated segment (19,542 bp; shaded box in fig. 2). This duplication is also observed in the partially sequenced mt genome of another *Ostreococcus* strain (*O. lucimarinus*; Palenik B, personal communication), thereby excluding erroneous genome assembly. The presence of such a duplicated sequence has not been observed in any other member of the Chlorophyta, except for *C. reinhardtii*, wherein its mt genome, which is linear instead of circular, terminal IRs of approximately 500 bp have been described (Vahrenholz et al, 1993; table 1). No duplication is present in the mt genome of the charophyte *Chara vulgaris* (Turmel et al, 2003). The only large repeated sequences previously reported are present in higher land plants (e.g., *Arabidopsis thaliana*: 366,924

bp, containing repeat sequences of 6.5 and 4.5 kb and *Beta vulgaris*: 368,799 bp, containing a repeat sequence of 6.2 kb) (Unseld et al, 1997 and Kubo et al, 2000). These repeated regions in the mt genome of angiosperms gave rise to a multipartite genome structure (Fauron et al, 1995) and lead to high-frequency intramolecular recombination. Indeed, a master circle, containing the complete genetic information, can lead to different subgenomic circles by homologous recombination via a repeated sequence motif (e.g., the tobacco mt genome can provide 6 different subgenomic circles by homologous recombination between the different repeated sequences) (Knoop 2004 and Sugiyama et al, 2005). The presence of this multipartite genome structure enables them to change their gene and genome copy number, resulting in an altered plant phenotype (Kanazawa et al, 1994 and Janska et al, 1998).

The unique repeated segment in the mt genome of *O. tauri* contains 5 protein-coding genes (*cob*, *cox1*, *atp4*, *atp8*, and *nad4L*), 2 tRNAs (trnMe and trnY), the SSU rRNA (rns), LSU rRNA (rnl), and 5S rRNA (rrn5) genes and the *orf129*. The duplicated nucleotide sequences are 100% identical over a length of 9,771 bp, covering 44% of the genome (see supplementary fig. S3, Supplementary Material online). Repeats that are a 100% identical in mt genomes are not exceptional. For instance, *Brassica napus* has 2 repeats with 1 mismatch over 2,427 bp (Handa, 2003), *B. vulgaris* has 2 repeats (Kubo et al, 2000), and *A. thaliana* has 3 repeats that are 100% identical (Unseld et al, 1997).

Short dispersed repeats (SDR) are also thought to play an important role in mt genome rearrangements, thereby altering the gene content and genome size. This is not only true for members of the Chlorophyta, but also in land plants, yeasts, and even animals, where they serve as hot spots for recombination (Pombert et al, 2004). Short dispersed repeats have been described in all known members of the Chlorophyta, although their abundance is highly variable. All Chlorophyta members hold SDRs of at least 15 bp in their genome. This number is reduced to 52 repeats in *N. olivacea* (Pombert et al, 2006b) and to only 11 in *O. tauri*. The largest repeats found in *O. tauri* and *N. olivacea* are rather short, being 34 bp and 42 bp, respectively. The GC content of the SDRs present in *O. tauri* does not differ much from the overall GC content of the mt genome (36% for the SDRs vs. 38% for the complete genome). In general, more derived

lineages show an increase of the number of SDRs: *O. viridis* contains 1,206 (Pombert et al, 2006b), *Scenedesmus obliquus* 4,086 (Nedelcu et al, 2000), and *P. akinetum* 8,002 (Pombert et al, 2004) SDRs of at least 15 bp long. It seems that after the split of the Prasinophyceae, an increase of SDRs took place (with the exception of the Chlamydomonadales), and it is tempting to correlate this increase with the gene rearrangements that took place within the other members of the Chlorophyta (see further).

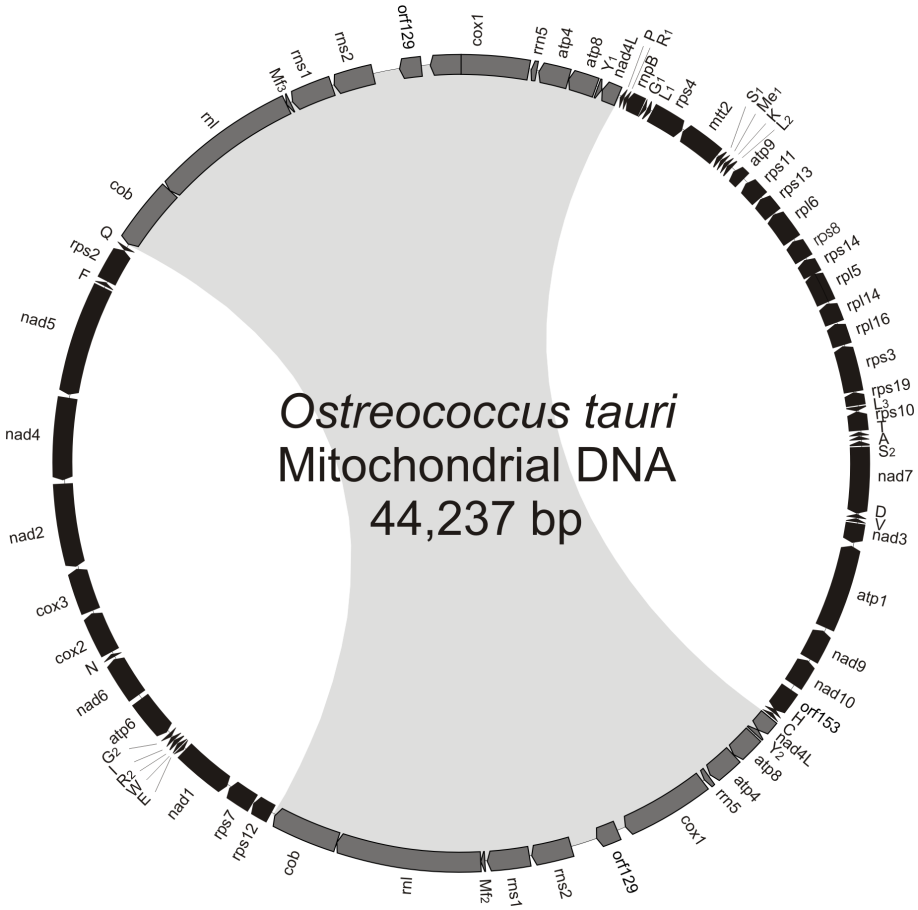


Figure 2. *Ostreococcus tauri* mt genome. Genes located in the unique duplicated region are colored in gray; single-copy genes are in black. The length of the boxes is proportional to their amino acid length. tRNA genes are represented by the 1-letter amino acid code, and the unique ORFs are indicated by orf followed by their amino acid length.

Table 2. Gene list of *Ostreococcus tauri* mitochondrial DNA

Gene Products	Genes
Small subunit ribosomal proteins (11)	rps 2,3,4,7,8,10,11,12,13,14,19
Large subunit ribosomal proteins (4)	rpl 5,6,14,16
NADH dehydrogenase (10)	nad 1,2,3,4,4L*,5,6,7,9,10
ATP synthase (5)	atp 1,4*, 6,8*,9
Cytochrome c oxidase (3)	cox 1*,2,3
Ubiquinol: cytochrome c oxidoreductase (1)	cob*
Sec-independent protein translocation pathway (1)	mtt2
Ribosomal RNAs (3)	rns*,rnl*, rrn5*
Transfer RNAs (26)	Y(gua)*, P(ugg), R1(ucu), G1(ucc), L1(uaa), S1(uga), Me1(cau), K(uuu), L2(uag), L3(gag), T(ggu), A(ugc), S2(gcu), D(guc), V(uac), H(gug), C(gca), Q(uug), Mf(gau)*, E(uuc), W(cca), R2(acg), I(gau), G2(gcc), N(guu), F(gaa)
Rnase P RNA (1)	rnpB
unknown ORF (2)	orf129*, orf153

* located in duplicated block

Numbers within parentheses indicate the number of genes in this class (duplicated genes were counted only once)

Comparison with Other mt Genomes.

Comparison of the *O. tauri* mt genome with 9 other species of the Viridiplantae lineage (Cr: *C. reinhardtii*, No: *N. olivacea*, Ov: *O. viridis*, Pa: *P. akinetum*, So: *S. obliquus*, At: *A. thaliana*, Mp: *M. polymorpha*, Cg: *C. globosum*, and Mv: *M. viride*) unveiled only 9 genes (not including tRNAs), which are common to all these species (table 3). However, when removing *C.reinhardtii* (Michaelis et al, 1990) and *S. obliquus*, 2 members of the Chlorophyceae, from this comparison, this number increases to 25 shared genes. When further removing the 2 ulvophyte green algae (*O. viridis* and *P. akinetum*), the number of conserved genes increases to 30, thus, representing the gene content conservation between the 2 prasinophytes and the land plants. However, when only considering the protein-coding genes of *O. tauri*, *N. olivacea*, and *M. viride*, 36 genes are shared, which represents 95% of the *O. tauri* and 92% of the *M. viride* protein-coding gene content. Apparently, the gene content conservation between these genomes, which are assumed to represent a more ancestral state, is still very high. One of the 7 protein-coding genes that are absent in the *O. tauri* mt genome, namely *rpl2*, could be uncovered in the nuclear genome (see supplementary table S3, Supplementary Material online).

Disregarding the unique ORFs and tRNAs (trnG[gcc] and trnL[gag] seem to have been lost in *N. olivacea* compared with *O. tauri*, whereas trnR[ucg] is lost in *O. tauri* compared with *N. olivacea*), the gene repertoires of *O. tauri*

and *N. olivacea* are identical (table 3). Furthermore, there is a high degree of synteny between these 2 algae, with 5 gene clusters of at least 5 genes and 1 of 2 genes, which are almost identical in both mt genomes (genes denoted in black in fig. 3). However, when one considers gene polarities, synteny is limited to only 2 gene clusters (12 genes extending from *rps11* to *rps10* and 5 genes extending from *atp6* to *cox3*). The major difference between both mt genomes is the duplication in *O. tauri* and the presence of 4 group I introns in *N. olivacea* (3 within the rnl and 1 in the *cob* gene) (Turmel et al, 1999b).

A certain degree of synteny can still be detected when adding *C. globosum* (charophyte) (Turmel et al, 2002a) and *Marchantia polymorpha* (streptophyte) (Oda et al, 1992) to the 2 previous species (genes in black and gray in fig. 3), except for cluster 3 where no clear synteny could be detected among the 4 organisms and; for cluster 2 (fig. 3), where the genes of *M. polymorpha* are divided into 2 parts ([*atp6*, *nad6*, and trnN] and [*cox2* and *cox3*]). In contrast, synteny conservation still exists in cluster 5 where 9 genes are present in the 4 organisms, all oriented in the same direction, indicating that although the genome size increases from green algae to higher land plants, the gene organization of some clusters are extremely well conserved in evolution.

Additionally, we estimated the number of gene inversions needed to transform the gene organization of one genome into another, thereby providing quantitative measurement of their evolutionary distances. Fifty-four conserved genes (duplicated genes were used only once) of 3 Chlorophyta (*O. tauri*, *N. olivacea*, and *P. akinetum*) and *M. viride* were used, showing that a minimum of 29 inversions are needed to transform the gene organization of *O. tauri* into that of *N. olivacea*. When comparing *O. tauri* with the other mt genomes, almost twice as many inversions are needed (50 for both *P. akinetum* and *M. viride*), again indicating the close relationship between the 2 Prasinophyceae.

Structure and Gene Content of the cp Genome.

With a circular cp genome of 71,666 bp long (fig. 4), *O. tauri* contains the smallest cp genome known so far within the Viridiplantae (except for the parasite *Helicosporidium sp.* [de Koning and Keeling, 2006]). Cp genome size in green algae ranges from 118,360 bp in *M. viride* (Lemieux et al, 2000) to 203,395 bp in *C. reinhardtii* (Maul et al, 2002; table 4). The GC content (39.9%) of the *O.*

Table 3. Comparison of gene Content in Green Algal and Land Plant Mitochondrial genomes

	Ov	Pa	So	Cr	No	Ot	Mv	Cg	At	Mp
NADH dehydrogenase										
nad1	*	*	*	*	*	*	*	*	*	*
nad2	*	*	*	*	*	*	*	*	*	*
nad3	*	*	*		*	*	*	*	*	*
nad4	*	*	*	*	*	*	*	*	*	*
nad4L	*	*	*		*	*	*	*	*	*
nad5	*	*	*	*	*	*	*	*	*	*
nad6	*	*	*	*	*	*	*	*	*	*
nad7	*	*			*	*	*	*	*	*+
nad9	*				*	*	*	*	*	*
nad10					*	*				
Cob-complex										
cob	*	*	*	*	*	*	*	*	*	*
Cytochrome c oxidase										
cox1	*	*	*	*	*	*	*	*	*	*
cox2	*	*	*		*	*	*	*	*	*
cox3	*	*	*		*	*	*	*	*	*
ATP synthase										
atp1	*	*			*	*	*	*	*	*
atp4	*	*			*	*	*	*	*	*
atp6	*	*	*		*	*	*	*	*	*
atp8	*	*			*	*	*	*	*	*
atp9	*	*	*		*	*	*	*	*	*
Conserved protein										
mtt2	*	*			*	*	*	*	*	*
LSU ribosomal proteins										
rpl2								*	*	*
rpl5		*			*	*	*	*	*	*
rpl6					*	*	*	*		*
rpl14					*	*	*			
rpl16	*	*			*	*	*	*	*	*
SSU ribosomal proteins										
rps1							*	*		*
rps2	*	*			*	*	*	*		*
rps3	*	*			*	*	*	*	*	*
rps4		*			*	*	*	*	*	*
rps7					*	*	*	*	*	*
rps8					*	*				*
rps10		*			*	*	*	*		*
rps11	*	*			*	*	*	*		*
rps12	*	*			*	*	*	*	*	*
rps13	*	*			*	*	*	*		*
rps14	*	*			*	*	*	*	*+	*
rps19	*	*			*	*	*	*	*+	*
Ribosomal RNAs										
rns	*	*	*	*	*	*	*	*	*	*
rnl	*	*	*	*	*	*	*	*	*	*
rrn5	*				*	*	*	*	*	*
Rnase P RNA										
rnpB					*	*				
Transfer RNAs	24	25	27	3	26	26	26	28	17	27

Cr: *Chlamydomonas,* No: *Nephroselmis* , Ot: *Ostreococcus* ,
Mv: *Mesostigma* , Ov: *Oltmannsiellopsis* ,
Pa: *Pseudendoclonium* , So: *Scenedesmus* ,
Cg: *Chaetosphaeridium* , Mp: *Marchantia,* At: *Arabidopsis*
*+ pseudogene
* present
Bold gene names indicate conserved genes

tauri cp genome is close to that of *N. olivacea* (42.1%) (Turmel et al, 1999a) and *O. viridis* (40.5%) (Pombert et al, 2006a), but higher than that of other chlorophytes, such as *Chlorella vulgaris* (31.6%) (Wakasugi et al, 1997), *C. reinhardtii* (34.6%), and *M. viride* (30.1%). Like all known members of the Chlorophyta, except *C. vulgaris*, the cp genome of *O. tauri* has a quadripartite structure containing 2 large IRs of 6,825 bp (covering 9.5% of the genome) separating a large single-copy (LSC) region (35,684 bp, covering 49.8%) and a small single-copy (SSC) region (22,332 bp, covering 31.2%) (fig. 4 and supplementary fig S4, Supplementary Material online). Despite the difference in size, both the LSC and the SSC contain 41 genes, whereas the IR sequences contain, next to *psbA*, the rRNA operon (rrs, trnI[gau], trnA[ugc], rrl, and rrf). Besides its ultrasmall cp genome, the gene content is reduced to a minimum: 86 genes (unique ORFs were not taken into account, and duplicated genes were counted only once) were identified, including 25 tRNAs and the rRNA gene cluster (rrf, rrl, and rrs). Two predicted proteins (*orf537* and *orf1260*) coding for 537 and 1,260 amino acids, respectively, show little similarity with known genes: *ycf1* and *ycf2*. These genes will be indicated as *orf537/ycf1* and *orf1260/ycf2* (table 5). This gene repertoire is the smallest known to date among the green algae: *C. reinhardtii* has a slightly higher number of genes (94 genes, not including the duplicated genes and unique ORFs), but this number is also much lower than the number of genes present in other Chlorophyta (e.g., *N. olivacea* contains 127 genes and *M. viride* contains 135 genes) (table 4). Twenty-five tRNAs could be detected, a number that is low compared with that of other members of the green lineage (e.g., *N. olivacea*: 32, *M. viride*: 37, and *A. thaliana*: 37 [Sato et al, 1999]), but enables the *O. tauri* cp genome to decode all 61 codons. Two of the 3 types of stop codons are present with UAA being the one most often used (95%) and with UGA being absent (see appendix table A4). The average length of intergenic regions is 116 bp, varying from 1 to 476 bp. There are 3 cases of overlapping genes (*psbC-psbD*, *rpoC1-rpoC2*, and *rps3-rpl16*), resulting in an average coding density (including conserved genes, unique ORFs, and introns) of 84.7%. One group II intron, present in *atpB*, could be detected.

Table 4. General features of cp-genomes of different green organisms

Feature	Ot	No	Cr	Mv	Cg	Mp	At
Size (bp)	71,666	200,799	203,827	118,360	131,183	121,024	154,478
GC (%)	39.9	42.1	34.5	30.1	29.6	28.2	36.3
Gene number[a]	86	128	94	136	125	120	87
Gene density (%)	84.7	68.7	50.1	?	?	?	?
Intron number	1 (II)	-	5 (I), 2 (II)	-	1 (I), 17 (II)	1 (I), 19 (II)	1 (I), 20 (II)
IR (bp)	6,825	46,137	22,211	6,057	12,430	10,058	26,264
SSC (bp)	22,332	16,399	78,088	22,619	17,640	19,813	17,780
LSC (bp)	35,684	92,126	81,307	83,627	88,683	81,095	84,170

Ot: Ostreococcus tauri, No: Nephroselmis olivacea, Cr: Chlamydomonas reinhardtii, Mv: Mesostigma viride,
Cg: Chaetosphaeridium globosum, Mp: Marchantia polymorpha, At: Arabidopsis thaliana
- not present
? Not known
[a] Unique ORFs were not taken into account and duplicated genes were counted only once

Comparison with Other cp Genomes

The gene repertoire of the cp genomes of 7 Chlorophyta (Cr: *C. reinhardtii*, Cv: *C. vulgaris*, No: *N. olivacea*, Ov: *O. viridis*, Pa: *P. akinetum*, So: *S. obliquus*, and Ot: *O. tauri*), 2 Streptophyta (At: *A. thaliana* and Nt: *N. tabacum*), and *M. viride* (Mv) were compared and the results shown in table 6. Fifty-three core genes are shared between both Chlorophyta and Streptophyta (bold gene names), whereas 4 additional core genes (*ycf12*, *tufA*, *rpl5*, and *rps9*) are present when only considering the Chlorophyta lineage. The 53 core cp genes are involved either in photosynthesis, energy metabolism, or some housekeeping functions. Gene loss and gene transfer to the nucleus is a common feature of cp genomes (Stegemann et al, (2003)), and (Grzebyk and Schofield, 2003) reported the loss of 7 genes (*rpl21*, *rpl22*, *rpl33*, *rps15*, *rps16*, *odpB*, and *ndhJ*) at the base of the Chlorophyta lineage. These genes were also not detected in the *O. tauri* cp genome, but 5 of them are present in the nuclear genome (see supplementary table S3, Supplementary Material online).

In *O. tauri*, 34 genes are lost in the cp genome compared with other Chlorophyta: 1) the 10 homologs of the mt *ndh* genes, subunits of the NADH:ubiquinone oxidoreductase. None of these genes were present in the nuclear genome; 2) the genes *chlB*, *chlI*, *chlL*, and *chlN* involved in the chlorophyl synthesis in dark. In almost all known green algal cp genomes, these 4 genes are present, but not in *O. tauri* where only *chlI* was found in the nuclear genome (on chromosome 2). The absence of *chlB*, *chlL*, and *chlN* in the cp or nuclear genome of *O. tauri* confirms the inability of this organism to produce chlorophyl in dark (Derelle et al, 2006); 3) both the *petL* and *petD* genes are absent in *O. tauri*, whereas they are present in all other studied organisms where they encode a small subunit of the cytochrome b6f complex. The *petL* has not been

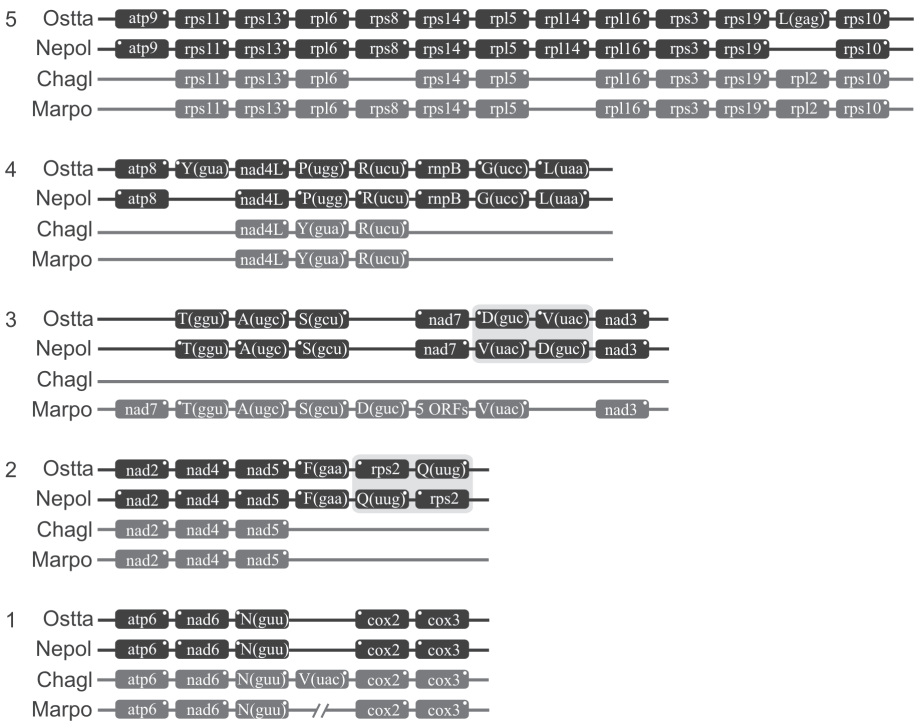


Figure 3. Overview of gene clusters conserved between *O. tauri* (Ostta), *Nephroselmis olivacea* (Nepol), *Chaetosphaeridium globosum* (Chagl), and *Marchantia polymorpha* (Marpo). Black boxes represent the genes of *O. tauri* and *N. olivacea*, and gray boxes represent the genes of both Streptophyta. White dots in the upper-left or upper-right corner indicate the orientation of the genes. Gaps represent the loss of genes. tRNA genes are represented by the one-letter amino acid code followed by the anticodon in parentheses. Light gray background boxes indicate inversions of gene order.

transferred to the nucleus, whereas *petD* could be located on chromosome 7. However, it has been shown in *C. reinhardtii* that a free *petL* N-terminus is not required for the b6f complex function (Zito et al, 2002); 4) *psbM*, a part of the photosystem II reaction center, is absent in the cp genome, but is present in the nucleus (chromosome 12); and 5) at least 13 additional genes (*petN*, *minE*, *minD*, *ftsI*, *ftsW*, *ftsH*, *rpl12*, *rpl19*, *accD*, *cemA*, *ccsA*, *cysA*, and *cysT*) and 3 unknown conserved genes (*ycf4*, *ycf6*, and *ycf10*) have been lost in *O. tauri*. However, 5 of them are present in the nuclear genome (*ycf4*, *minD*, *rpl12*, *rpl19*, and *cemA/ycf10*) (see supplementary table S3, Supplementary Material online).

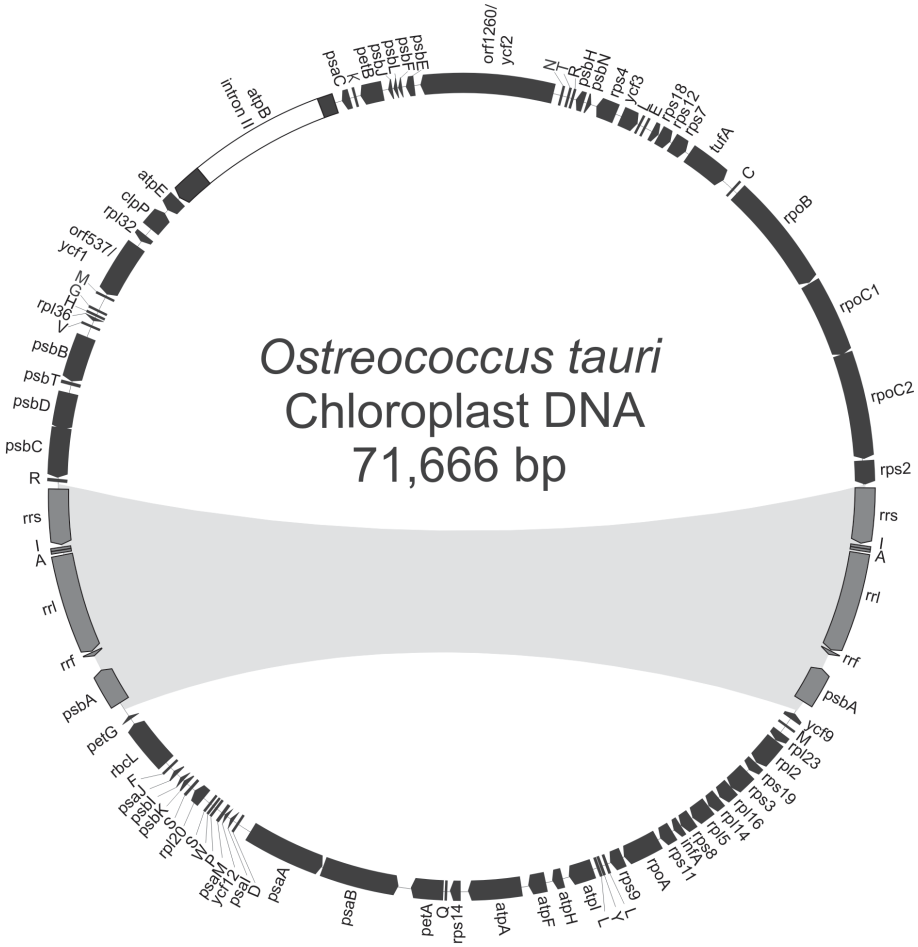


Figure 4. *Ostreococcus tauri* plastid genome. Genes located in the inverted repeat sequences are colored in gray; genes in the single copy regions are black. The single intron, located in *atpB*, is shown as a white box. The length of the boxes is proportional to their amino acid length. tRNA genes are represented by the 1-letter amino acid code and the unique ORFs are indicated by orf followed by their amino acid length.

Despite these differences in gene content, 10 conserved blocks, ranging from 2 to 12 genes are shared between *O. tauri* and *N. olivacea*, 11 between *O. tauri* and *C. vulgaris*, and 12 between *O. tauri* and *M. viride*. When aligning the 4 genomes together, 9 conserved blocks of at least 2 genes can be unveiled. However, when adding the cp genome of *C. reinhardtii*, whose genome is structurally the most comparable to that of *O. tauri* (see below), almost no conserved blocks shared by all species, can be detected. Comparison of the cp genome of *O. tauri* with the one of *O. viridis*, a member of the Ulvophyceae, also showed shared gene clusters. So in general, without considering *C.*

Table 5. Gene list of *Ostreococcus tauri* chloroplast DNA

Gene Products	Genes
Photosystem I (6)	psa A,B,C,I,J,M
Photosystem II (14)	psb A,B*,C,D,E,F,H,I,J,K,L,N,T,Z
Cytochrome b6/f (3)	pet A,B,G
ATP synthase (6)	atp A,B[a],E,F,H,I
Rubisco (1)	rbcL
Large subunit ribosomal proteins (8)	rpl 2,5,14,16,20,23,32,36
Small subunit ribosomal proteins (11)	rps 2,3,4,7,8,9,11,12,14,18,19
RNA polymerase (4)	rpo A,B,C1,C2
Translation factors (2)	infA, tufA
Miscellaneous proteins (1)	ClpP
Proteins of unknown functions (2)	ycf 3,12
Ribosomal RNAs (3)	5S*,16S*, 23S*
Transfer RNAs (25)	A(ugc)*, I(gau)*, R(acg), V(uac), H(gug), G(ucc), M(cau), K(uuu), N(guu), T(ggu), R(ucu), L(uag), E(uuc), C(gca), fM(cau), L(gag), Y(gua), L(uaa), Q(uug), D(guc), P(ugg), W(cca), S(gcu), S(uga), F(gaa)
unknown ORF (2)	orf537/ycf1, orf1260/ycf2

* Located in duplicated block

[a] Genes containing intron

Numbers within parentheses indicate the number of genes in this class (duplicated genes were counted only once)

reinhardtii, 9 conserved blocks of at least 2 genes can be unveiled between different members of the Chlorophyta, representing 33 genes (for *O. tauri* 37% of its gene content), indicating the importance of maintaining certain gene clusters throughout evolution. However, if we compare the gene order of *O. tauri* cp genome with the 24 "ancestral" gene clusters present in *N. olivacea* and *M. viride* (de Cambiaire et al, 2006), only 7 of them are completely present in *O. tauri*, indicating the loss of its ancestral characteristics.

The number of gene inversions necessary to transform the gene organization of one genome into another has been estimated for 4 Chlorophyta (*O. tauri*, *N. olivacea*, *O. viridis*, and *C. vulgaris*) and for *M. viride*. An average of 50 inversions is needed to transform the gene organization of *O. tauri* into that of any other of these cp genomes.

Although some genes and gene clusters are well conserved among green algae, the overall structure of the cp genomes can show remarkable differences. First, both the LSC and the SSC region of *O. tauri* cp genome contain 41 genes, in contrast to the cp genomes of other green algae (*N. olivacea*, *M. viride*, *O. viridis*, and *P. akinetum*), where most of the genes are located in the LSC region (Pombert et al, 2006a). Second, the difference in length between the 2 SSCs is much smaller than in other Chlorophyta (e.g., in *N. olivacea*, the LSC region is

5.6 times larger than its SSC region) or even Streptophyta (e.g., in *A. thaliana*, the LSC region is 4.7 times larger than its SSC region) (table 4). In this respect, the cp genome of *O. tauri* is more similar to the cp genome of *C. reinhardtii* (Maul et al, 2002) for 2 reasons: 1) the SSCs have almost identical lengths and both contain an almost identical number of genes (81 and 78, respectively) and 2) the IRs, which in both cases cover almost 20% of the genome, contain exactly the same genes, orientated in the same direction.

The distribution of different genes over the LSC and SSC regions is highly conserved, not only in the entire streptophyte lineage (*M. viride* and land plant genomes share essentially the same gene partitioning), but also in the early diverging *N. olivacea*, indicating that the last common ancestor of all chlorophytes featured a gene partitioning very similar to that observed in land plants. In this respect, Pombert et al, (2006a) created an ancestral cp genome based on the genomes of *O. viridis* and *P. akinetum* (both Chlorophyta, belonging to the Ulvophyceae) and compared that with the genome of *N. olivacea*, which is a prasinophyte and can be considered as ancestral to the 2 ulvophyte. They concluded that the LSC region of the ancestral genome of both Ulvophyceae contained only genes characteristic of the LSC region of *N. olivacea* and that the SSC region contained genes usually found in the SSC and LSC region of *N. olivacea*. However, in the *O. tauri* cp genome, the genes are scattered across the LSC and SSC region, and the previous assumption made by Pombert (2006a) holds no longer true for *O. tauri*. Because the Prasinophyceae are not a monophyletic group, it is not surprising that the *O. tauri* cp genome differs significantly from the *N. olivacea* cp genome and that changes in gene partitioning have occurred independently in *O. tauri* from those observed in ulvophycean and chlorophycean algae. With the availability of more cp genomes it will become clearer whether *O. tauri* is an exception to the rule and has undergone specific genome reshuffling or whether different species all have their own independent evolutionary history regarding their cp genome structure.

Also in the cp genome, we looked for the presence of SDRs. Sixty-four repeats larger than 15 bp are present, but none of the detected repeats exceed the length of 25 bp. Almost all these SDRs are located in the coding region of 5 protein-coding genes (*rpl23*, *psbD*, *psaB*, *psaA*, and *psbA*) and 5 tRNAs (fig. 5). The GC content of the SDRs is comparable to the overall GC content of the cp

Table 6. Comparison of gene content in green algal and land plant chloroplast genomes

	Ov	Pa	So	Cr	Cv	No	Ot	Mv	At	Nt
Photosystem I										
psaA	*	*	*	*	*	*	*	*	*	*
psaB	*	*	*	*	*	*	*	*	*	*
psaC	*	*	*	*	*	*	*	*	*	*
psaI	*	*			*	*	*	*	*	*
psaJ	*	*	*	*	*	*	*	*	*	*
psaM	*	*			*		*	*	*	
Photosystem II										
psbA	*	*	*	*	*	*	*	*	*	*
psbB	*	*	*	*	*	*	*	*	*	*
psbC	*	*	*	*	*	*	*	*	*	*
psbD	*	*	*	*	*	*	*	*	*	*
psbE	*	*	*	*	*	*	*	*	*	*
psbF	*	*	*	*	*	*	*	*	*	*
PsbG									*	
psbH	*	*	*	*	*	*	*	*	*	*
psbI	*	*	*	*	*	*	*	*	*	*
psbJ	*	*	*	*	*	*	*	*	*	*
psbK	*	*	*	*	*	*	*	*	*	*
psbL	*	*	*	*	*	*	*	*	*	*
psbM	*	*	*	*	*	*		*	*	*
psbN	*	*	*	*	*	*	*	*	*	*
psbT	*	*	*	*	*	*	*	*	*	*
psbZ (ycf9)	*	*	*	*	*	*	*	*	*	*
Cytochromeb6/F										
petA	*	*	*	*	*	*	*	*	*	*
petB	*	*	*	*	*	*	*	*	*	*
petD	*	*	*	*	*	*		*	*	*
petG	*	*	*	*	*	*	*	*	*	*
petL	*	*	*	*	*	*		*	*	*
petN						*		*	*	*
ATP synthase										
atpA	*	*	*	*	*	*	*	*	*	*
atpB	*	*	*	*	*	*	*	*	*	*
atpE	*	*	*	*	*	*	*	*	*	*
atpF	*	*	*	*	*	*	*	*	*	*
atpH	*	*	*	*	*	*	*	*	*	*
atpI	*	*	*	*	*	*	*	*	*	*
Chlorophyl biosynthesis										
chlB	*		*	*	*	*		*		
chlI	*	*		*	*	*		*		
chlL	*		*	*	*	*		*		
chlN	*		*	*	*	*		*		
Rubisco										
rbcL	*	*	*	*	*	*	*	*	*	*
RNA polymerase										
rpoA	*	*	*	*	*	*	*	*	*	*
rpoB	*	*	*	*	*	*	*	*	*	*
rpoC1a				*						
rpoC1b				*						
rpoC1	*	*	*		*	*	*	*	*	*
rpoC2	*	*	*	*	*	*	*	*	*	*
Ribosomal RNAs										
23S	*	*	*	*	*	*	*	*	*	*
16S	*	*	*	*	*	*	*	*	*	*
5S	*	*	*	*	*	*	*	*	*	*
TRNAs	25	28	27	31	33	32	27	37	37	30
rnpB						*				

	Ov	Pa	So	Cr	Cv	No	Ot	Mv	At	Nt
NADH oxidoreductase										
ndhA - I						*		*	*	*
ndhJ								*	*	*
ndhK						*		*		*
LSU ribosomal proteins										
rpl2	*	*	*	*	*	*	*	*	*	*
rpl5	*	*	*	*	*	*	*	*		
rpl12	*	*	*		*	*				
rpl14	*	*	*	*	*	*	*	*	*	*
rpl16	*	*	*	*	*	*	*	*	*	*
rpl19	*	*			*	*		*		
rpl20	*	*	*	*	*	*	*	*	*	*
rpl22								*	*	*
rpl23	*	*	*	*	*	*	*	*	*	*
rpl32	*	*			*	*	*	*	*	*
rpl33								*	*	*
rpl36	*	*	*	*	*	*	*	*	*	*
SSU ribosomal proteins										
rps2	*	*	*	*	*	*	*	*	*	*
rps3	*	*	*	*	*	*	*	*	*	*
rps4	*	*	*	*	*	*	*	*	*	*
rps7	*	*	*	*	*	*	*	*	*	*
rps8	*	*	*	*	*	*	*	*	*	*
rps9	*	*	*	*	*	*	*	*		
rps11	*	*	*		*	*	*	*	*	*
rps12	*	*	*	*	*	*	*	*	*	*
rps14	*	*	*	*	*	*	*	*	*	*
rps15								*	*	*
rps16								*	*	*
rps18	*	*	*	*	*	*	*	*	*	*
rps19	*	*	*	*	*	*	*	*	*	*
Translation factors										
infA	*	*	*		*	*	*	*		*
tufA	*	*	*	*	*	*	*	*		
Division										
ftsI						*		*		
ftsW						*		*		
minD	*	*			*	*		*		
minE					*					
Miscellaneous proteins										
accD	*	*			*	*			*	*
cemA	*	*	*	*	*	*		*	*	*
clpP	*	*	*	*	*	*	*	*	*	*
ccsA	*	*	*	*	*	*		*		*
cysA					*	*		*		
cysT					*	*		*		
I-CvuI					*					
Conserved proteins										
ycf1	*	*	*		*	*	*	*	*	*
ycf2					*	*	*	*	*	*
ycf3	*	*	*	*	*	*	*	*	*	*
ycf4	*	*	*	*	*	*		*	*	*
ycf5					*				*	
ycf6/petN						*		*	*	
ycf9/psbZ	*	*	*	*	*	*	*	*	*	*
ycf10/cemA					*				*	
ycf12	*	*	*	*	*	*	*	*		

Cr: *Chlamydomonas,* Cv: *Chlorella* , No: *Nephroselmis* , Ot: *Ostreococcus* , Ov: *Oltmannsiellopsis* , So: *Scenedesmus* , Pa: Pseudendoclonium, Mv: *Mesostigma* , At: *Arabidopsis* , Nt: *Nicotiana*

names indicate

* present

genome (38% for the SDRs vs. 39.9% for the cp genome). The number of SDRs in *N. olivacea* is similar, but substantially differs from *C. reinhardtii*, which cp genome is more similar to the *O. tauri* cp genome regarding its

structure (see above). In the *O. tauri* cp genome, no direct link can be made between the major reshuffling that took place and the abundance of SDRs, whereas for *C. reinhardtii* the major rearrangements could be explained by the huge collection of SDRs present in its cp genome. Consequently, another mechanism is probably responsible for the large number of rearrangements present in the cp genome of *O. tauri*.

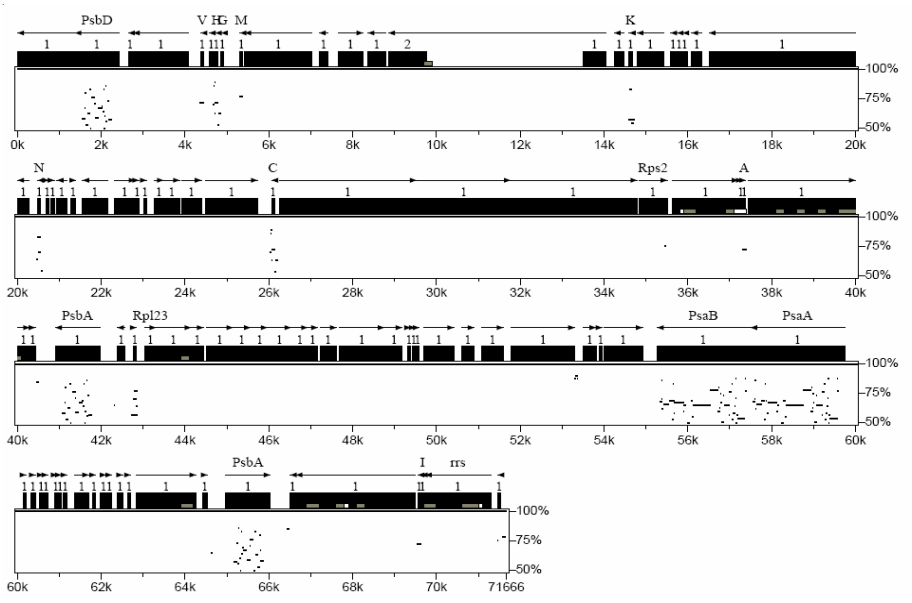


Figure 5. Location of the SDRs in the *Ostreococcus tauri* cp genome. PIPMAKER was used to align the *O. tauri* cp genome against itself, thereby visualising regions containing SDRs by clusters of dots. Genes and their polarities are indicated by horizontal arrows, while their coding sequences are black boxes. Only genes containing SDRs are indicated. Similarities between aligned regions are shown as an average percent identity.

CONCLUSION

Ostreococcus tauri is the smallest eukaryotic organism known to date, and recently, its small (12.56 Mb), but gene dense nuclear genome has been described (Derelle et al, 2006). Here, we present its mt and cp genome, which makes *O. tauri* one of the very few green lineage organisms for which the 3 genome sequences are available. The 2 *O. tauri* organellar genomes are small and display both common and special features compared with their closest relatives.

The main difference between the *O. tauri* and the other Chlorophyta mt genomes is the presence of a unique duplication, previously unobserved in the Chlorophytae. On the other hand, the mt genome of *O. tauri*, which is the most gene dense among all known green algae, closely resembles the one of *Nephroselmis olivacea*, another member of the Prasinophyceae. This is illustrated by a number of common characteristics: 1) the gene content is almost identical in both genomes; 2) there is a high degree of synteny between the 2 genomes, which is illustrated by the presence of a number of conserved gene blocks and by a low number of gene inversions necessary to transform the *O. tauri* gene structure into the one of *N. olivacea*; and finally 3) Pombert (2006b) showed that there is an increase in the number of Short Dispersed Repeats (SDR) when moving in the tree from *N. olivacea* to the more derived lineages within the Chlorophyta. These analyses were confirmed by *O. tauri*, which contains even fewer SDRs than *N. olivacea*. All these data clearly show that the mt genome of *O. tauri* shares the "ancestral" pattern of evolution typified by the *N. olivacea* genome. This conclusion for *N. olivacea* representing an ancestral state (Turmel et al, 1999b) was based on its basal phylogenetic position in the chlorophyte lineage, on the presence of 3 genes (*nad10*, *rpl14*, and *rnpB*) that had not been identified at that time in any other mt genome (today, *rpl14* is also identified in *P. akinetum*), and on its ancestral organizational pattern. These arguments also hold for the *O. tauri* mt genome, and most likely, both the *O. tauri* and *N. olivacea* mt genome represent the most ancestral form known to date for the green lineage. Whether the unique duplication seen in *Ostreococcus* is restricted to this organism will hopefully become clear with the availability of more mt genomes of basal green algae (e.g., the one of *Micromonas pusilla*,

another prasinophyte which is currently being sequenced; Worden A, personal communication).

The *O. tauri* cp genome is very compact, and both the genome size and the gene number are the smallest known among the green plants and green algae. Looking at the gene content, the *O. tauri* cp genome lost many genes compared with other prasinophyte green algae or to *M. viride*. This is well illustrated by the small number of ancestral gene clusters still present in the *O. tauri* cp genome where only 7 of the 24 *Mesostigma/Nephroselmis* gene clusters (de Cambiaire et al, 2006) could be uncovered. Finally, although gene partitioning among LSC and SSC regions is well conserved in all Streptophyta and early-diverging Chlorophyta, the genes in the *O. tauri* cp genome are randomly distributed between both regions. All these data strongly suggest that, in contrast to its mt genome, the *O. tauri* cp genome seems to have lost most of the ancestral features observed in the *M. viride* and *N. olivacea* genomes.

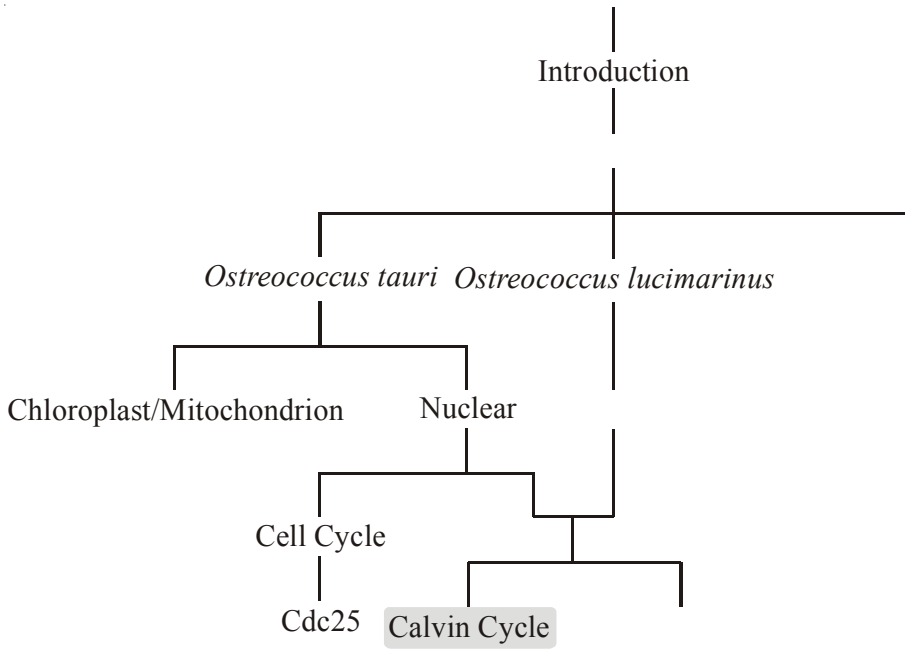
SUPPLEMENTARY MATERIALS

The genome data have been submitted to the European Molecular Biology Laboratory, www.embl.org (accession numbers CR954200 [mt genome] and CR954199 [cp genome]) or can be found at http://bioinformatics.psb.ugent.be/ Supplementary tables and figures are available at Molecular Biology and Evolution online (http://www.mbe.oxfordjournals.org/).

ACKNOWLEDGEMENTS

We would like to thank Sasker Grootjans for his help in the phylogenetic analyses, Jeroen Raes for discussions, Yvan Saeys for help with the figures, and Igor Grigoriev, Brian Palenik, and the Joint Genome Institute for prior access to the *O. lucimarinus* data. S.R. is indebted to the Institute for the Promotion of Innovation by Science and Technology in Flanders for a predoctoral fellowship. This work was supported by the Génopole Languedoc-Roussillon and the French research ministry, and was conducted within the framework of the "Marine Genomics Europe'' European Network of Excellence (GOCE-CT-2004-505403).

Chapter 6


Introduction
Ostreococcus tauri
Ostreococcus lucimarinus
Chloroplast/Mitochondrion
Nuclear
Cell Cycle
Cdc25
Calvin Cycle

J Mol Evol 64, 601-604, 2007*

Unique Regulation of the Calvin Cycle in the Ultrasmall Green Alga *Ostreococcus*

Steven Robbens[1,2*], Jörn Petersen[3*], Henner Brinkmann[4], Pierre Rouzé[1,5] and Yves Van de Peer[1,2,#]

[1] Department of Plant Systems Biology, VIB, B-9052, Ghent, Belgium
[2] Department of Molecular Genetics, Ghent University, B-9052, Ghent, Belgium
[3] Institut für Genetik, Technische Universität Braunschweig, D-38106 Braunschweig, Germany
[4] Département de Biochimie, Université de Montréal, C.P. 6128, Montréal, Canada
[5] Laboratoire Associé de l'INRA (France), Ghent University, Technologiepark 927, B-9052 Ghent, Belgium

[*] Steven Robbens and Jörn Petersen have participated equally to this work
[#] Correspondence to: yves.vandepeer@psb.ugent.be

Key Words: *Ostreococcus tauri*, *Ostreococcus lucimarinus,* Plant evolution, Glyceraldehyde-3phosphate dehydrogenase, CP12, Calvin cycle

Abstract

Glyceraldehyde-3-phosphate dehydrogenase (GapAB) and CP12 are two major players in controlling the inactivation of the Calvin cycle in land plants at night. GapB originated from a GapA gene duplication and differs from GapA by the presence of a specific C-terminal extension that was recruited from CP12. While GapA and CP12 are assumed to be generally present in the Plantae (glaucophytes, red and green algae, and plants), up to now GapB was exclusively found in Streptophyta, including the enigmatic green alga *Mesostigma viride*. However, here we show that two closely related prasinophycean green algae, *Ostreococcus tauri* and *Ostreococcus lucimarinus*, also possess a *GapB* gene, while *CP12* is missing. This remarkable finding either antedates the *GapA/B* gene duplication or indicates a lateral recruitment. Moreover, *Ostreococcus* is the first case where the crucial CP12 function may be completely replaced by GapB-mediated GapA/B aggregation.

Short Communication

During photosynthesis, plastids of land plants and algae transform light energy into ATP and NADPH. This chemical energy fuels the Calvin cycle, where carbon dioxide gets fixed to produce sugar compounds that are used for fatty acid, isoprenoid, and amino acid synthesis (Bassham, 2003). Following the circadian light/dark rhythm, chloroplasts switch between anabolic and catabolic metabolism exemplified by starch production and degradation. A general metabolic transition in green plants is the inactivation of the reductive Calvin cycle and the activation of the oxidative pentose phosphate pathway (OPPP) for NADPH generation at night (Klein, 1986; Schnarrenberger et al, 1995; Martin and Herrmann, 1998; and Michels et al, 2005). Especially the thioredoxin system is responsible for the reversible redox regulation of the Calvin cycle and it is mediated by a small regulator named CP12 (Wedel et al, 1997). The nuclear-encoded CP12 protein is 75 amino acids long, contains at least two crucial cysteine residues (Pohlmeyer et al, 1996 and Petersen et al, 2006), and, together with glyceraldehyde-3-phosphate dehydrogenase (GAPDH; GapA) and phosphoribulokinase (PRK), oligomerizes into a stable protein complex (Cerff, 1979 and Wedel et al, 1997). This mechanism completely blocks the whole cycle at night and is assumed to be generally conserved in cyanobacteria and Plantae, comprising glaucophytes, rhodophytes, chlorophytes, and land plants (Wedel and Soll, 1998 and Petersen et al, 2006). However, land plants contain an additional inactivation complex. Ordinary GapA is redox-insensitive, but a duplicate named GapB recruited the regulatory redox domain, a characteristic C-terminal extension, from *CP12* by a gene fusion (Pohlmeyer et al, 1996). As a consequence, the Calvin cycle inactivation is tightened by a second mechanism that is exclusively based on GAPDH association (Scheibe et al, 2002).

Recently, Petersen et al. (2006) determined the sequence for *GapA*, *GapB*, and *CP12* of different green plants. Their analyses revealed that *GapB* sequences can be unequivocally identified by the CTE as well as a specific sequence pattern including two insertions. Petersen and coworkers identified *GapA* and *GapB* sequences from several charophytes, but especially the presence of *GapB* within the unicellular green alga *Mesostigma viride* dates the *GapA/B* gene duplication at least to the common ancestor of all Streptophyta,

about 700 to 1150 million years ago (mya) (Yoon et al, 2004). Comprehensive analyses of all orders of Chlorophyta, representing prasino-, trebouxio-, ulvo-, and chlorophycean species, including the most prominent and completely sequenced green alga, *Chlamydomonas reinhardtii*, exclusively uncovered *GapA* sequences. Phylogenetic analyses showed *GapA* sequences of Chlorophyta to form a weakly supported group diverging prior to the distinct *GapA* and *GapB* subtrees of streptophytes, hence allocating the *GapA/B* gene duplication to an early stage of streptophycean evolution.

Here we present the distribution of nuclear-encoded plastid *GAPDH* and *CP12* genes from the first completely sequenced prasinophycean green alga, *Ostreococcus tauri* (Derelle et al, 2006), and the closely related strain, *Ostreococcus lucimarinus* (Brian Palenik, personal communication). This unicellular green alga is the smallest free-living eukaryote known to date (Courties et al, 1994 and 1998) and has a genome size of 12.56 Mb, distributed over 20 chromosomes (Derelle et al, 2006). Two genes with high sequence similarity to known *GapA/B* genes could be detected in both *Ostreococcus* species. A typical *GapA* sequence is located on chromosome 10, shows 72% amino acid identity to the *GapA* sequence of *Chlamydomonas reinhardtii*, another member of the Chlorophyta, and clusters within the *GapA* subtree of Chlorophyta (fig. 1). Unexpectedly, the second GAPDH homologue, which is located on chromosome 1, seems to be a genuine *GapB*. It contains the typical C-terminal extension (CTE) with two regulatory cysteine residues (fig. 2) and exhibits the *GapB* specific sequence pattern including two characteristic insertions (fig. 3). Even if the statistic support is weak (fig. 1), in particular, the latter observation authenticates the common origin of *GapB* and rules out CTE recruitment via a second independent gene fusion between *GapA* and *CP12* duplicates. However, the presence of *GapB* in a prasinophyte is surprising, because this gene could previously not be identified in any chlorophyte (prasino-, trebouxio-, ulvo- or chlorophyceae [Petersen et al, 2006]), and it is definitely absent from the completely sequenced chlorophycean genomes of *Chlamydomonas* and *Volvox* (http://www.jgi.doe. gov/).

At least two scenarios can explain the presence of *GapB* in *Ostreococcus*. First, the *GapA/B* gene duplication may have occurred much earlier in green plant evolution than previously assumed (Petersen et al, 2006). Since Chlorophyta

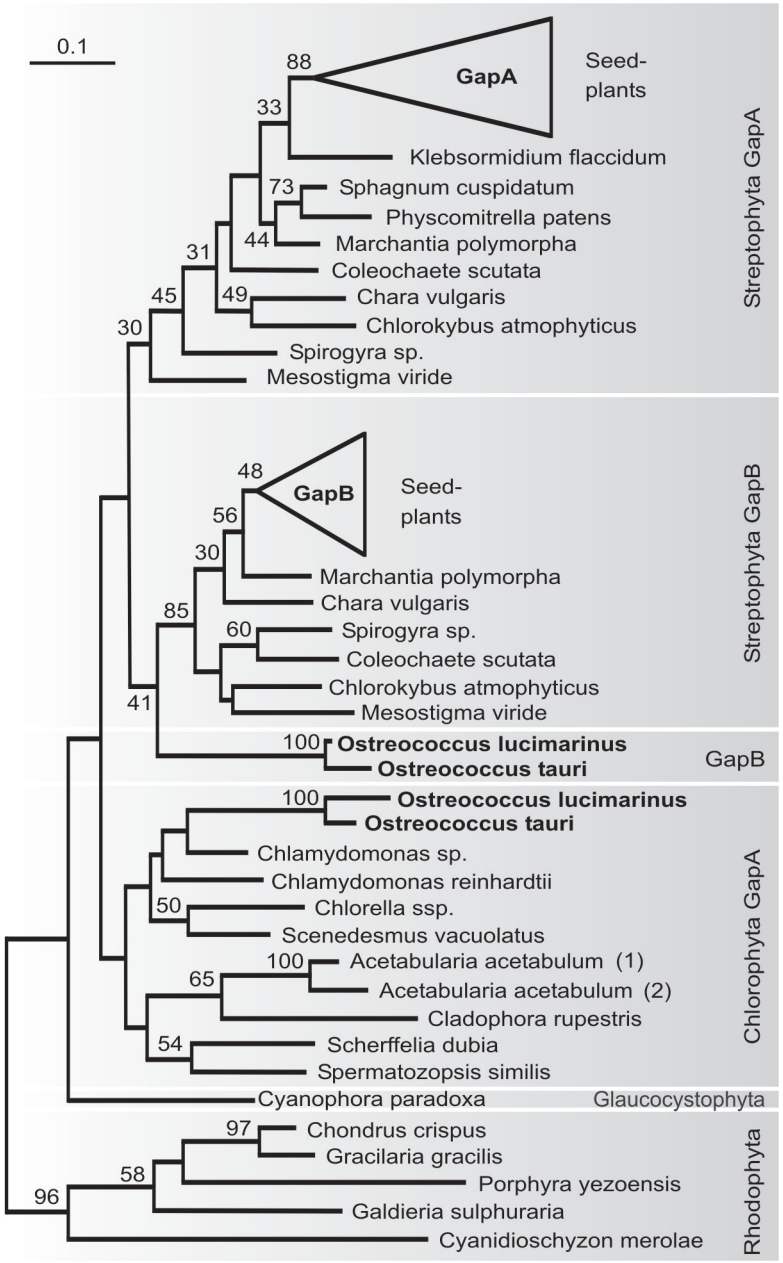


Figure 1. The best maximum likelihood tree based on 71 plastid GAPDH sequences of Plantae and 326 amino acid positions inferred by the program Treefinder under a WAG+ G4 model. Numbers given at internal nodes correspond to nonparametric bootstrap values (100 replicates). Bootstrap values lower than 30% are not indicated. The new GapA and GapB sequences from the two *Ostreococcus strains* are shown in boldface.

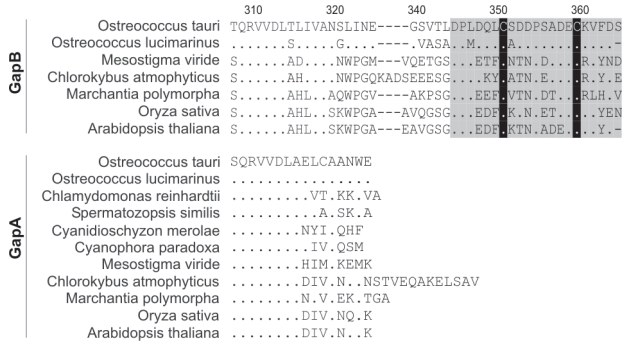


Figure 2. Multiple sequence alignment of the C-terminal end of GapA and GapB proteins from plants. Black shaded boxes indicate the regulatory cysteine residues while the gray shaded box indicate the GapB specific CTE; dashes (-) indicate gaps; dots (.) indicate conserved amino acid positions. Numbering is based on the protein sequence of *Geobacillus stearothermophilus*.

and Streptophyta form two deep and distinct green lineages, and prasinophytes represent the most ancient lineage of the former clade (Rodríguez-Ezpeleta et al, 2007), the *GapA/B* gene duplication would have occurred in a common ancestor of present-day chloro- and streptophytes (Viridiplantae). This premise would imply secondary losses of *GapB* in chlorophyceae (e.g., *Chlamydomonas*, *Volvox*), but also in ulvo- and trebouxiouphyceae, where this gene has not been detected so far (Petersen et al, 2006). In addition, in its simpliest version (one duplication event) it would demand the monophyly of all green plant *GapA* sequences, thus suggesting a phylogenetic artifact in the current tree. Second, it cannot be excluded that an ancestor of *Ostreococcus* recruited the *GapB* via horizontal gene transfer (HGT), for instance, from a charophycean green alga. Mixotrophy has been reported for some prasinophytes (Graham and Wilcox, 1999) and a certain proportion of *Ostreococcus* genes is closely related to marine algae and not to green plants as one would expect. A striking example is the nuclear-encoded Calvin cycle sedoheptulose-1,7-bisphosphatase (SBP; *O. tauri*, chromosome 3; accession no. CAL53197 [wrongly annotated FBP]), which is closely related to the SBP of the diatom *Phaeodactylum tricornutum* (data not shown). Moreover, a unique finding is the replacement of the cytosolic GAPDH (*GapC*), one of the most prominent housekeeping genes that is otherwise universally present in Plantae, Metaozoa, and Fungi, by a *gap3* gene (*O. tauri*, chromosome 2; accession no. CAL52398), which was previously

exclusively identified from bacteria and diplonemids (Figge and Cerff, 2001 and Qian and Keeling 2001). If the assumption of HGT were also true for *GapB*, the evolutionary rate of the *Ostreococcus GapB* sequences might have been accelerated in the context of recruitment (fig. 1), resulting in an artifactual basal position (Brinkmann et al, 2005). Taken together, the discovery of additional *GapB* genes within more distantly related chlorophytes would substantiate the former scenario, whereas a sporadic occurrence in *Ostreococcus* would support the HGT explanation in accordance with a mixotrophic lifestyle.

		GapB specific				
		Insertions		Positions		
		143	252	103	245	250
GapB	Ostreococcus tauri	N	T	P	N	G
	Ostreococcus lucimarinus	N	T	P	N	G
	Mesostigma viride	N	T	P	N	G
	Chlorokybus atmophyticus	A	T	P	N	G
	Marchantia polymorpha	N	T	P	N	G
	Oryza sativa	N	T	P	N	G
	Arabidopsis thaliana	N	T	P	N	G
GapA	Ostreococcus tauri	-	-	P	Q	T
	Ostreococcus lucimarinus	-	-	P	Q	T
	Chlamydomonas reinhardtii	-	-	V	T	T
	Spermatozopsis similis	-	-	A	Q	T
	Cyanidioschyzon merolae	-	-	D	Q	T
	Cyanophora paradoxa	-	-	P	Q	T
	Mesostigma viride	-	-	A	Q	T
	Chlorokybus atmophyticus	-	-	A	Q	T
	Marchantia polymorpha	-	-	E	Q	T
	Oryza sativa	-	-	D	Q	T
	Arabidopsis thaliana	-	-	E	Q	T

Figure 3. GapB specific insertions and positions compared to GapA sequences from different green algae and land plants. The numbering of the conserved positions and insertions is performed with respect to the sequence of *Bacillus stearothermophilus*. Dashes (-) indicate gaps. Both *Ostreococcus* strains are highlighted with light shaded boxes.

Apart from the presence of *GapB*, the investigation of both *Ostreococcus* species also revealed that the *CP12* genes are absent from their genomes. Comprehensive BLAST analyses yielded two weak hits with the N- and C-terminal domain of CP12, located on chromosomes 17 and 11, respectively. Thus, it can be ruled out that these sequences belong to one *CP12* gene that is separated by introns. Since *CP12* was previously assumed to be generally present in cyanobacteria and Plantae (Pohlmeyer et al, 1996 and Petersen et al, 2006), its absence from both *Ostreococcus* genomes might have drastic consequences for GAPDH inactivation as well as plastid metabolism.

Cyanobacterial *CP12* knockout mutants accordingly show significantly reduced growth rates (Tamoi et al, 2005). The lack of *CP12* in complex algae such as diatoms (Armbrust et al, 2004), which obtained their plastids through secondary endosymbiosis, correlates with the absence of the plastid oxidative pentose phosphate pathway (OPPP), probably due to the missing inactivation of the Calvin cycle at night, which would result in futile cycling (Michels et al, 2005 and Petersen et al, 2006). We analyzed the distribution of glucose-6-phosphate dehydrogenase (G6PDH), a key enzyme of the OPPP (Martin and Herrmann 1998), in *Ostreococcus* and identified a single gene for the respective plastid protein (in both species; data not shown). If OPPP and Calvin cycle are present in the chloroplasts of these ultrasmall algae, the presence of GapB is probably essential to ensure GAPDH aggregation at night. Thus, in contrast to streptophytes that harbor two regulatory complexes based on CP12 and GapB, the prasinophyte *Ostreococcus* would be the first example where the Calvin cycle is exclusively inactivated by the formation of GapAB complexes.

METHODS.

Homologous *Ostreococcus* sequences were identified using BLAST (Altschul et al, 1990) and added to the dataset of Petersen et al, (2006). The candidate gene products were manually added to the existing dataset using the EDIT option of the MUST package (Philippe, 1993). Manual annotation was performed with ARTEMIS (Rutherford et al, 2000). In the final alignments, HMMer (Eddy, 1998) was used to generate specific profiles for each gene family with hidden Markov models.

Phylogenetic analyses.

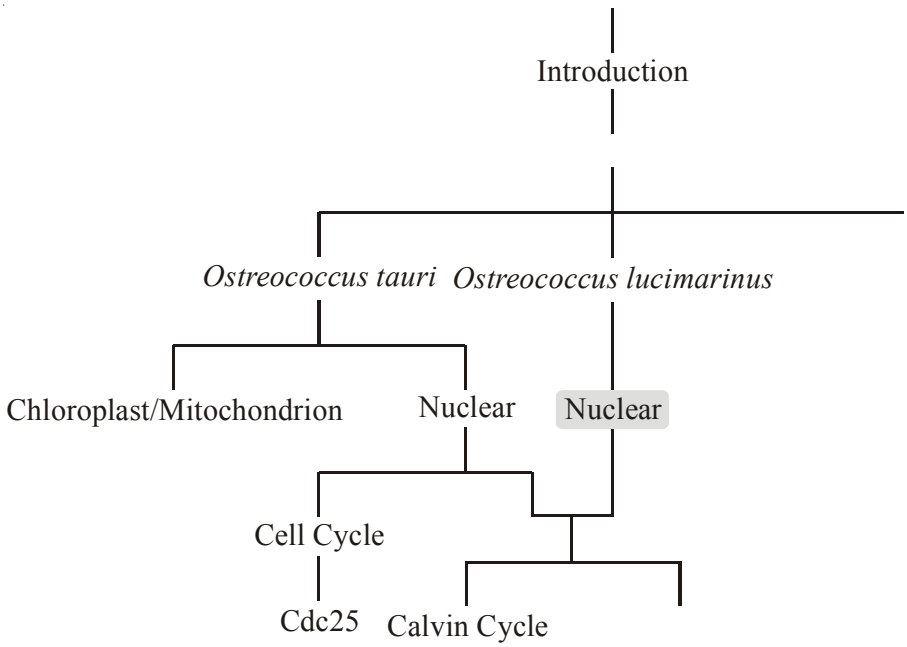
All new sequences reported in this letter have been submitted to GenBank under the following accession numbers: *Ostreococcus tauri GapA* and *GapB* (DQ649076 and DQ649078) and *Ostreococcus lucimarinus GapA* and *GapB* (DQ649077 and DQ649079). The final alignment consists of 71 sequences that all belong to the Plantae with the red algae sequences as outgroup. G-blocks was used to eliminate all ambigiously aligned positions resulting in a dataset with 326 amino acid positons (Castresana, 2000). The best maximum likelihood

(ML) tree was obtained using TREEFINDER under a WAG+Γ4 model (Jobb et al, 2004). In order to estimate the statistical support of the internal nodes, nonparametric bootstrapping (Felsenstein, 1985) on 100 replicates was performed in Treefinder using the same model.

ACKNOWLEDGEMENTS

The authors would like to thank Igor Grigoriev, Brian Palenik, and the JGI for the prior access to the *Ostreococcus lucimarinus* data. S.R. is indebted to the Institute for the Promotion of Innovation by Science and Technology in Flanders for a predoctoral fellowship. The authors also want to thank two anonymous reviewers for careful reading of the manuscript and constructive criticisms.

Chapter 7


Introduction
Ostreococcus tauri
Ostreococcus lucimarinus
Chloroplast/Mitochondrion
Nuclear
Nuclear
Cell Cycle
Cdc25
Calvin Cycle

PNAS 104 (18), 7705-7710, 2007*

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

Brian Palenik[1,18], Jane Grimwood[2], Andrea Aerts[3], Pierre Rouzé[4,8], Asaf Salamov[3], Nicholas Putnam[3], Chris Dupont[1], Richard Jorgensen[5], Evelyne Derelle[6], Stephane Rombauts[7,8], Kemin Zhou[3], Robert Otillar[3], Sabeeha S. Merchant[9], Sheila Podell[10], Terry Gaasterland[10], Carolyn Napoli[5], Karla Gendler[5], Andrea Manuell[11], Vera Tai[1], Olivier Vallon[12], Gwenael Piganeau[6], Séverine Jancek[6], Marc Heijde[13], Kamel Jabbari[13], Chris Bowler[13], Martin Lohr[14], Steven Robbens[7,8], Gregory Werner[3], Inna Dubchak[3], Gregory J. Pazour[15], Qinghu Ren[16], Ian Paulsen[16], Chuck Delwiche[17], Jeremy Schmutz[2], Daniel Rokhsar[3], Yves Van de Peer[7,8], Hervé Moreau[6], and Igor V. Grigoriev[3,18]

[1] Scripps Institution of Oceanography, University of California, USA
[2] Joint Genome Institute and Stanford Human Genome Center, USA
[3] U.S. Department of Energy Joint Genome Institute, USA
[4] Associé de l'Institut National de la Recherche Agronomique, France
[5] Department of Plant Sciences, University of Arizona, USA
[6] Observatoire Océanologique, Laboratoire Arago, France
[7] Department of Plant Systems Biology, VIB, B-9052, Ghent, Belgium
[8] Department of Molecular Genetics, Ghent University, Ghent, Belgium
[9] Department of Chemistry and Biochemistry, University of California, USA
[10] Scripps Genome Center, Scripps Institution of Oceanography, USA
[11] Department of Cell Biology, The Scripps Research Institute, USA
[12] Institut de Biologie Physico-Chimique, Paris, France
[13] Département de Biologie, Paris, France
[14] Institut für Allgemeine Botanik, Germany
[15] Program in Molecular Medicine, University of Massachusetts, USA
[16] The Institute for Genomic Research, Rockville, USA
[17] Cell Biology and Molecular Genetics, University of Maryland, USA
[18] Correspondence to: bpalenik@ucsd.edu

Author contribution see appendix page 227

* modified version

Abstract

The smallest known eukaryotes, at 1μm 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.

INTRODUCTION

Phytoplankton living in the oceans perform nearly half of total global photosynthesis (Behrenfeld and Falkowski, 1997). 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 (Hutchinson, 1961). 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 (Courties et al, 1994 and Chrétiennot-Dinet et al, 1995) and have been isolated or detected from samples of diverse geographical origins (Diez et al, 2001; Guillou et al, 2004; Worden et al, 2004; and Countway and Caron, 2006). 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 (Chrétiennot-Dinet et al, 1995). 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 (Guillou et al, 2004 and Rodríguez et al, 2005). Here we report on the gene content, genome organization, and deduced metabolic capacity of the complete genome of *Ostreococcus* sp. strain CCE9901 (Worden et al, 2004), 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 (Derelle et al, 2006). 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 chromosomes (Derelle et al, 2006) (table 1). For comparison here, both genomes were annotated by using the same tools, as described in Methods.

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	8,166	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.

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* (Derelle et al, 2006). 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, and Chr 21) (figs. 1 and 2) are very distinct, not only from the core genome but also between the species.

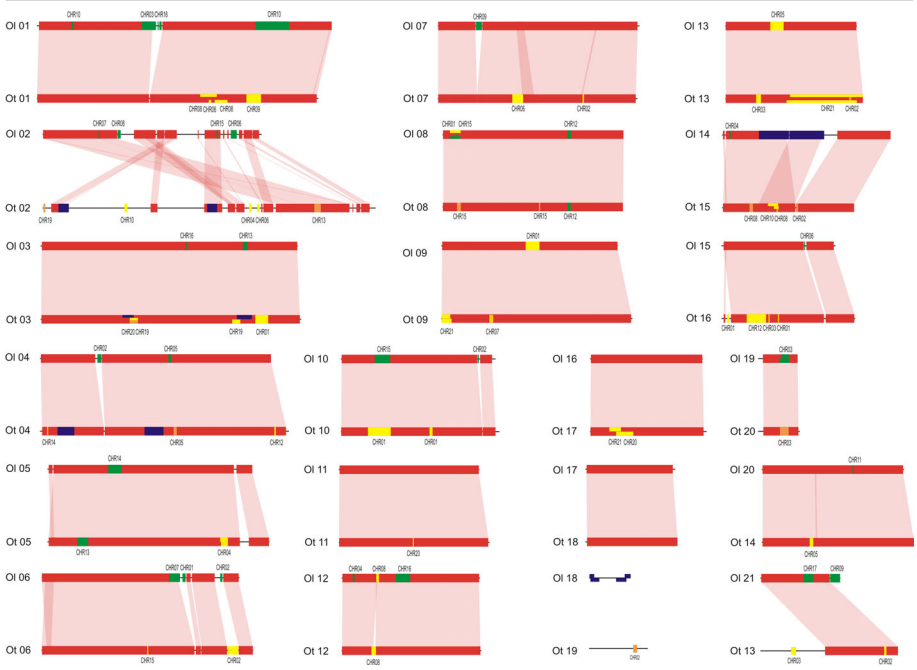


Figure 1. Synteny between the chromosomes of *O. tauri* (Ot) and *O. lucimarinus* (Ol). Depicted areas in red show collinear regions (conserved gene order and content) as described in Methods. Blocks of different colors denote different sorts of duplications: blue, an internally duplicated segment; green,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

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* (González et al, 2002), 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 Derelle et al, 2006), 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* (Derelle et al, 2006) 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.

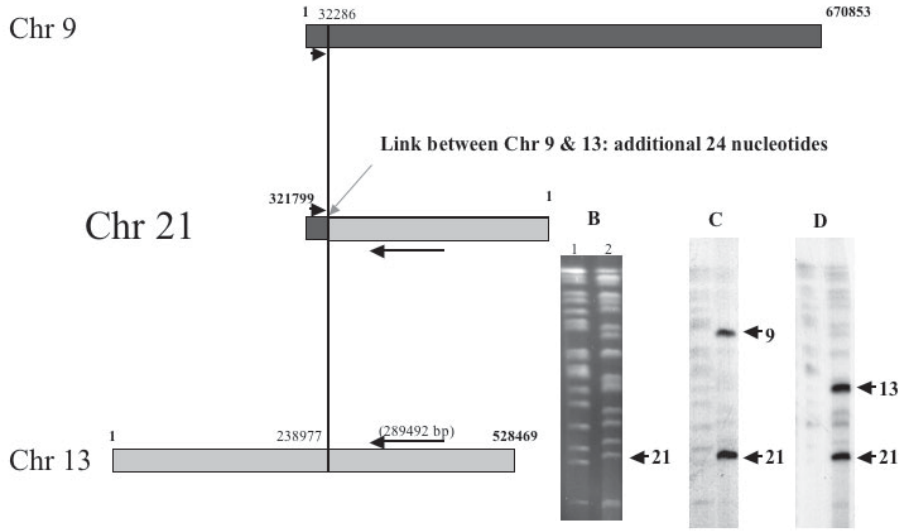


Figure 2. Origin of the new *O. lucimarinus* chromosome, Chr 21. This chromosome was recently formed from pieces of Chr 9 and Chr 13. (A) Map of Chr 21. (B) Pulsed-field gel electrophoresis analysis of the *Ostreococcus* sp. genome migration for 72 h. Lane 1, *O. tauri* genome; lane 2, *O. lucimarinus*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

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 (Frazer et al, 2004) 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 (Palenik et al, 2003). 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 (Palenik et al, 2003), 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* (Derelle et al, 2006). 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.

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 (Koszul et al, 2004), 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.

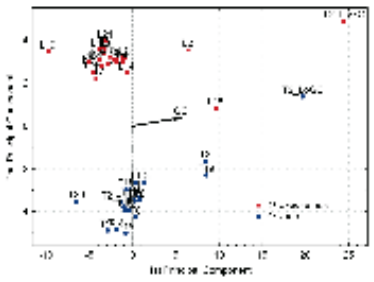


Figure 3. Principal component analysis of *Ostreococcus* genomes.

Most of the characterization of phytoplankton diversity traditionally has focused on pigment and morphological characteristics, and occasionally the utilization of nutrients, for example (Peers and Price, 2006). 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 (Novoselov et al, 2002) 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 peroxidise family (Gladyshev and Kryukov, 2001). 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 (Kim et al, 2006)] 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 (Martin et al, 1994). 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 (Askwith at al, 1996; Armbrust et al, 2004; and La Fontaine et al, 2002). 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 (Curie and Briat, 2003 and Kosman, 2003). 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 Nicontaining SOD, but not a Fe-SOD (Kliebenstein et al, 1998).

Copper concentrations have been shown to affect community composition in coastal ecosystems (Moffett et al, 1997); therefore, it came as some surprise to find that *Ostreococcus* lacks a gene for phytochelatin synthase for ameliorating copper toxicity (Ahner et al, 1995 and Cobbett, 1999). 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 contains 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 (Croft et al, 2005).

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 (Mandal and Breaker, 2004). 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 (Venter et al 2004). 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 (Oliveira and Huynh, 1990). 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 thiolmetabolizing components like CcdA (Kranz et al, 1998). 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 (Hamel et al, 2003), 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.

CONCLUSIONS

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/O*lucimarinus* and www.jgi.doe.gov/O*tauri*. *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 (Worden et al, 2004). 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 Myers (1999) and Aparicio et al (2002). To perform finishing, initial read layouts from the *O. lucimarinus* whole-genome shotgun assembly were converted into our Phred/Phrap/Consed pipeline (Gordon et al, 1988). 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 (2-5x10^7 cells) were agarose-embedded and analyzed by pulsed-field gel electrophoresis as described previously (Rodríguez et al, 2005; Mead et al, 1988; and Wöhl et al, 1995). 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 (Salamov and Solovyev, 2000), homology-based Fgenesh (SoftBerry), Genewise (Birney et al, 2004), MAGPIE (Gaasterland and Sensen, 1996), EST-based estExt (I.V.G., unpublished data), and a combined-approach EuGene (Schiex et al, 2001). 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 (Mulder et al, 2005), and their mappings to Gene Ontology (GOC, 2001), eukaryotic clusters of orthologous groups [KOGs (Koonin et al, 2004)], and KEGG metabolic pathways (Solovyev et al, 2004). 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 (Kent, 2002). 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 (Altschul et al, 1990) ($e<1e^{-8}$). 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 (Simillion et al, 2004). 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 (Frazer et al, 2004) with the constructed genomewide pairwise alignments accessible from http://pipeline.lbl.gov.

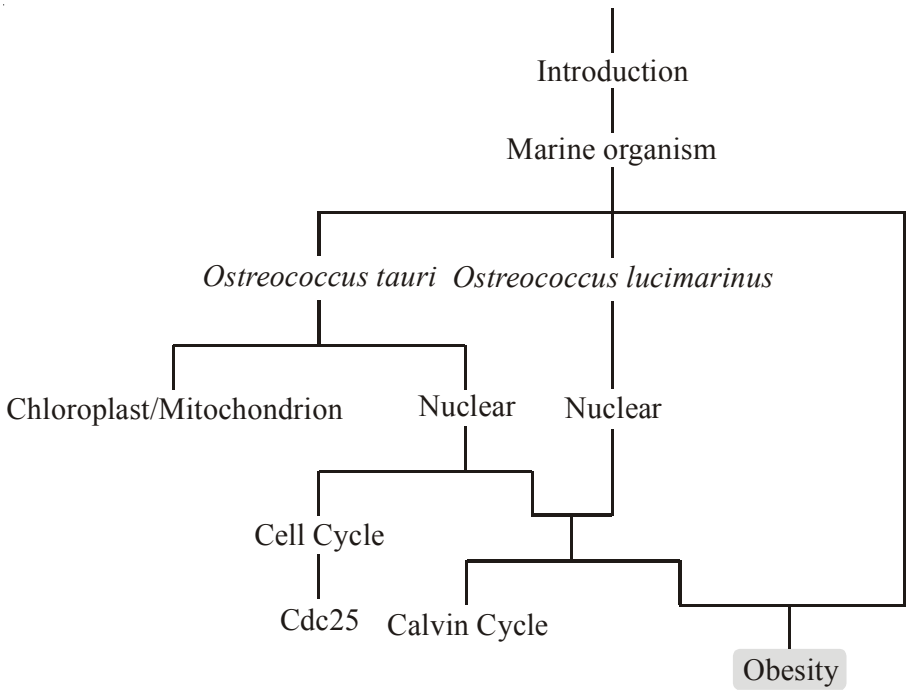
Analysis of Codon Usage.

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 (Rice et al, 2000). Codonfrequency principal components, using correlations, were then calculated with each chromosome as a case and each codon frequency as a variable (Venables and Ripley, 2002). 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.

ACKNOWLEDGEMENTS

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.

Chapter 8


Introduction
Marine organism
Ostreococcus tauri
Ostreococcus lucimarinus
Chloroplast/Mitochondrion
Nuclear
Nuclear
Cell Cycle
Cdc25
Calvin Cycle
Obesity

J Mol Evol, accepted (in press)

The *FTO* gene, implicated in human obesity, is found only in vertebrates and marine algae

Steven Robbens[1,2], Pierre Rouzé[1,3], J. Mark Cock[4], Jürg Spring[5], Alexandra Z. Worden[6], and Yves Van de Peer[1,2,#]

[1] Department of Plant Systems Biology, VIB, B-9052, Ghent, Belgium
[2] Department of Molecular Genetics, Ghent University, B-9052, Ghent, Belgium
[3] Laboratoire Associé de l'INRA (France), Ghent University, B-9052 Ghent, Belgium
[4] UMR 7139 CNRS -UPMC, Végétaux Marins et Biomolécules, Station Biologique, BP74, 29682 Roscoff, France
[5] Institute of Zoology, University of Basel, Klingelbergstrasse 50, CH-4056 Basel, Switzerland
[6] Rosenstiel School of Marine and Atmospheric Science, University of Miami, Miami, FL 33149 USA.

[#] Correspondence to: yves.vandepeer@psb.ugent.be

Key words: FTO, FATSO, obesity, marine algae

Abstract

Obesity is a major societal issue contributing to increased levels of morbidity and mortality. Recently, several studies have demonstrated an association between the *FTO* gene locus and early onset and severe obesity. To date, the *FTO* gene has only been discovered in vertebrates. We identified *FTO* homologs in the complete genome sequences of several evolutionary diverse marine eukaryotic algae, ranging from unicellular photosynthetic picoplankton to a multicellular seaweed. However, *FTO* homologs appear to be absent from all other completely sequenced genomes of plants, fungi and invertebrate animals. Although the biological roles of these marine algal *FTO* homologs are still unknown, these genes will be useful for exploring basic protein features and could hence help unravel the function of the *FTO* gene in vertebrates and its inferred link with obesity in humans.

Obesity is a major societal issue contributing to increased morbidity and mortality, as well as rising health care costs. In 2003-2004, 66% of human population in the USA was classified as overweight (body mass index (BMI) >= 25 kg/m²), and 32% was classified as obese (BMI >= 30 kg/m²) (Ogden et al, 2006). Excessive weight is often associated with an increased risk of several life threatening diseases, including cancer, heart diseases and type 2 diabetes mellitus (Frayling et al, 2007). Unfortunately, the number of obese people continues to increase every day, probably as a result of a modified lifestyle (more food and less exercise). An improved understanding of the genetic basis, and the associated risk factors, is necessary if society is to proactively address this epidemic. Recently, several studies have demonstrated an association between the *FTO* gene locus and early onset and severe obesity in both children and adults (Dina et al, 2007; Field, 2007; Frayling et al, 2007; Frayling, 2007; Groop, 2007; Scott et al, 2007; and Scuteri et al, 2007). *FTO,* also known as FATSO, was originally identified as one of the six genes deleted in the fused toe (*Ft*) mutant mouse (van der Hoeven et al, 1994). Heterozygous animals showed fused toes on their limbs and a thymic hyperplasia, while homozygous mice exhibited a lethal malformation of the developing brain; the embryos lost genetic control of left-right asymmetry; and finally the mice died around the tenth day of their embryonic development (Peters et al, 2002). The *Ft* deletion spans several genes, of which quite a few remain of uncharacterized function. Peters and co-workers (1999) showed that one of these genes, *FTO* (FATSO), which is completely deleted in the *Ft* mutation, is expressed throughout embryonic development and at a high level in most organs in wild type mice. In mouse, this novel gene spans at least 250 kb and encodes a protein of 502 amino acid residues of unknown function. It is still not known whether loss of *FTO* is a causal factor for the phenotype observed in *Ft* mutant mice. Furthermore, no deviations in BMI have been reported in *Ft* mutant mice. However, in human, unlike the associations with BMI initially reported for *GAD*, *ENPPI* and *INSIG2*, which have not been reproduced consistently, association between the *FTO* locus and BMI is strongly supported. Frayling and co-workers (2007) studied almost 40,000 Europeans for variants of the *FTO* gene and identified an obesity risk allele. Depending on the presence of specific single nucleotide polymorphisms (SNPs) in the first intron of *FTO*, individuals weighed 1.2 to 3

kg more and had a 1.67 fold higher rate of obesity than those lacking the risk allele. Similar findings were reported by Dina et al (2007) who studied 2,900 individuals of European ancestry, and potential Type 2 diabetes susceptibility has been correlated with another FATSO intron 1 SNP (Scott et al, 2007).

Until recently, homology searches using the mouse FATSO gene as a query, only recovered sequences from vertebrates. However, with the complete genome sequencing of several marine algae, these results have been dramatically altered. While no clear homolog is found in invertebrate animals, fungi, plants, heterotrophic protists, bacteria or archaea, we identified *FTO* homologs in the genomes of a diverse array of eukaryotic marine algae, ranging from unicellular photosynthetic picoplankton to a multicellular seaweed (Fig. 1). Specifically, *FTO* homologs were retrieved from three species within the Prasinophyceae (*Micromonas pusilla*, *Ostreococcus tauri* and *Ostreococcus lucimarinus*) and two diatom species (*Phaeodactylum tricornutum* and *Thalassiosira pseudonana*), all of which are unicellular, and which represent the only completely sequenced members of their respective lineages. Two copies of the *FTO* homolog were identified in the multicellular brown alga, *Ectocarpus*

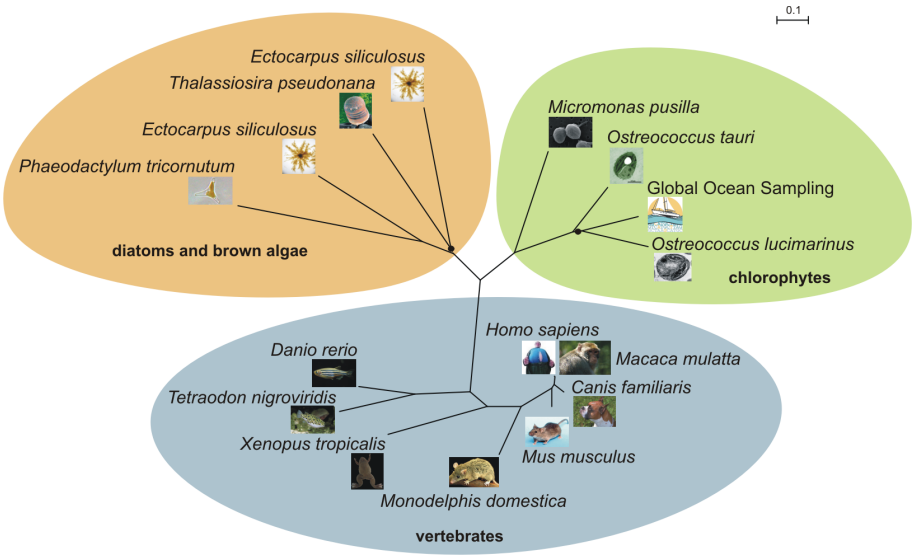


Figure. 1. Maximum likelihood tree showing the distribution of the *FTO* gene. Three major clades can be discerned: the previously described *FTO* genes in the vertebrates; the newly detected genes in diatoms and brown alga; and those of the chlorophytes and GOS sequences. All nodes are highly bootstrap supported (>70%) except two (indicated by a black dot, 50% < BS < 70%).

siliculosus. Furthermore, we scanned the Global Ocean Survey (GOS) dataset (Rusch et al, 2007), and recovered two additional *FTO* genes. These two sequences appear to be derived from the marine prasinophytes, due to high similarity to *FTO* homologs in the prasinophyte genomes supported by the presence of *Ostreococcus* and *Micromonas* 18S rRNA gene sequences in the same GOS sample. Strikingly, all the algae found to harbor *FTO* homologs live in marine environments, given that no *FTO* homologs were recovered from freshwater algae. We performed additional searches for *FTO* in freshwater algae using the *Chlamydomonas reinhardtii* genome sequence (Merchant et al, 2007) but to no avail. We also performed additional searches of the finished genome sequence of the red alga *Cyanidioschyzon merolae*, which thrives in acidic hot springs (Matsuzaki et al, 2004 and Nozaki et al, 2007). Moreover, we performed these searches iteratively, using the newly discovered marine *FTO* sequences as queries, and still detected no homologs in invertebrate animals, fungi, plants, heterotrophic protists, bacteria or archaea, confirming our initial findings.

As mentioned above, the function of *FTO* is still not known. Dina et al. (2007) detected *FTO* expression in 11 out of 11 human tissue types tested, with the highest expression levels being in the hypothalamus, pituitary and adrenal glands. These findings have promoted the hypothesis that *FTO* plays a role in body weight regulation through the hypothalamic-pituitary adrenal axis. *FTO* is also expressed in rat and mouse. EST data indicates that the marine *FTO* homologs in the diatom *P. tricornutum* and the prasinophyte *M. pusilla* are expressed under standard growth conditions. Although the biological roles of the algal *FTO* homologs are still unknown, these genes can be used, together with the vertebrate sequences, to explore basic protein features. Based on primary sequence characteristics, *FTO* proteins are unlikely to be targeted to either membranes or to organelles, but rather are predicted to be globular, cytosolic proteins with mixed alpha/beta structures. Looking at conserved positions shows a drop from 195 positions conserved amongst animal sequences to only 44 conserved over all sequences, likely pinpointing the functionally essential residues. Among the most widely divergent *FTO* sequences, three amino acid residues (W, Y and H) are strikingly over-represented amongst the 44 absolutely conserved positions (see Supplementary Fig. 1). *In silico* predictions indicate

that these residues are more likely to be located at an active site than to be at a protein-protein interface or to be surface interacting residues (Ma et al, 2003). This suggests that *FTO* may have an enzymatic function rather than be involved in protein-protein interactions. Three of the conserved positions have high prediction scores for phosphorylated residues, indicating a potential role for phosphorylation in regulation of *FTO*.

Our findings do not negate the association between *FTO* intron 1 SNPs and obesity. While identification of risk factors has advanced tremendously, for the most part, the functional ramifications of these genetic variations remain uncharacterized. In the case of *FTO*, Frayling and colleagues (2007) raised the alternative hypothesis that the intron 1 SNP might serve to alter regulation of another gene, as opposed to having a specific affect on the product encoded by *FTO*. While risk factors carry value in preventative medicine it is mechanistic knowledge that fosters therapeutic innovation. Why marine algae harbour and express *FTO* is unclear, as is the link with obesity in humans. However, previous studies have demonstrated that algal research can be applied to investigation of vertebrate gene function. For instance, *Chlamydomonas* is often referred to as "the green yeast" because it is an easy to work with eukaryotic model organism which also performs photosynthesis (see Li et al, 2004). None of the highly developed but easy to use (i.e. not involving animal work) model organisms (e.g. *Chlamydomonas*, *Arabidopsis*, Yeast, *Drosophila* and *C. elegans*) possesses an *FTO* gene. Thus, here we identify alternative systems for functional studies, such as the genetically tractable diatom *Phaeodactylum* (Siaut et al, 2007). These in turn will shed light on *FTO* function and, should that function be relevant to vertebrate homologs, thereby streamline research on genetic factors contributing to human obesity.

METHODS

We initially scanned all publicly available non-redundant databases, as well as our in-house data for homologs of the mouse *FTO* gene, using BLASTP (Altschul et al, 1997). Because there was a very clear drop-off in E-value between homologs and non-homologs (significant values from E^{-82} to E^{-27}, then dropping to non-significant E-values of 0.71 or higher), selection of *FTO* homologs was

straightforward. No (distantly related) genes homologous (or partially homologous) to the *FTO* genes could be identified. Next, HMMer (Eddy, 1998) was used to generate a specific profile of the *FTO* gene family with hidden Markov Models, using all available sequences and we searched NCBI EST and genome databases using TBLASTN. However, no new candidate *FTO* genes were detected.

Annotation of the *FTO* gene sequences was manually checked and corrected using ARTEMIS (Rutherford et al, 2000) when necessary. Protein sequences were aligned with CLUSTALW (Thompson et al, 1994) and after manual improvement of the alignments using BIOEDIT (Hall, 1999), only 266 well-aligned positions were taken into account for tree construction. Pairwise distance trees were constructed using TREECON (Van de Peer and Wachter, 1994), based on Poisson-corrected distances, while PHYML 2.4.4 (Guindon and Gascuel, 2003) was used to compute the maximum likelihood tree. Bootstrap analyses with 500 replicates were performed to test the significance of the nodes. Both methods gave identical tree topologies and similar bootstrap support.

ACKNOWLEDGEMENTS

Sequence data of *Phaeodactylum* were produced by the Joint Genome Institute (http://www.jgi.doe.gov/). Sequence data of *Ectocarpus* were produced by Genoscope (http://www.cns.fr/). *Micromonas* culture work and genome sequencing were supported by the USDOE and a Gordon & Betty Moore Foundation grant to AZW. S.R. is indebted to the Institute for the Promotion of Innovation by Science and Technology in Flanders for a predoctoral fellowship.

DATA

Accession numbers:

Ostreococcus lucimarinus: XP_001420808

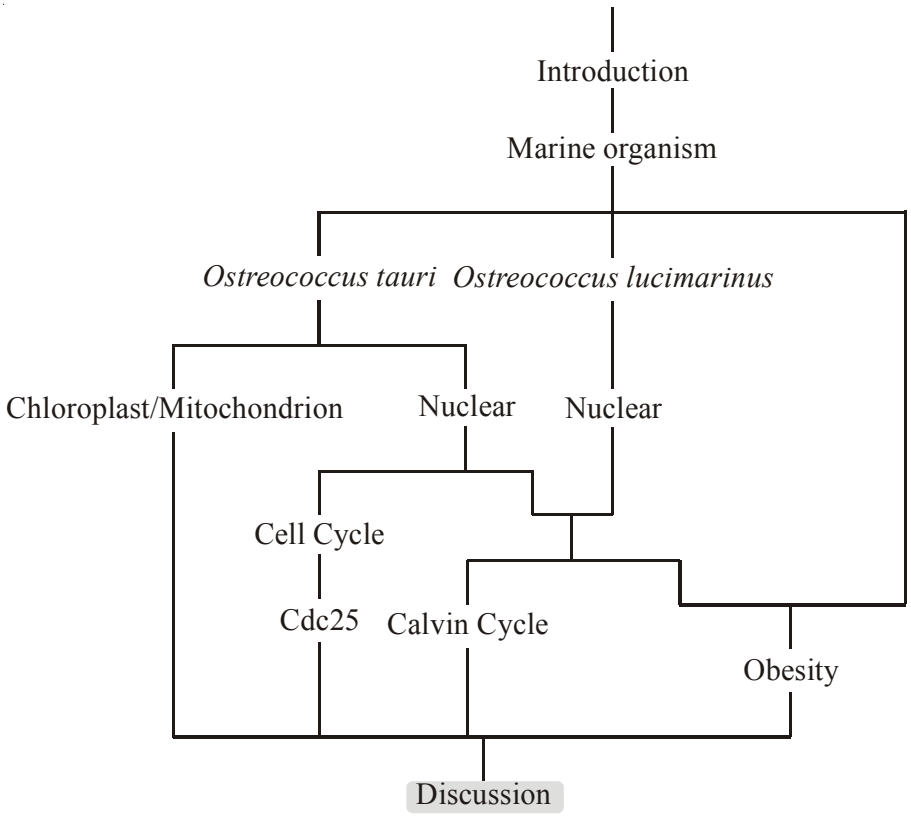
Ostreococcus tauri: CAL57236

Thalassiosira pseudonana: jgi|Thaps3|261481|thaps1_ua_kg.chr_2000305 (http://genome.jgi-psf.org/Thaps3/Thaps3.home.html)

Phaeodactylum tricornutum: jgi|Phatr2|41429|fgenesh1_pg.C_chr_30000044 (http://genome.jgi-psf.org/Phatr2/Phatr2.home.html)

FTO sequences from *Micromonas pusilla* and *Ectocarpus siliculosus* can be obtained from the authors upon request.

Chapter 9


Introduction
Marine organism
Ostreococcus tauri
Ostreococcus lucimarinus
Chloroplast/Mitochondrion
Nuclear
Nuclear
Cell Cycle
Cdc25
Calvin Cycle
Obesity
Discussion

The title of my PhD (Barely visible but highly unique: the *Ostreococcus* genome unveils its secrets) tries to give us the impression that *Ostreococcus* is now a well understood organism, thereby providing us insight into the architecture of its genome and the function of its genes. This is of course partially true, as the two genome sequencing projects unravelled the characteristics of both genomes and (a few) of its special features. However, it also provided us a lot of unanswered questions. Sequencing new genomes will always deliver some species-specific or unique/unseen trends, which serve as food for further discussions and analyses.

Starting this genome project of *Ostreococcus tauri* immediately provided us a first unique/unseen characteristic within the green lineage. The presence of an in silico predicted and experimental validated Cdc25 protein was and still is the first real green lineage Cdc25 dual-specificity phosphatase described (chapter 2). Why it was only present in *O. tauri* (Prasinophyceae) and not in other green algae like *Chlamydomonas reinhardtii* (Chlorophyceae) or higher land plants like *Arabidopsis thaliana* and *Oryza sativa* could not be explained. Today however, previous analyses can be extended by the detection of two other cdc25 orthologs in *O. lucimarinus* and *Micromonas pusilla* (Alexandra Z. Worden, personal communication), two members of the Prasinophyceae. Still, no Cdc25 could be detected in algae belonging to other classes of green algae. The detection of cdc25 orthologs in different green algae clearly shows that Cdc25 appeared before the divergence of the ophistokont (metazoans and fungi) and the green lineage and that it was not a recent acquisition specific to the ophistokont lineage. Why only prasinophyte green algae harbour a Cdc25 and other green algae and land plants don't, is yet not really understood, but different groups, using the available data, tried to reconstruct the history of Cdc25. After publication of the cdc25 ortholog in *O. tauri*, Landrieu and co-workers (2004) published in the same year two papers describing a "true" cdc25 ortholog in the land plant *A. thaliana*, based on its tertiary structure and its ability of binding and activating *in vitro* phosphorylated CDKs. However, great sequence divergence and the absent of the important N-terminal domain (Landrieu et al, 2004), its involvement in arsenate reduction (Bleeker at al, 2006 and Dhankher et al, 2006), and its inability to complement *cdc25-22* yeast mutants (Landrieu et al, 2004), made Boudolf and co-workers (2006) argue

that this isoform might not be a *bona fide* cdc25. In contrast, they postulated an evolutionary replacement of Cdc25 by the plant-specific B-type CDKs. This would implement the speciation of CDKBs (resulted from a gene duplication) and the consequently loss of the cd25 gene. In this evolutionary model, *O. tauri* served as an ideal intermediated organism as it contains both a B-type CDK, having partially the characteristics of an A- and B-type CDK (chapter 3 and Corellou et al, 2005), and a true cdc25. This could indicate that within *O. tauri* cdc25 is still needed as its B-type CDK doesn't contain yet all the typical higher plant B-type CDK characteristics. Whether this evolutionary model is correct or not (no experimental evidence supports this model) the strength of having data from organisms located at different nodes on the tree of life already made an impact.

Above example shows that the choice of which organism's genome will be sequenced is not a light step to take. One can ask why sequencing the genome of a tiny unicellular green alga, which has (at first sight) no economical or biomedical relevance (to humans). In this respect, sequencing the genomes of cereal crops, like rice, maize and wheat, and mammalians like human, mouse and rat, seems more relevant to do. This is of course partially true, because the knowledge that will be gained by sequencing these 'relevant' genomes will (hopefully) have direct influence on healing certain diseases and reducing the food shortage problem in certain countries. However, if we want to fully understand the evolutionary dynamics of these plant (and mammalian) genome organizations and the functional and evolutionary processes that are fundamental to plant life and to plant reproduction, a broader taxonomic sampling is needed. We should not only focus on the economical important plant species, but also concentrate on species that occupy crucial evolutionary nodes within the green lineage. Comparing different genomes located at different places in the green tree can produce vital information needed to fully understand the organization, function and evolution of the plant genome in general.

Besides their phylogenetic position, the genome size should also be taken into account when sequencing new species. Comparing genomes of different sizes can increase our understanding of duplicated and mostly redundant regions within the bigger genome. Their phylogenetic position combined with its genome size, are essential criteria for sequencing the genome of new species. In this

respect, sequencing the genomes of *O. tauri* and *O. lucimarinus* (chapter 4 and 7) agrees with above requirements: as members of the Prasinophyceae they occupy a key position at the base of the green lineage and with a genome size of less than 13 Mb, they possess a highly dense packed and small genome. This reduced genome size stands in big contrast to the genome size of another unicellular green alga, *C. reinhardtii*, which is estimated around 100 Mbp. Having such a small genome resulted in a reduction of overall gene family size, nicely illustrated in chapter 3, where we describe a complete, yet highly reduced set of genes involved in cell cycle control. Initially we stated that each gene family contained only one member, but after the completion of its genome, we had to adjust these numbers: apparently *O. tauri* harbours two D-type cyclins, and besides a WEE1 kinase we also found a Myt1 kinase. Missing these genes was caused by the large amounts of gaps (> 2,000) still present at the time of the initial analyses. As both genes were located at the border of such a contig, we were not able to detect the complete sequence of both genes.

When we compare this updated dataset with the cell cycle genes detected in the large genome of *C. reinhardtii* (Bisova et al, 2005), almost an identical gene set is present. *C. reinhardtii* also contains a highly reduced set of cell cycle genes, including different D-type cyclins, but with the absent of a true Cdc25 ortholog and with only one Wee1 kinase. In contrast, *C. reinhardtii* encodes for six additional genes, which are not present in *O. tauri*. When adding the cell cycle genes of land plants, we see that for *A. thaliana* a lot of its core cell cycle genes are duplicated, leading to a much bigger dataset (Vandepoele et al, 2002). Despite these differences in numbers, the cell cycle proteins found in both algae are remarkably similar to those found in higher plants and metazoans. As these genes, and the pathways they are involved in, are well conserved among all kingdoms, studying them in *Ostreococcus*, where we don't have functional redundancy, will provide the necessary knowledge for studying the cell cycle in other, commercially more important species.

However, sequencing and annotating genomes of green algae is not only useful for gaining knowledge in order to better understand other, more "relevant" species, but they themselves are also becoming more and more important. Understanding the mechanisms behind controlling the cell cycle within green algae for example can provide solutions for an increasing problem the world is facing today: huge

lakes of densely packed algae forming algal blooms. With the ever growing number of people occupying this world, humans have doubled the natural rate of nitrogen input in terrestrial systems, thereby significantly increasing the nitrate levels in all aquatic systems. The result of an excess of nitrogen (and other nutrients) into waters cause increased growth of algae and green plants. As more algae and plants grow, others die. This dead organic matter becomes food for bacteria that decompose it. With more food available, the bacteria increase in number and use up the dissolved oxygen in the water. When the dissolved oxygen content decreases, many fish and aquatic insects cannot survive. This results in a dead area. As these blooms also produce some toxins which have severe biological impacts on wildlife and even humans, the better we understand their genetic make-up, and more specific how these cells multiply, the better we can find solutions to prevent these problems.

Besides these negative side affects of having too many algae concentrated in one place, algae do have ecological and economical importance. Ecologically, seaweeds are essential because they are involved in CO_2 fixation (Calvin Cycle, Chapter 8), they act as one of the primary producers in the marine food chain and they are involved in global cycling of the elements N, O, S, P and C. Economically, green algae seem to manifest themselves more and more in different areas. First, with the increasing oil prices and the need for alternative energy sources, algae can play an important role as deliverers of biofuels (liquid or gas transportation fuel derived from biomass) (Haag, 2007; and Gordon and Polle, 2007). One hectare of algae can produce much more litres of biodiesel (90,000 litres) compared to soya (450 litres), canola (1,200 litres) or oil palm (6,000 litres). Second, algae are becoming more important in producing recombinant therapeutic proteins when acting as protein factories (Mayfield et al, 2003 and 2007; Franklin and Mayfield, 2004; and Mayfield and Franklin, 2005). There are several advantages of using green algae instead of bacteria, yeast, insect or mammalian cell cultures for creating therapeutic proteins: chloroplasts of green algae correctly fold and assemble mammalian proteins, using a minimum length of time; algae can be grown in closed environments without having the risk of contaminating the environment; and many algae are save to eat, so oral delivery of the therapeutic proteins is possible. Finally, chapter 8 describes the presence of a potential obesity-linked gene in *Ostreococcus*

and other marine organisms. While identification of risk factors has advanced tremendously, for the most part, the functional ramifications of these genetic variations remain uncharacterized. Risk factors carry value in preventative medicine while mechanistic knowledge fosters therapeutic innovation. Why marine algae harbour and express FTO is unclear, as is the link with obesity in humans. However, because of the relative ease of work with these algae, future experiments should shed light on FTO function and thereby streamline research on genetic factors contributing to human obesity.

All these important (ecological, economical and biomedical) aspects surrounding green algae should encourage scientists to completely unravel their genomes as they will become more and more important in different aspects of our society. Besides these money- or world related advantages, studying these genomes also provided us some unexpected biological knowledge. With the presence of different forms of heterogeneity within both *Ostreococcus* genomes, we were challenged in two different ways. First we had to be able to overcome the computational problems we encountered while annotating these genomes. Creating different methods and models for the different parts of the chromosomes finally enabled us to map all the structural components present in these genomes. Second, besides the computational know-how we gained, we also had to broaden our biological view on how genomes are or can be build up. Till now we were only able to speculate on the reason why *Ostreococcus* carries these "odd" chromosomes, but none of these speculations (sex chromosome related) can be proved. Hopefully, with the presence of more green genomes, we will be able to explain and comprehend their odd structure.

Generally we can state that sequencing new genomes provides much new, sometimes unexpected information which can lead to some nice publications. The big challenge however lies now in interpreting all this amount of data. Generating new data is one thing, but answering and explaining certain problems is the most difficult part. Experimental, wet lab evidence will always be needed to verify certain hypothesis, but I feel that bioinformatics will become even more important in handling all the already available and in the future coming data. A balanced mix between biologists/biotechnologists and mathematics/physicist will provide the necessary tools to create order in the chaos of A, C, G, and T's.

Future perspectives

As already mentioned, once the sequence is provided and the genes are annotated, we are only half way: creating is one thing, explaining another. Finding biological answers in the newly obtained data is not only the most interesting part, but also the most difficult one. Comparing two genomes already provided us insight into the genetic toolkit of two different, yet closely related organisms (chapter 7), but when also including data from other sequenced and annotated genomes, a much broader picture will be delivered. When we would only focus on highly similar or closely related species, we will not be able to extrapolate our findings towards other member of a certain class or lineage. In this respect, we should not only focus on both *Ostreococcus* species, but also integrate data from species located at different places on the tree of live. By doing this, we will be able to gain insight in the evolution of plant genomes, and more particular in the evolution of gene families across green plants. Performing these types of analyses can be done on two different levels: or we try to get an overall view on gene family evolution/distribution among the green lineage (green algae and land plants) or we specify a bit more and focus only on members of the Chlorophyta.

When looking at the general picture, we see that already several green genomes are completely sequenced and annotated: the green algae *Ostreococcus tauri*, *Ostreococcus lucimarinus* and *Chlamydomonas reinhardtii*, the moss *Physcomitrella patens* (in house data), the monocot *Oryza sativa* and the dicots *Arabidopsis thaliana and Populus trichocarpa*, representing a broad taxonomic sampling. Creating clusters of gene families, based on similar function, will enable us to create phylogenetic profiles. These are profiles describing the presence or absence and number of homologous genes present in the different genomes. Based on these profiles, trends among the different organisms can be spotted: which gene families are conserved among all green species and can be considered as universally important for the green plants. Are these genes present as single copy genes, or has there been some species- or lineage-specific duplications? However, when looking at general gene function conservation, we can expand our dataset and also include representatives of other important eukaryotic lineages: human for the mammalians, *Saccharomyces cerevisae* as

a yeast and *Cyanidioschyzon merolae* as a red alga. This will give us an idea of the number and function of the genes that are universally conserved among eukaryotes. Moreover, as we noticed that *Ostreococcus* harbours certain genes that are not present in other plants but do have an ortholog in human and/or yeast (chapter 2 and 8), this will give us an idea whether these findings were rather unique or whether there exist more of these types of genes families. Overall we will get a more in depth view on gene family distribution among different clades. Besides looking at conservation, trying to identify certain specific features will provide us data on lineage- or even species-specific characteristics: which genes are only present within the green algae and have consequently been lost within the land plants and visa versa. Can we link the presence, absence or expansion of certain gene families to certain morphological differences (uni- versus multi-cellular organisms, water versus land plants, etc.) or to big events that happened in history? Can we draw some conclusions regarding the structure of certain genes within a family: have there been domain fusions or domain rearrangements, have certain genes become longer or smaller during evolution and why did this happen? Finally, classifying the species specific genes (genes only present in one organisms and absent in all other used organisms) will be the hardest part because as they are unique, there is no data available to compare with. One can ask whether these genes are really unique (to more data that will become available, the better we will be able to answer this) or whether they are derived from another organism through horizontal gene transfer. Horizontal gene transfer is well known between prokaryotes and eukaryotes, but recent studies have shown the existence of transfer between two members of the eukaryotes. Phylogenetic analyses combined with taxonomic distribution analyses will probably shed light on the amount and the origin of the transferred genes within the green lineage. As we see, a lot of questions will pop up, but should be able to be answered when using above data.

Besides analysing gene family evolution within the entire green lineage, we can also only focus on the green algae. Today, more than ten new sequencing projects, focusing on green algae, already have been started: four members of the Chlorophyceae, three members of the Trebouxiophyceae and finally five prasinophyte green algae (www.genomesonline.org) are being sequenced. Asking the same questions as above will enable us to get some insight in the

gene family architecture and evolution within the green algae and maybe provide us some answers that we were not able to answer yet: why is there a Cdc25 still present in the prasinophyte *Ostreococcus* and not in other classes of the green algae, why does *Ostreococcus* harbour these heterogeneous chromosomes, why does the mitochondrial genome of *Ostreococcus* contain a duplication, which was only observed in within members of the higher plants, why does the regulation of the Calvin Cycle differ in *Ostreococcus* compared to other green algae and why does the smallest described eukaryotic organisms contain an obesity-linked gene?

All above future perspectives are highly interesting, but I will say goodbye to my green friend, thank him/her for the nice collaboration and focus on a new organism. Merci et adieu.

Summary/samenvatting

English summary

Algae can range in size from tiny single-celled organisms of only 1 micron in diameter to the 65 meter long giant brown kelps, forming underwater forests. In between these two extreme life forms lies a huge morphological diversity: unicellular organisms, branched or unbranched filaments, colonies of loosely packed or highly organized cells and finally complex parenchymatous algae. Depending on the amount and composition of pigments like chlorophyll a, which is needed for photosynthesis, eukaryotic algae can roughly be subdivided into 3 different groups: the red algae (Rhodophyta), the brown algae (Phaeophyta) and the green algae, whereas the latter group is the largest of all. The green algae are considered to be "primitive" plants that gave rise to the huge amount of land plants known to date. They originated around 1,500 million years ago and they diverged from the land plants around 425-490 million years ago. The green algae can be split up into two lineages: one known as the chlorophyte green algae (Chlorophyta), which includes the majority of what have been called green algae, and a second lineage entitled Charophyta, containing a smaller number of green taxa. Finally, a third phylogenetic green plant lineage, the Prasinophyceae constitutes a particularly interesting algal class, holding a basal position in the evolution of the extant green lineage. As a result, today the green lineage can be subdivided into two monophyletic lineages: Streptophyta (land plants and the charophyte green algae) and Chlorophyta (the rest of the green algae, including the prasinophyte green algae). Members of the class Prasinophyceae have long been identified as being scaly green flagellates, but today, species with or without flagella and/or scales and even coccoids have been added to this class. These morphological heterogeneities combined with phylogenetic evidence based on SSU rDNA sequences, led to the conclusion that the Prasinophyceae are a paraphyletic group. In this thesis, we describe the first sequenced and annotated genomes of two members of the Prasinophyceae: *Ostreococcus tauri* and *Ostreococcus lucimarinus*.

O. tauri is a unicellular green alga that was discovered in the Mediterranean Thau lagoon (France) in 1994. With a size less than 1 μm, comparable to that of a bacterium, it is the smallest eukaryotic organism currently described. Its cellular organization is rather simple with a relatively large nucleus, a single chloroplast,

1 mitochondrion, 1 Golgi body, and a highly reduced cytoplasm compartment. A membrane surrounds the cells, but no cell wall can be observed.

The main goal of this thesis was to get insight into the genome organization within the green algae by annotating different genomes of the Chlorophyta and by comparing the annotated genes among members of the green lineage. Before annotating the entire genome, a case study was performed to get a first glimp on how the genes look like within *O. tauri*. As there is quite some knowledge about the cell cycle within our lab, this was consequently the pathway we looked at first. Chapter 2 describes for the first time the presence of a Cdc25 dual specificity phosphatase in the green lineage. The Cdc25 protein phosphatase is a key enzyme involved in the regulation of the G2/M transition during the cell cycle in metazoans and yeast. However, no Cdc25 ortholog has been identified in plants, although functional studies have shown that an activating dephosphorylation of the CDK-cyclin complex regulates the G2/M transition. The Cdc25 detected in *O. tauri* encodes a protein which is able to rescue the yeast *S. pombe cdc25-22* conditional mutant. Furthermore, microinjection of GST-tagged *O. tauri* Cdc25 specifically activates prophase-arrested starfish oocytes. In vitro histone H1 kinase assays and anti-phosphotyrosine Western Blotting confirmed the in vivo activating dephosphorylation of starfish CDK1-cyclinB by recombinant *O. tauri* Cdc25.

When looking at all the core genes involved in the control of the cell cycle, chapter 3 highlights the minimal yet complete set of core cell cycle genes described to date. *O. tauri* has only one homolog of CDKA, CDKB, CDKD, cyclin A, cyclin B, cyclin H, Cks, Rb, E_2F, DP, DEL and Cdc25; and two copies of cyclin D and Wee1/Myt1. Interestingly, the plant specific B-type CDK contains a motif that lays between a CDKA and CDKB type motif. Phylogenetic analyses however still place the *O. tauri* ortholog together with other B-type CDKs.

Once enough genes were manually annotated that could serve as a training set to train gene prediction software, the genome of *O. tauri* was completely automatically annotated (chapter 4). Overall, the 12.56-Mb nuclear genome has an extremely high gene density (8,166 protein-coding genes are spread over 20 chromosomes), in part because of extensive reduction of intergenic regions and other forms of compaction such as gene fusion. However, the

genome is structurally complex. It exhibits previously unobserved levels of heterogeneity for an eukaryote. Two chromosomes differ structurally from the other eighteen. Both have a significantly biased G+C content, and, remarkably, they contain the majority of transposable elements. Many chromosome 2 genes also have unique codon usage and splicing, but phylogenetic analysis and composition do not support alien gene origin. In contrast, most chromosome 19 genes show no similarity to green lineage genes and a large number of them are specialized in cell surface processes.

Besides sequencing its nuclear genome, both the complete nucleotide sequence of the chloroplast and mitochondrial genomes have been determined and manually annotated (chapter 5). The mitochondrial genome assembles as a circle of 44,237 bp and contains 65 genes. With an overall average length of only 42 bp for the intergenic regions, this is the most gene-dense mitochondrial genome of all Chlorophyta. Furthermore, it is characterized by a unique segmental duplication, encompassing 22 genes and covering 44% of the genome. Such a duplication has not been observed before in green algae, although it is also present in the mitochondrial genomes of higher plants. The quadripartite chloroplast genome forms a circle of 71,666 bp, containing 86 genes divided over a larger and a smaller single-copy region, separated by 2 inverted repeat sequences. Based on genome size and number of genes, the *Ostreococcus* chloroplast genome is the smallest known among the green algae. Phylogenetic analyses based on a concatenated alignment of chloroplast, mitochondrial, and nuclear genes confirm the position of *O. tauri* within the Prasinophyceae, an early branch of the Chlorophyta.

Once annotation data was provided, specific analysis could be performed. In this respect and in collaboration with Jörn Petersen, comparison of genes involved in the regulation of the Calvin Cycle (or carbon fixation cycle) within the green lineage provided us some unexpected results (chapter 6). Glyceraldehyde-3-phosphate dehydrogenase (GapAB) and CP12 are two major players in controlling the inactivation of the Calvin cycle in land plants at night. GapB originated from a *GapA* gene duplication and differs from GapA by the presence of a specific C-terminal extension that was recruited from CP12. While GapA and CP12 are assumed to be generally present in the Plantae (glaucophytes, red and green algae, and plants), up to now GapB was exclusively found in

Streptophyta, including the enigmatic green alga *Mesostigma viride*. However, we show that both *Ostreococcus* species also possess a *GapB* gene, while *CP12* is missing. This remarkable finding either antedates the *GapA/B* gene duplication or indicates a lateral recruitment. Moreover, *Ostreococcus* is the first case where the crucial CP12 function may be completely replaced by GapB-mediated GapA/B aggregation.

In order to get a better insight in the genome organisation within the green algae and to see whether above findings are *O. tauri*-specific, the genome of another strain, *Ostreococcus lucimarinus*, was sequenced and annotated (chapter 7). *O. lucimarinus* has a nuclear genome size of 13.2 million base pairs spread over 21 chromosomes, including 7,651 annotated genes. The comparison between both species reveals surprising differences across orthologous chromosomes, 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.

To finalize this thesis, chapter 8 describes a rather peculiar finding within different marine organisms. Obesity is a major societal issue contributing to increased morbidity and mortality. Recently, several studies have demonstrated an association between the *FTO* gene locus and early onset and severe obesity. Till recently, the *FTO* gene has only been discovered in vertebrate organisms. We have now identified *FTO* homologs in the genomes of marine eukaryotic algae, ranging from unicellular photosynthetic picoplankton to a multicellular sea weed. Although the biological roles of these *FTO* homologs are still unknown, these genes are helpful in exploring basic protein features that could potentially help unravelling the function of the *FTO* gene in vertebrates and its link with obesity in humans.

To conclude we can state that sequencing the genomes of two different members of the Prasinophyceae led to some unique and unexpected findings, but provided us also with a lot of unanswered questions. Future genome projects within the green algae will shed light on these issues and tell us whether these findings are really unique or can be linked to some group of species or even events in time.

Nederlandse samenvatting

Algen kunnen variëren in grootte, gaande van zeer kleine ééncellige organismen met een diameter van slechts 1 micro meter, tot de gigantische bruine wieren, die met een lengte tot 65 meter in staat zijn om als het ware onderzeese bossen te vormen. Tussen deze twee extreme levensvormen kan een enorme morfologische diversiteit onder de algen gedetecteerd worden: ééncellige organismen, vertakte of onvertakte filamenten, kolonies ontstaan uit een verzameling van individuele, al dan niet georganiseerde algen, en tenslotte de complexe *parenchymatous* algen. Afhankelijk van de hoeveelheid en samenstelling van hun pigmenten, zoals chlorofyl a dat noodzakelijk is voor fotosynthese, kunnen de eukaryote algen ruwweg onderverdeeld worden in drie grote groepen: de rode algen (Rhodophyta), de bruine algen (Phaeophyta) en de groenen algen. Deze laatste vormen de grootste groep van de drie. De groene algen worden beschouwd als de primitieve planten, die aan de basis van het ontstaan van de talrijke landplanten liggen die we vandaag de dag kennen. Ze zijn ongeveer 1,500 miljoen jaar geleden ontstaan en ze hebben zich ongeveer 425 tot 490 miljoen jaar geleden afgesplitst van de land planten. De groene algen kunnen onderverdeeld worden in twee groepen: de chlorophyte groene algen en de charophyte groene algen. Deze laatste zijn het nauwst verwant met de landplanten. Ten slotte bestaat er een derde groep groene algen, de Prasinophyceae, die een uitermate interessante groep vormen daar ze, aan de hand van fylogenetische studies, zich aan de basis van de groene plantenlijn bevinden. Door het voorkomen van deze verschillende groepen, kan vandaag de dag de groene plantenlijn onderverdeeld worden in twee grote groepen: de Streptophyta (landplanten en de charophyte groene algen) en de Chlorophyta (al de andere groene algen, inclusief de prasinophyte groene algen).

Algen behorend tot de Prasinophyceae zijn gedurende geruime tijd geïdentificeerd op basis van de aanwezigheid van schubben en een flagellum, die hen in staat stelt om te bewegen. Vandaag de dag echter zien we dat ook algen met of zonder flagellum en/of schubben tot deze groep behoren. Deze talrijke verschillen in morfologie, in combinatie met resultaten bekomen door fylogenetische studies gebaseerd op SSU rDNA sequenties, gaf aanleiding tot het definiëren van de Prasinophyceae als een parafyletische groep. In deze thesis beschrijven we

voor het eerst de sequentie en bijhorende annotatie van twee leden van de Prasinophyceae: *Ostreococcus tauri* en *Ostreococcus lucimarinus*.

O. tauri is een ééncellige groene alg die in de Mediterrane Thau lagune (Frankrijk) in 1994 werd ontdekt. Met een grootte van minder dan 1 micro meter, vergelijkbaar met dat van een bacterie, is *O. tauri* het kleinste eukaryote organisme dat tot op heden beschreven is. Zijn cellulaire opbouw is eerder eenvoudig met een vrij grote kern, één enkele chloroplast, één mitochondrium, één Golgi lichaam, en een gereduceerd cytoplasmatisch gedeelte. Een membraan omringt de cellen, maar geen celwand kan worden waargenomen.

Het belangrijkste doel van deze thesis was om aan de hand van het annoteren van de genomen van verschillende leden van de Chlorophyta, inzicht te verkrijgen in de genoomorganisatie van de groene algen. Alvorens echter het volledige genoom van *O. tauri* te annoteren en daarbij inzicht te verkrijgen in de samenstelling en vorm van zijn genen, werd een pilootstudie uitgevoerd. Daar er reeds heel wat kennis in ons labo aanwezig is over de celcyclus, werd deze aanvankelijk manueel bestudeerd in *O. tauri*. Hoofdstuk 2 beschrijft voor het eerst de aanwezigheid van een Cdc25 *dual specificity phosphatase* in de groene plantenlijn. Het Ccd25 fosfatase is een zeer belangrijk enzym in metazoans en gisten dat betrokken is bij de controle van de G2/M overgang tijdens de celcyclus. Tot voor kort was er nog in geen enkel lid van de groene plantenlijn een Cdc25 gen geïdentificeerd. Het Cdc25 gen dat in *O. tauri* ontdekt werd, codeert voor een proteïne dat in staat is om in *S. pombe* een *cdc25-22* mutant te herstellen. Bijkomende experimentele analyses hebben aangetoond dat het Ccd25 gen, dat gedetecteerd werd in *O. tauri,* alle eigenschappen bezit van de reeds gekende Cdc25 genen in andere organismen.

Wanneer we echter naar alle genen gaan kijken die een bepaalde rol spelen in de regulatie van de celcyclus, zien we dat in *O. tauri* alle reeds gekende genen aanwezig zijn (hoofdstuk 3). Opvallend hier is dat voor iedere genfamilie slechts 1 of maximaal 2 vertegenwoordigers aanwezig zijn: *O. tauri* bezit slechts 1 CDKA, CDKB, CDKD, cycline A, cycline B, cycline H, Cks, Rb, Dp, E_2F, DP, DEL en Cdc25 homoloog en twee cycline D en WEE1/Myt1 homologen. Deze bevindingen staan in groot contrast met de hoeveelheid celcyclus genen die eerder beschreven werden in de landplant *Arabidopsis thaliana*.

Eenmaal er voldoende kennis was opgedaan dankzij bovenvermelde pilootstudies

en we voldoende genen manueel geannoteerd hadden die gebruikt zouden worden als trainingsset voor het trainen van de nodige software, werd overgegaan tot het automatisch annoteren van het volledige genoom van *O. tauri* (hoofdstuk 4). Het nucleair genoom is 12,56 miljoen basen lang en bezit een enorm grote densiteit in genen: er werden 8,166 eiwit coderende genen geïdentificeerd die over 20 chromosomen verspreid zitten. Dit komt deels door de beperkte grootte van de intergenische gebieden en door de aanwezigheid van fusies tussen verschillende genen. Algemeen kan men echter stellen dat het genoom een complexe structuur bezit met een heel hoge graad van heterogeniteit, iets dat op heden nog nooit geobserveerd is binnen de eukaryote organismen. Twee van de 20 chromosomen verschillen sterk van de anderen. Beide hebben een afwijkende G+C inhoud en ze bezitten de overgrote meerderheid van de aanwezige *transposons*. Tevens hebben de meeste genen die gelokaliseerd zijn op chromosoom 2 een uniek codon gebruik en hebben ze een verschillend *splicing* systeem. Fylogenetische studies tonen echter aan dat deze genen geen vreemde (niet *O. tauri*) oorsprong hebben. De genen aanwezig op chromosoom 19 bezitten echter nauwelijks enige vorm van similariteit met de genen aanwezig in de groene plantenlijn. De meeste van deze genen spelen een rol in processen die zich afspelen aan de oppervlakte van de cel.

Naast het sequeneren van het nucleaire genoom werden de sequenties van het chloroplast en mitochondriaal genoom bepaald en manueel geannoteerd (hoofdstuk 5). Het mitochondriaal genoom komt voor als een cirkel, bestaande uit 44,237 base paren en bevat 65 genen. Het mitochondriale genoom bezit, dankzij de aanwezigheid van heel kleine intergenische gebieden van gemiddeld 42 bp lang, het meest compacte genoom van alle Chlorophyta. Voorts wordt het gekenmerkt door de aanwezigheid van een unieke segmentale duplicatie, die 22 genen bevat en 44% van het genoom uitmaakt. Een dergelijke duplicatie werd voordien nog niet waargenomen in de groene algen, hoewel het ook aanwezig is in de mitochondriale genomen van hogere landplanten. Het vierdelige chloroplast genoom vormt een cirkel van 71.666 bp en bevat 86 genen. Rekening houdend met de genoom grootte en het aantal aanwezige genen, blijkt dat het chloroplast genoom van *O. tauri* het kleinste is binnen de groene algen. Fylogenetische studies, gebaseerd op genen aanwezig in het chloroplast en/of mitochondriaal genoom, bevestigen de positie van *O. tauri* binnen de

Prasinophyceae.

Eenmaal de verschillende annotatie projecten afgelopen waren, konden meer specifiekere analyses uitgevoerd worden. In dit opzicht en in samenwerking met Jörn Petersen, verstrekte de vergelijking van de genen betrokken bij de regulatie van de Calvin Cyclus binnen de groene plantenlijn, ons enkele onverwachte resultaten (hoofdstuk6). Glyceraldehyde-3-fosfaat dehydrogenase (GapAB) en CP12 zijn twee belangrijke genen die betrokken zijn bij de controle van de inactivatie van de Calvin cyclus gedurende de nacht. GapB is ontstaan na een gen duplicatie van GapA en onderscheidt zich ervan door de aanwezigheid van een specifieke extensie, die afkomstig is van CP12. Terwijl algemeen aangenomen wordt dat GapA en CP12 universeel aanwezig zijn binnen de Plantae (glaucophyten, rode en groene algen en landplanten), werd GapB tot op heden enkel geïdentificeerd bij leden van de Streptophyta, inclusief de groene alg *Mesostigma viride*. Wij kunnen nu echter aantonen dat beide *Ostreococcus* species in het bezit zijn van een *GapA* en een *GapB* gen, terwijl CP12 niet gedetecteerd kon worden. Door deze unieke samenstelling moet ofwel het tijdstip van de *GapA/B* duplicatie herzien worden ofwel is dit een mooi voorbeeld van het lateraal verwerven van nieuwe genen. Tevens beschrijven wij hier voor het eerst de afwezigheid van CP12 en de mogelijke vervanging van de functie van CP12 door GapB.

Om vervolgens een beter inzicht te verwerven in de genoom organisatie van de groene algen en om bovenstaande *Ostreococcus*-specifieke gevallen te verifiëren, werd het genoom van *Ostreococcus lucimarinus* gesequeneerd (hoofdstuk 7). Het nucleaire genoom van *O. lucimarinus* bedraagt 13,2 miljoen bp. 7,651 genen werden geannoteerd, verspreid over 21 chromosomen. De vergelijkende studie tussen de orthologe chromosomen van beide soorten openbaart verrassende verschillen, gaande van bijna identieke chromosomen tot totaal verschillende chromosomen.

Om deze thesis te beëindigen, beschrijft hoofdstuk 8 een eerder bizar verhaal over een gen dat enkel aanwezig is in vertebraten en marine organismen. Obesitas of zwaarlijvigheid is een groeiend hedendaags probleem dat de dood als gevolg kan hebben. Onlangs hebben verscheidene studies een verband aangetoond tussen het FTO genlocus en het vroege begin van zwaarlijvigheid. Het *FTO* gen, dat tot op heden enkel geïdentificeerd was in vertebraten, blijkt

nu ook aanwezig te zijn in verschillende marine organismen, gaande van ééncellige groene algen tot meercellige bruine wieren. Hoewel de biologische werking van dit *FTO* gen tot op heden niet gekend is, kan de aanwezigheid van *FTO* binnen deze marine organismen ons menig inzicht verwerven over zijn algemene opbouw. De kennis die hierdoor verworven wordt, kan dan eventueel gebruikt worden om de functie van *FTO* binnen de vertebraten te verklaren en zijn eventuele link met zwaarlijvigheid bij mensen bloot te leggen.

Finaal kunnen we besluiten dat het sequeneren van de genomen van twee leden van de Prasinophyceae ons enkele unieke en onverwachte resultaten heeft opgeleverd, maar dat het ons vooral heeft opgezadeld met een groot aantal onbeantwoorde vragen. Toekomstige projecten waar meerdere genomen van groene algen gesequeneerd zullen worden, zullen ons hopelijk kunnen zeggen of deze feiten echt wel *Ostreococcus* specifiek zijn of we ze eerder kunnen toeschrijven aan een grotere groep organismen of zelfs aan een bepaalde gebeurtenis uit het verleden.

Bibliography

Abrahamsen, M. S., T. J. Templeton, S. Enomoto, J. E. Abrahante, G. Zhu, C. A. Lancto, M. Deng, C. Liu, G. Widmer, S. Tzipori, G. A. Buck, P. Xu, A. T. Bankier, P. H. Dear, B. A. Konfortov, H. F. Spriggs, L. Iyer, V. Anantharaman, L. Aravind, and V. Kapur. 2004. Complete genome sequence of the apicomplexan, *Cryptosporidium parvum*. Science 304:441-445.

Ahner, B. A., S. Kong, and F. M. M. Morel. 1995. Phytochelatin production in marine algae: I. An interspecies comparison. Limnol. Oceanogr. 40:649–657.

Altschul, S. F., W. Gish, W. Miller, E. W. Myers, and D. J. Lipman. 1990. Basic local alignment search tool. J Mol Biol 215:403-410.

Altschul, S. F., T. L. Madden, A. A. Schaffer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25:3389-3402.

Aparicio, S., J. Chapman, E. Stupka, N. Putnam, J. M. Chia, P. Dehal, A. Christoffels, S. Rash, S. Hoon, A. Smit, M. D. Gelpke, J. Roach, T. Oh, I. Y. Ho, M. Wong, C. Detter, F. Verhoef, P. Predki, A. Tay, S. Lucas, P. Richardson, S. F. Smith, M. S. Clark, Y. J. Edwards, N. Doggett, A. Zharkikh, S. V. Tavtigian, D. Pruss, M. Barnstead, C. Evans, H. Baden, J. Powell, G. Glusman, L. Rowen, L. Hood, Y. H. Tan, G. Elgar, T. Hawkins, B. Venkatesh, D. Rokhsar, and S. Brenner. 2002. Whole-genome shotgun assembly and analysis of the genome of *Fugu rubripes*. Science 297:1301-1310.

Apweiler, R., T. K. Attwood, A. Bairoch, A. Bateman, E. Birney, M. Biswas, P. Bucher, L. Cerutti, F. Corpet, M. D. Croning, R. Durbin, L. Falquet, W. Fleischmann, J. Gouzy, H. Hermjakob, N. Hulo, I. Jonassen, D. Kahn, A. Kanapin, Y. Karavidopoulou, R. Lopez, B. Marx, N. J. Mulder, T. M. Oinn, M. Pagni, F. Servant, C. J. Sigrist, and E. M. Zdobnov. 2001. The InterPro database, an integrated documentation resource for protein families, domains and functional sites. Nucleic Acids Res 29:37-40.

Armbrust, E. V., J. A. Berges, C. Bowler, B. R. Green, D. Martinez, N. H. Putnam, S. Zhou, A. E. Allen, K. E. Apt, M. Bechner, M. A. Brzezinski, B. K. Chaal, A. Chiovitti, A. K. Davis, M. S. Demarest, J. C. Detter, T. Glavina, D. Goodstein, M. Z. Hadi, U. Hellsten, M. Hildebrand, B. D. Jenkins, J. Jurka, V. V. Kapitonov, N. Kroger, W. W. Lau, T. W. Lane, F. W. Larimer, J. C. Lippmeier, S. Lucas, M. Medina, A. Montsant, M. Obornik, M. S. Parker, B. Palenik, G. J. Pazour, P. M. Richardson, T. A. Rynearson, M. A. Saito, D. C. Schwartz, K. Thamatrakoln, K. Valentin, A. Vardi, F. P. Wilkerson, and D. S. Rokhsar. 2004. The genome of the diatom *Thalassiosira pseudonana*: ecology, evolution, and metabolism. Science 306:79-86.

Ashburner, M., C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin, and G. Sherlock. 2000. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet 25:25-29.

Askwith, C. C., D. de Silva, and J. Kaplan. 1996. Molecular biology of iron acquisition in *Saccharomyces cerevisiae*. Mol Microbiol 20:27-34.

Baldauf, S. L. 2003. The deep roots of eukaryotes. Science 300:1703-1706.

Barroco, R. M., L. De Veylder, Z. Magyar, G. Engler, D. Inze, and V. Mironov. 2003. Novel complexes of cyclin-dependent kinases and a cyclin-like protein from *Arabidopsis thaliana* with a function unrelated to cell division. Cell Mol Life Sci 60:401-412.

Bassham, J. A. 2003. Mapping the carbon reduction cycle: a personal retrospective. Photosynth Res 76:35-52.

Behrenfeld, M., and P. Falkowski. 1997. A consumer's guide to phytoplankton primary productivity models. Limnol. Oceanogr 42:1479-1491.

Ben Ali, A., R. De Baere, G. Van der Auwera, R. De Wachter, and Y. Van de Peer. 2001. Phylogenetic relationship among algae based on complete large-subunit rRNA sequences. International Journal of Systematic and Evolutionary Microbiology 51:737-749.

Benson, D. A., I. Karsch-Mizrachi, D. J. Lipman, J. Ostell, B. A. Rapp, and D. L. Wheeler. 2002. GenBank. Nucleic Acids Res 30:17-20.

Berry, L. D., and K. L. Gould. 1996. Regulation of Cdc2 activity by phosphorylation at T14/Y15. Prog Cell Cycle Res 2:99-105.

Bhattacharya, D., and L. K. Medlin. 1998. Algal Phylogeny and the Origin of Land Plants. Plant Physiol 116:9-15.

Bhattacharya, D., K. Weber, S. S. An, and W. Berning-Koch. 1998. Actin phylogeny identifies *Mesostigma viride* as a flagellate ancestor of the land plants. J Mol Evol 47:544-550.

Birney, E., M. Clamp, and R. Durbin. 2004. GeneWise and Genomewise. Genome Res 14:988-995.

Bisova, K., D. M. Krylov, and J. G. Umen. 2005. Genome-wide annotation and expression profiling of cell cycle regulatory genes in *Chlamydomonas reinhardtii*. Plant Physiol 137:475-491.

Bleeker, P. M., H. W. Hakvoort, M. Bliek, E. Souer, and H. Schat. 2006. Enhanced arsenate reduction by a CDC25-like tyrosine phosphatase explains increased phytochelatin accumulation in arsenate-tolerant *Holcus lanatus*. Plant J 45:917-929.

Blom, N., S. Gammeltoft, and S. Brunak. 1999. Sequence and structure-based prediction of eukaryotic protein phosphorylation sites. J Mol Biol 294:1351-1362.

Bordo, D., and P. Bork. 2002. The rhodanese/Cdc25 phosphatase superfamily. Sequence-structure-function relations. EMBO Rep 3:741-746.

Borgne, A., A. C. Ostvold, S. Flament, and L. Meijer. 1999. Intra-M phase-promoting factor phosphorylation of cyclin B at the prophase/metaphase transition. J Biol Chem 274:11977-11986.

Borodovsky, M., K. E. Rudd, and E. V. Koonin. 1994. Intrinsic and extrinsic approaches for detecting genes in a bacterial genome. Nucleic Acids Res 22:4756-4767.

Boudolf, V., D. Inze, and L. De Veylder. 2006. What if higher plants lack a CDC25 phosphatase? Trends Plant Sci 11:474-479.

Bowes, G., S. Rao, G. Estavillo, and J. Reiskind. 2002. C4 mechanisms in aquatic angiosperms: comparisons with terrestrial C 4 systems. Functional Plant Biology 29:379-392.

Brennicke, A., L. Grohmann, R. Hiesel, V. Knoop, and W. Schuster. 1993. The mitochondrial genome on its way to the nucleus: different stages of gene transfer in higher plants. FEBS Lett 325:140-145.

Brown, S. L., M. R. Laundry, S. Christensen, D. Garrison, M. M. Gowing, R. R. Bidigare, and L. Campbell. 2002. Microbial community dynamics and taxon-specific phytoplankton production in the Arabian Sea during the 1995 monsoon seasons. Deep Sea Research Part II: Topical Studies in Oceanography 49:2345-2376.

Buchanan, B. B., and Y. Balmer. 2005. Redox regulation: a broadening horizon. Annu Rev Plant Biol 56:187-220.

Burke, D. T., G. F. Carle, and M. V. Olson. 1987. Cloning of large segments of exogenous DNA into yeast by means of artificial chromosome vectors. Science 236:806-812.

Campbell, L., H. Nolla, and D. Vaulot. 1994. The importance of *Prochlorococcus* to community structure in the central North Pacific Ocean. Limnology & Oceanography 39:955-961.

Capron, A., L. Okresz, and P. Genschik. 2003. First glance at the plant APC/C, a highly conserved ubiquitin-protein ligase. Trends Plant Sci 8:83-89.

Castresana, J. 2000. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. Mol Biol Evol 17:540-552.

Cerff, R. 1979. Quaternary structure of higher plant glyceraldehyde-3-phosphate dehydrogenases. Eur J Biochem 94:243-247.

Chepurnov, V. A., D. G. Mann, K. Sabbe, and W. Vyverman. 2004. Experimental studies on sexual reproduction in diatoms. Int Rev Cytol 237:91-154.

Chrétiennot-Dinet, M. J., C. Courties, A. Vaquer, J. Neveux, H. Claustre, J. Lautier, and M. C. Machado. 1995. A new marine picoeukaryote: *Ostreococcus tauri* gen. et sp. Nov. (Chlorophyta, Prasinophyceae). Phycologia 4:285-292.

Cobbett, C. S. 1999. A family of phytochelatin synthase genes from plant, fungal and animal species. Trends Plant Sci 4:335-337.

Cope, G. A., and R. J. Deshaies. 2003. COP9 signalosome: a multifunctional regulator of SCF and other cullin-based ubiquitin ligases. Cell 114:663-671.

Coppin, A., J. S. Varre, L. Lienard, D. Dauvillee, Y. Guerardel, M. O. Soyer-Gobillard, A. Buleon, S. Ball, and S. Tomavo. 2005. Evolution of plant-like crystalline storage polysaccharide in the protozoan parasite Toxoplasma gondii argues for a red alga ancestry. J Mol Evol 60:257-267.

Corellou, F., C. Brownlee, L. Detivaud, B. Kloareg, and F. Y. Bouget. 2001. Cell cycle in the fucus zygote parallels a somatic cell cycle but displays a unique translational regulation of cyclin-dependent kinases. Plant Cell 13:585-598.

Corellou, F., A. Camasses, L. Ligat, G. Peaucellier, and F. Y. Bouget. 2005. Atypical regulation of a green lineage-specific B-type cyclin-dependent kinase. Plant Physiol 138:1627-1636.

Costanzo, M., J. L. Nishikawa, X. Tang, J. S. Millman, O. Schub, K. Breitkreuz, D. Dewar, I. Rupes, B. Andrews, and M. Tyers. 2004. CDK activity antagonizes Whi5, an inhibitor of G1/S transcription in yeast. Cell 117:899-913.

Countway, P. D., and D. A. Caron. 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., R. Perasso, M. J. Chrétiennot-Dinet, M. Gouy, L. Guillou, and M. Troussellier. 1998. Phylogenetic analysis and genome size of *Ostreococcus tauri* (Chlorophyta, Prasinophyceae). J. Phycol 34:844-849.

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

Criqui, M. C., and P. Genschik. 2002. Mitosis in plants: how far we have come at the molecular level? Curr Opin Plant Biol 5:487-493.

Croft, M. T., A. D. Lawrence, E. Raux-Deery, M. J. Warren, and A. G. Smith. 2005. Algae acquire vitamin B12 through a symbiotic relationship with bacteria. Nature 438:90-93.

Cross, F. R., and J. M. Roberts. 2001. Retinoblastoma protein: combating algal bloom. Curr Biol 11:R824-827.

Curie, C., and J. F. Briat. 2003. Iron transport and signaling in plants. Annu Rev Plant Biol 54:183-206.

de Cambiaire, J. C., C. Otis, C. Lemieux, and M. Turmel. 2006. The complete chloroplast genome sequence of the chlorophycean green alga *Scenedesmus obliquus* reveals a compact gene organization and a biased distribution of genes on the two DNA strands. BMC Evol Biol 6:37.

de Koning, A. P., and P. J. Keeling. 2006. The complete plastid genome sequence of the parasitic green alga *Helicosporidium sp*. is highly reduced and structured. BMC Biol 21:12.

De Rijk, P., and R. De Wachter. 1993. DSCE, an interactive tool for sequence alignment and secondary structure research. Comput Appl Biosci 9:735-740.

De Rijk, P., J. Wuyts, and R. De Wachter. 2003. RNAVIZ 2: an improved representation of RNA secondary structure. Bioinformatics 19:299-300.

De Veylder, L., T. Beeckman, G. T. Beemster, L. Krols, F. Terras, I. Landrieu, E. van der Schueren, S. Maes, M. Naudts, and D. Inze. 2001. Functional analysis of cyclin-dependent kinase inhibitors of *Arabidopsis*. Plant Cell 13:1653-1668.

Degroeve, S., Y. Saeys, B. De Baets, P. Rouze, and Y. Van de Peer. 2005. SpliceMachine: predicting splice sites from high-dimensional local context representations. Bioinformatics 21:1332-1338.

del Pozo, J. C., M. B. Boniotti, and C. Gutierrez. 2002. *Arabidopsis* E2Fc functions in cell division and is degraded by the ubiquitin-SCF(AtSKP2) pathway in response to light. Plant Cell 14:3057-3071.

Derelle, E., C. Ferraz, P. Lagoda, S. Eychenié, R. Cooke, F. Regad, X. Sabau, C. Courties, M. Delseny, J. Demaille, A. Picard, and H. Moreau. 2002. DNA libraries for sequencing the genome of *Ostreococcus tauri* (Chlorophytae,

Prasinophyceae) : the smallest free-living eukaryotic cell. J. Phycology 38:1150-1156.

Derelle, E., C. Ferraz, S. Rombauts, P. Rouze, A. Z. Worden, S. Robbens, F. Partensky, S. Degroeve, S. Echeynie, R. Cooke, Y. Saeys, J. Wuyts, K. Jabbari, C. Bowler, O. Panaud, B. Piegu, S. G. Ball, J. P. Ral, F. Y. Bouget, G. Piganeau, B. De Baets, A. Picard, M. Delseny, J. Demaille, Y. Van de Peer, and H. Moreau. 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.

Deshaies, R. J., and J. E. Ferrell, Jr. 2001. Multisite phosphorylation and the countdown to S phase. Cell 107:819-822.

Dewitte, W., and J. A. Murray. 2003. The plant cell cycle. Annu Rev Plant Biol 54:235-264.

Dhankher, O. P., B. P. Rosen, E. C. McKinney, and R. B. Meagher. 2006. Hyperaccumulation of arsenic in the shoots of *Arabidopsis* silenced for arsenate reductase (ACR2). Proc Natl Acad Sci U S A 103:5413-5418.

Di Stefano, L., M. R. Jensen, and K. Helin. 2003. E2F7, a novel E2F featuring DP-independent repression of a subset of E2F-regulated genes. Embo J 22:6289-6298.

Dickerson, R. E. 1971. The structures of cytochrome c and the rates of molecular evolution. J Mol Evol 1:26-45.

Dietrich, F. S., S. Voegeli, S. Brachat, A. Lerch, K. Gates, S. Steiner, C. Mohr, R. Pohlmann, P. Luedi, S. Choi, R. A. Wing, A. Flavier, T. D. Gaffney, and P. Philippsen. 2004. The *Ashbya gossypii* genome as a tool for mapping the ancient *Saccharomyces cerevisiae* genome. Science 304:304-307.

Díez, B., C. Pedrós-Alió, and R. Massana. 2001. Study of Genetic Diversity of Eukaryotic Picoplankton in Different Oceanic Regions by Small-Subunit rRNA Gene Cloning and Sequencing. Applied and Environmental Microbiology 67:2932-2941.

Dina, C., D. Meyre, S. Gallina, E. Durand, A. Korner, P. Jacobson, L. M. Carlsson, W. Kiess, V. Vatin, C. Lecoeur, J. Delplanque, E. Vaillant, F. Pattou, J. Ruiz, J. Weill, C. Levy-Marchal, F. Horber, N. Potoczna, S. Hercberg, C. Le Stunff, P. Bougneres, P. Kovacs, M. Marre, B. Balkau, S. Cauchi, J. C. Chevre, and P. Froguel. 2007. Variation in *FTO* contributes to childhood obesity and severe adult obesity. Nat Genet 39:724-726.

Do, J. H., and D. K. Choi. 2006. Computational approaches to gene prediction. J Microbiol 44:137-144.

Doree, M., and T. Hunt. 2002. From Cdc2 to Cdk1: when did the cell cycle kinase join its cyclin partner? J Cell Sci 115:2461-2464.

Eddy, S. R. 2002. A memory-efficient dynamic programming algorithm for optimal alignment of a sequence to an RNA secondary structure. BMC Bioinformatics 3:18.

Eddy, S. R. 1998. Profile hidden Markov models. Bioinformatics 14:755-763.

Everitt, D. A., S. W. Wright, J. K. Volkman, D. P. Thomas, and E. J. Lindstrom. 1990. Phytoplankton community compositions in the western equatorial Pacific determined from chlorophyll and carotenoid pigment distributions. Deep-Sea Res. 37:975-997.

Ewing, B., L. Hillier, M. C. Wendl, and P. Green. 1998. Base-calling of automated sequencer traces using phred. I. Accuracy assessment. Genome Res 8:175-185.

Fauron, C., M. Casper, Y. Gao, and B. Moore. 1995. The maize mitochondrial genome: dynamic, yet functional. Trends Genet 11:228-235.

Fawley, M. W., Y. Yun, and M. Qin. 2000. Phylogenetic analyses of 18s rdna sequences reveal a new coccoid lineage of the prasinophyceae (Chlorophyta). Journal of Phycology 36:387-393.

Feschotte, C., and S. R. Wessler. 2002. Mariner-like transposases are widespread and diverse in flowering plants. Proc Natl Acad Sci U S A 99:280-285.

Fickett, J. W. 1996. The gene identification problem: an overview for developers. Comput Chem 20:103-118.

Field, S. F., J. M. Howson, N. M. Walker, D. B. Dunger, and J. A. Todd. 2007. Analysis of the obesity gene *FTO* in 14,803 type 1 diabetes cases and controls. Diabetologia 50:2218-2220.

Forsburg, S. L. 1993. Comparison of *Schizosaccharomyces pombe* expression systems. Nucleic Acids Res 21:2955-2956.

Fouilland, E., C. Descolas-Gros, C. Courties, Y. Collos, A. Vaquer, and A. Gasc. 2004. Productivity and growth of a natural population of the smallest free-living eukaryote under nitrogen deficiency and sufficiency. Microb Ecol 48:103-110.

Franklin, S. E., and S. P. Mayfield. 2004. Prospects for molecular farming in the green alga *Chlamydomonas*. Curr Opin Plant Biol 7:159-165.

Fraser, J. A., and J. Heitman. 2004. Evolution of fungal sex chromosomes. Mol Microbiol 51:299-306.

Frayling, T. M. 2007. Genome-wide association studies provide new insights into type 2 diabetes aetiology. Nat Rev Genet 8:657-662.

Frayling, T. M., N. J. Timpson, M. N. Weedon, E. Zeggini, R. M. Freathy, C. M. Lindgren, J. R. Perry, K. S. Elliott, H. Lango, N. W. Rayner, B. Shields, L. W. Harries, J. C. Barrett, S. Ellard, C. J. Groves, B. Knight, A. M. Patch, A. R. Ness, S. Ebrahim, D. A. Lawlor, S. M. Ring, Y. Ben-Shlomo, M. R. Jarvelin, U. Sovio, A. J. Bennett, D. Melzer, L. Ferrucci, R. J. Loos, I. Barroso, N. J. Wareham, F. Karpe, K. R. Owen, L. R. Cardon, M. Walker, G. A. Hitman, C. N. Palmer, A. S. Doney, A. D. Morris, G. D. Smith, A. T. Hattersley, and M. I. McCarthy. 2007. A common variant in the *FTO* gene is associated with body mass index and predisposes to childhood and adult obesity. Science 316:889-894.

Frazer, K. A., L. Pachter, A. Poliakov, E. M. Rubin, and I. Dubchak. 2004. VISTA: computational tools for comparative genomics. Nucleic Acids Res 32:W273-279.

Futcher, B. 1996. Cyclins and the wiring of the yeast cell cycle. Yeast 12:1635-1646.

Gaasterland, T., and C. W. Sensen. 1996. Fully automated genome analysis that reflects user needs and preferences. A detailed introduction to the MAGPIE system architecture. Biochimie 78:302-310.

Gardner, R. C., A. J. Howarth, P. Hahn, M. Brown-Luedi, R. J. Shepherd, and J. Messing. 1981. The complete nucleotide sequence of an infectious clone of

cauliflower mosaic virus by M13mp7 shotgun sequencing. Nucleic Acids Res 9:2871-2888.
Gene Ontology Consortium. 2001. Creating the gene ontology resource: design and implementation. Genome Res 11:1425-1433.
Gilson, P. R. 2001. Nucleomorph genomes: much ado about practically nothing. Genome Biol 2:REVIEWS1022.
Giordano, M., J. Beardall, and J. A. Raven. 2005. CO2 concentrating mechanisms in algae: mechanisms, environmental modulation, and evolution. Annu Rev Plant Biol 56:99-131.
Gladyshev, V. N., and G. V. Kryukov. 2001. Evolution of selenocysteine-containing proteins: significance of identification and functional characterization of selenoproteins. Biofactors 14:87-92.
Goh, C. S., A. A. Bogan, M. Joachimiak, D. Walther, and F. E. Cohen. 2000. Co-evolution of proteins with their interaction partners. J Mol Biol 299:283-293.
Gonzalez, J., J. M. Ranz, and A. Ruiz. 2002. Chromosomal elements evolve at different rates in the *Drosophila* genome. Genetics 161:1137-1154.
Gordon, D., C. Abajian, and P. Green. 1998. CONSED: a graphical tool for sequence finishing. Genome Res 8:195-202.
Gordon, J. M., and J. E. Polle. 2007. Ultrahigh bioproductivity from algae. Appl Microbiol Biotechnol.
Graham, E. L., and W. L. Wilcox. 1999. Algae. Prentice Hall, Upper Sadle River.
Green, E. D. 2001. Strategies for the systematic sequencing of complex genomes. Nat Rev Genet 2:573-583.
Griffiths-Jones, S., S. Moxon, M. Marshall, A. Khanna, S. R. Eddy, and A. Bateman. 2005. Rfam: annotating non-coding RNAs in complete genomes. Nucleic Acids Res 33:D121-124.
Groop, L. 2007. From fused toes in mice to human obesity. Nat Genet 39:706-707.
Grzebyk, D., and O. Schofield. 2003. The Mesozoic radiation of eukaryotic algae: the portable plastid hypothesis. J. Phycol 39:259-267.
Guillou, L., W. Eikrem, M. J. Chretiennot-Dinet, F. Le Gall, R. Massana, K. Romari, C. Pedros-Alio, and D. Vaulot. 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.
Guindon, S., and O. Gascuel. 2003. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. Syst Biol 52:696-704.
Gustafson, M. P., C. F. Thomas, Jr., F. Rusnak, A. H. Limper, and E. B. Leof. 2001. Differential regulation of growth and checkpoint control mediated by a Cdc25 mitotic phosphatase from *Pneumocystis carinii*. J Biol Chem 276:835-843.
Haag, A. L. 2007. Algae bloom again. Nature 447:520-521.
Haas, B. J., J. R. Wortman, C. M. Ronning, L. I. Hannick, R. K. Smith, Jr., R. Maiti, A. P. Chan, C. Yu, M. Farzad, D. Wu, O. White, and C. D. Town. 2005. Complete reannotation of the *Arabidopsis* genome: methods, tools, protocols and the final release. BMC Biol 3:7.

Hall, T. A. 1999. BIOEDIT: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. Nucleic Acids Symposium Series 41:95-98.

Hamel, P. P., B. W. Dreyfuss, Z. Xie, S. T. Gabilly, and S. Merchant. 2003. Essential histidine and tryptophan residues in CcsA, a system II polytopic cytochrome c biogenesis protein. J Biol Chem 278:2593-2603.

Handa, H. 2003. The complete nucleotide sequence and RNA editing content of the mitochondrial genome of rapeseed (*Brassica napus L.*): comparative analysis of the mitochondrial genomes of rapeseed and *Arabidopsis thaliana*. Nucleic Acids Res 31:5907-5916.

Harris, M. A., J. Clark, A. Ireland, J. Lomax, M. Ashburner, R. Foulger, K. Eilbeck, S. Lewis, B. Marshall, C. Mungall, J. Richter, G. M. Rubin, J. A. Blake, C. Bult, M. Dolan, H. Drabkin, J. T. Eppig, D. P. Hill, L. Ni, M. Ringwald, R. Balakrishnan, J. M. Cherry, K. R. Christie, M. C. Costanzo, S. S. Dwight, S. Engel, D. G. Fisk, J. E. Hirschman, E. L. Hong, R. S. Nash, A. Sethuraman, C. L. Theesfeld, D. Botstein, K. Dolinski, B. Feierbach, T. Berardini, S. Mundodi, S. Y. Rhee, R. Apweiler, D. Barrell, E. Camon, E. Dimmer, V. Lee, R. Chisholm, P. Gaudet, W. Kibbe, R. Kishore, E. M. Schwarz, P. Sternberg, M. Gwinn, L. Hannick, J. Wortman, M. Berriman, V. Wood, N. de la Cruz, P. Tonellato, P. Jaiswal, T. Seigfried, and R. White. 2004. The Gene Ontology (GO) database and informatics resource. Nucleic Acids Res 32:D258-261.

Henderson, G. P., L. Gan, and G. J. Jensen. 2007. 3-D ultrastructure of *O. tauri*: electron cryotomography of an entire eukaryotic cell. PLoS ONE 2:e749.

Hiramoto, Y. 1974. A method of microinjection. Exp Cell Res 87:403-406.

Huelsenbeck, J. P., F. Ronquist, R. Nielsen, and J. P. Bollback. 2001. Bayesian inference of phylogeny and its impact on evolutionary biology. Science 294:2310-2314.

Hutchinson, G. E. 1961. The paradox of the plankton. Am. Nat. 95:137-145.

Irniger, S. 2002. Cyclin destruction in mitosis: a crucial task of Cdc20. FEBS Lett 532:7-11.

Janska, H., R. Sarria, M. Woloszynska, M. Arrieta-Montiel, and S. A. Mackenzie. 1998. Stoichiometric shifts in the common bean mitochondrial genome leading to male sterility and spontaneous reversion to fertility. Plant Cell 10:1163-1180.

Jeffrey, P. D., A. A. Russo, K. Polyak, E. Gibbs, J. Hurwitz, J. Massague, and N. P. Pavletich. 1995. Mechanism of CDK activation revealed by the structure of a cyclinA-CDK2 complex. Nature 376:313-320.

Jimenez, J., L. Alphey, P. Nurse, and D. M. Glover. 1990. Complementation of fission yeast cdc2ts and cdc25ts mutants identifies two cell cycle genes from *Drosophila*: a cdc2 homologue and string. Embo J 9:3565-3571.

Jobb, G., A. von Haeseler, and K. Strimmer. 2004. TREEFINDER: a powerful graphical analysis environment for molecular phylogenetics. BMC Evol Biol 4:18.

Johnson, P. W., and J. Sieburth. 1979. Chroococcoid cyanobacteria in the sea: a ubiquitous and diverse phototrophic biomass. Limnology and Oceanography 24:928-935.

Joint, I. R., N. J. P. Owen, and A. J. Pomroy. 1986. Seasonal production of photosynthetic picoplankton and nanoplankton in the Celtic sea. Mar. Ecol. Prog. Ser 28:251-258.

Joint, I. R., and R. K. Pipe. 1984. An electron microscopic study of a natural population of picoplanktonic from the Celtic Sea. Mar. Ecol. Prog. Ser. 20:113-118.

Jones, D. T., W. R. Taylor, and J. M. Thornton. 1992. The rapid generation of mutation data matrices from protein sequences. Comput Appl Biosci 8:275-282.

Joubes, J., C. Chevalier, D. Dudits, E. Heberle-Bors, D. Inze, M. Umeda, and J. P. Renaudin. 2000. CDK-related protein kinases in plants. Plant Mol Biol 43:607-620.

Jukes, T. H. 1969. Recent advances in studies of evolutionary relationships between proteins and nucleic acids. Space Life Sci 1:469-490.

Kanazawa, A., N. Tsutsumi, and A. Hirai. 1994. Reversible changes in the composition of the population of mtDNAs during dedifferentiation and regeneration in tobacco. Genetics 138:865-870.

Kanehisa, M., S. Goto, S. Kawashima, Y. Okuno, and M. Hattori. 2004. The KEGG resource for deciphering the genome. Nucleic Acids Res 32:D277-280.

Kapraun, D. F. 2007. Nuclear DNA content estimates in green algal lineages: chlorophyta and streptophyta. Ann Bot (Lond) 99:677-701.

Karol, K. G., R. M. McCourt, M. T. Cimino, and C. F. Delwiche. 2001. The closest living relatives of land plants. Science 294:2351-2353.

Kent, W. J. 2002. BLAT—the BLAST-like alignment tool. Genome Res 12:656-664.

Khadaroo, B., S. Robbens, C. Ferraz, E. Derelle, S. Eychenie, R. Cooke, G. Peaucellier, M. Delseny, J. Demaille, Y. Van de Peer, A. Picard, and H. Moreau. 2004. The first green lineage cdc25 dual-specificity phosphatase. Cell Cycle 3:513-518.

Kim, H. Y., D. E. Fomenko, Y. E. Yoon, and V. N. Gladyshev. 2006. Catalytic advantages provided by selenocysteine in methionine-S-sulfoxide reductases. Biochemistry 45:13697-13704.

Kimura, M. 1983. The neutral theory of molecular evolution.

Kliebenstein, D. J., R. A. Monde, and R. L. Last. 1998. Superoxide dismutase in *Arabidopsis*: an eclectic enzyme family with disparate regulation and protein localization. Plant Physiol 118:637-650.

Knoop, V. 2004. The mitochondrial DNA of land plants: peculiarities in phylogenetic perspective. Curr Genet 46:123-139.

Koonin, E. V., N. D. Fedorova, J. D. Jackson, A. R. Jacobs, D. M. Krylov, K. S. Makarova, R. Mazumder, S. L. Mekhedov, A. N. Nikolskaya, B. S. Rao, I. B. Rogozin, S. Smirnov, A. V. Sorokin, A. V. Sverdlov, S. Vasudevan, Y. I. Wolf, J. J. Yin, and D. A. Natale. 2004. A comprehensive evolutionary classification of proteins encoded in complete eukaryotic genomes. Genome Biol 5:R7.

Kosman, D. J. 2003. Molecular mechanisms of iron uptake in fungi. Mol Microbiol 47:1185-1197.

Koszul, R., S. Caburet, B. Dujon, and G. Fischer. 2004. Eucaryotic genome evolution through the spontaneous duplication of large chromosomal segments. Embo J 23:234-243.

Kraft, C. 2003. Mitotic entry: tipping the balance. Curr Biol 13:R445-446.

Kranz, R., R. Lill, B. Goldman, G. Bonnard, and S. Merchant. 1998. Molecular mechanisms of cytochrome c biogenesis: three distinct systems. Mol Microbiol 29:383-396.

Kubo, T., S. Nishizawa, A. Sugawara, N. Itchoda, A. Estiati, and T. Mikami. 2000. The complete nucleotide sequence of the mitochondrial genome of sugar beet (*Beta vulgaris L.*) reveals a novel gene for tRNA(Cys)(GCA). Nucleic Acids Res 28:2571-2576.

Kurtz, S., J. V. Choudhuri, E. Ohlebusch, C. Schleiermacher, J. Stoye, and R. Giegerich. 2001. REPUTER: the manifold applications of repeat analysis on a genomic scale. Nucleic Acids Res 29:4633-4642.

Kurtz, S., A. Phillippy, A. L. Delcher, M. Smoot, M. Shumway, C. Antonescu, and S. L. Salzberg. 2004. Versatile and open software for comparing large genomes. Genome Biol 5:R12.

La Fontaine, S., J. M. Quinn, S. S. Nakamoto, M. D. Page, V. Gohre, J. L. Moseley, J. Kropat, and S. Merchant. 2002. Copper-dependent iron assimilation pathway in the model photosynthetic eukaryote *Chlamydomonas reinhardtii*. Eukaryot Cell 1:736-757.

Landrieu, I., M. da Costa, L. De Veylder, F. Dewitte, K. Vandepoele, S. Hassan, J. M. Wieruszeski, F. Corellou, J. D. Faure, M. Van Montagu, D. Inze, and G. Lippens. 2004. A small CDC25 dual-specificity tyrosine-phosphatase isoform in *Arabidopsis thaliana*. Proc Natl Acad Sci U S A 101:13380-13385.

Landrieu, I., S. Hassan, M. Sauty, F. Dewitte, J. M. Wieruszeski, D. Inze, L. De Veylder, and G. Lippens. 2004. Characterization of the *Arabidopsis thaliana* Arath;CDC25 dual-specificity tyrosine phosphatase. Biochem Biophys Res Commun 322:734-739.

Lemieux, C., C. Otis, and M. Turmel. 2000. Ancestral chloroplast genome in *Mesostigma viride* reveals an early branch of green plant evolution. Nature 403:649-652.

Lewis, A. L., and M. R. McCourt. 2004. Green algae and the origin of land plants. American Journal of Botany 91:1535-1556.

Li, J. B., J. M. Gerdes, C. J. Haycraft, Y. Fan, T. M. Teslovich, H. May-Simera, H. Li, O. E. Blacque, L. Li, C. C. Leitch, R. A. Lewis, J. S. Green, P. S. Parfrey, M. R. Leroux, W. S. Davidson, P. L. Beales, L. M. Guay-Woodford, B. K. Yoder, G. D. Stormo, N. Katsanis, and S. K. Dutcher. 2004. Comparative genomics identifies a flagellar and basal body proteome that includes the BBS5 human disease gene. Cell 117:541-552.

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

Li, W. K. W. 1994. Phytoplankton biomass and chlorophyll concentration across the North Atlantic. Scientia Marina 58:67-79.

Lister, D. L., J. M. Bateman, S. Purton, and C. J. Howe. 2003. DNA transfer from chloroplast to nucleus is much rarer in *Chlamydomonas* than in tobacco. Gene 316:33-38.

López-García, P., F. Rodríguez-Valera, C. Pedrós-Alió, and D. Moreira. 2001. Unexpected diversity of small eukaryotes in deep-sea Antarctic plankton. Nature 409:603-607.

Lowe, T. M., and S. R. Eddy. 1997. TRNASCAN-SE: a program for improved detection of transfer RNA genes in genomic sequence. Nucleic Acids Res 25:955-964.

Lundgren, K., N. Walworth, R. Booher, M. Dembski, M. Kirschner, and D. Beach. 1991. mik1 and wee1 cooperate in the inhibitory tyrosine phosphorylation of cdc2. Cell 64:1111-1122.

Ma, B., T. Elkayam, H. Wolfson, and R. Nussinov. 2003. Protein-protein interactions: structurally conserved residues distinguish between binding sites and exposed protein surfaces. Proc Natl Acad Sci U S A 100:5772-5777.

Mandal, M., and R. R. Breaker. 2004. Gene regulation by riboswitches. Nat Rev Mol Cell Biol 5:451-463.

Martin, J., K. Coale, K. Johnson, S. Fitzwater, R. Gordon, S. Tanner, and C. Hunter. 1994. Testing the iron hypothesis in ecosystems of the equatorial Pacific Ocean. Nature 371:123-129.

Martin, W., and R. G. Herrmann. 1998. Gene transfer from organelles to the nucleus: how much, what happens, and Why? Plant Physiol 118:9-17.

Martin, W., T. Rujan, E. Richly, A. Hansen, S. Cornelsen, T. Lins, D. Leister, B. Stoebe, M. Hasegawa, and D. Penny. 2002. Evolutionary analysis of *Arabidopsis*, cyanobacterial, and chloroplast genomes reveals plastid phylogeny and thousands of cyanobacterial genes in the nucleus. Proc Natl Acad Sci U S A 99:12246-12251.

Mathé, C., M. F. Sagot, T. Schiex, and P. Rouze. 2002. Current methods of gene prediction, their strengths and weaknesses. Nucleic Acids Res 30:4103-4117.

Matsuzaki, M., O. Misumi, I. T. Shin, S. Maruyama, M. Takahara, S. Y. Miyagishima, T. Mori, K. Nishida, F. Yagisawa, K. Nishida, Y. Yoshida, Y. Nishimura, S. Nakao, T. Kobayashi, Y. Momoyama, T. Higashiyama, A. Minoda, M. Sano, H. Nomoto, K. Oishi, H. Hayashi, F. Ohta, S. Nishizaka, S. Haga, S. Miura, T. Morishita, Y. Kabeya, K. Terasawa, Y. Suzuki, Y. Ishii, S. Asakawa, H. Takano, N. Ohta, H. Kuroiwa, K. Tanaka, N. Shimizu, S. Sugano, N. Sato, H. Nozaki, N. Ogasawara, Y. Kohara, and T. Kuroiwa. 2004. Genome sequence of the ultrasmall unicellular red alga *Cyanidioschyzon merolae 10D*. Nature 428:653-657.

Maul, J. E., J. W. Lilly, L. Cui, C. W. dePamphilis, W. Miller, E. H. Harris, and D. B. Stern. 2002. The *Chlamydomonas reinhardtii* plastid chromosome: islands of genes in a sea of repeats. Plant Cell 14:2659-2679.

Maundrell, K. 1990. nmt1 of fission yeast. A highly transcribed gene completely repressed by thiamine. J Biol Chem 265:10857-10864.

Mayfield, S. P., and S. E. Franklin. 2005. Expression of human antibodies in eukaryotic micro-algae. Vaccine 23:1828-1832.

Mayfield, S. P., S. E. Franklin, and R. A. Lerner. 2003. Expression and assembly of a fully active antibody in algae. Proc Natl Acad Sci U S A 100:438-442.

Mayfield, S. P., A. L. Manuell, S. Chen, J. Wu, M. Tran, D. Siefker, M. Muto, and J. Marin-Navarro. 2007. *Chlamydomonas reinhardtii* chloroplasts as protein factories. Curr Opin Biotechnol 18:126-133.

McCourt, M. R., K. G. Karol, S. Kaplan, and R. W. Hoshaw. 1995. Using rbcL sequences to test hypotheses of chloroplast and thallus evolution in conjugating green algae (Zygnematales, Charophyceae). Journal of Phycology 36:989-995.

McCourt, R. M., C. F. Delwiche, and K. G. Karol. 2004. Charophyte algae and land plant origins. Trends Ecol Evol 19:661-666.

McKibbin, R. S., N. G. Halford, and D. Francis. 1998. Expression of fission yeast cdc25 alters the frequency of lateral root formation in transgenic tobacco. Plant Mol Biol 36:601-612.

Mead, J. R., M. J. Arrowood, W. L. Current, and C. R. Sterling. 1988. Field inversion gel electrophoretic separation of *Cryptosporidium spp.* chromosome-sized DNA. J Parasitol 74:366-369.

Melkonian, M. 1989. Flagellar apparatus ultrastructure in *Mesostigma viride* (Prasinophyceae). Plant Syst. Evol. 164:93-122.

Mendenhall, M. D., and A. E. Hodge. 1998. Regulation of Cdc28 cyclin-dependent protein kinase activity during the cell cycle of the yeast *Saccharomyces cerevisiae*. Microbiol Mol Biol Rev 62:1191-1243.

Merchant, S. S.S. E. ProchnikO. VallonE. H. HarrisS. J. KarpowiczG. B. WitmanA. TerryA. SalamovL. K. Fritz-LaylinL. Marechal-DrouardW. F. MarshallL. H. QuD. R. NelsonA. A. SanderfootM. H. SpaldingV. V. KapitonovQ. RenP. FerrisE. LindquistH. ShapiroS. M. LucasJ. GrimwoodJ. SchmutzP. CardolH. CeruttiG. ChanfreauC. L. ChenV. CognatM. T. CroftR. DentS. DutcherE. FernandezH. FukuzawaD. Gonzalez-BallesterD. Gonzalez-HalphenA. HallmannM. HanikenneM. HipplerW. InwoodK. JabbariM. KalanonR. KurasP. A. LefebvreS. D. LemaireA. V. LobanovM. LohrA. ManuellI. MeierL. MetsM. MittagT. MittelmeierJ. V. MoroneyJ. MoseleyC. NapoliA. M. NedelcuK. NiyogiS. V. NovoselovI. T. PaulsenG. PazourS. PurtonJ. P. RalD. M. Riano-PachonW. RiekhofL. RymarquisM. SchrodaD. SternJ. UmenR. WillowsN. WilsonS. L. ZimmerJ. AllmerJ. BalkK. BisovaC. J. ChenM. EliasK. GendlerC. HauserM. R. LambH. LedfordJ. C. LongJ. MinagawaM. D. PageJ. PanW. PootakhamS. RojeA. RoseE. StahlbergA. M. TerauchiP. YangS. BallC. BowlerC. L. DieckmannV. N. GladyshevP. GreenR. JorgensenS. MayfieldB. Mueller-RoeberS. RajamaniR. T. SayreP. BroksteinI. DubchakD. GoodsteinL. HornickY. W. HuangJ. JhaveriY. LuoD. MartinezW. C. NgauB. OtillarA. PoliakovA. PorterL. SzajkowskiG. WernerK. ZhouI. V. GrigorievD. S. Rokhsar, and A. R. Grossman. 2007. The *Chlamydomonas* genome reveals the evolution of key animal and plant functions. Science 318:245-250.

Meyers, B. C., S. Scalabrin, and M. Morgante. 2004. Mapping and sequencing complex genomes: let's get physical! Nat Rev Genet 5:578-588.

Michaelis, G., C. Vahrenholz, and E. Pratje. 1990. Mitochondrial DNA of *Chlamydomonas reinhardtii*: the gene for apocytochrome b and the

complete functional map of the 15.8 kb DNA. Mol Gen Genet 223:211-216.

Michels, A. K., N. Wedel, and P. G. Kroth. 2005. Diatom plastids possess a phosphoribulokinase with an altered regulation and no oxidative pentose phosphate pathway. Plant Physiol 137:911-920.

Mironov, V. V., L. De Veylder, M. Van Montagu, and D. Inze. 1999. Cyclin-dependent kinases and cell division in plants- the nexus. Plant Cell 11:509-522.

Moffett, J., L. Brand, P. Croot, and K. Barbeau. 1997. Cu speciation and cyanobacterial distribution in harbours subject to anthropogenic Cu inputs. Limnol. Oceanogr. 42:789-799.

Moon-van der Staay, S. Y., R. De Wachter, and D. Vaulot. 2001. Oceanic 18S rDNA sequences from picoplankton reveal unsuspected eukaryotic diversity. Nature 409:607-610.

Moore, L. R., G. Rocap, and S. W. Chisholm. 1998. Physiology and molecular phylogeny of coexisting *Prochlorococcus* ecotypes. Nature 393:464-467.

Morris, M. C., P. Kaiser, S. Rudyak, C. Baskerville, M. H. Watson, and S. I. Reed. 2003. Cks1-dependent proteasome recruitment and activation of CDC20 transcription in budding yeast. Nature 423:1009-1013.

Mulder, N. J., R. Apweiler, T. K. Attwood, A. Bairoch, A. Bateman, D. Binns, P. Bork, V. Buillard, L. Cerutti, R. Copley, E. Courcelle, U. Das, L. Daugherty, M. Dibley, R. Finn, W. Fleischmann, J. Gough, D. Haft, N. Hulo, S. Hunter, D. Kahn, A. Kanapin, A. Kejariwal, A. Labarga, P. S. Langendijk-Genevaux, D. Lonsdale, R. Lopez, I. Letunic, M. Madera, J. Maslen, C. McAnulla, J. McDowall, J. Mistry, A. Mitchell, A. N. Nikolskaya, S. Orchard, C. Orengo, R. Petryszak, J. D. Selengut, C. J. Sigrist, P. D. Thomas, F. Valentin, D. Wilson, C. H. Wu, and C. Yeats. 2007. New developments in the InterPro database. Nucleic Acids Res 35:D224-228.

Mulder, N. J., R. Apweiler, T. K. Attwood, A. Bairoch, A. Bateman, D. Binns, P. Bradley, P. Bork, P. Bucher, L. Cerutti, R. Copley, E. Courcelle, U. Das, R. Durbin, W. Fleischmann, J. Gough, D. Haft, N. Harte, N. Hulo, D. Kahn, A. Kanapin, M. Krestyaninova, D. Lonsdale, R. Lopez, I. Letunic, M. Madera, J. Maslen, J. McDowall, A. Mitchell, A. N. Nikolskaya, S. Orchard, M. Pagni, C. P. Ponting, E. Quevillon, J. Selengut, C. J. Sigrist, V. Silventoinen, D. J. Studholme, R. Vaughan, and C. H. Wu. 2005. InterPro, progress and status in 2005. Nucleic Acids Res 33:D201-205.

Muller, T., and M. Vingron. 2000. Modeling amino acid replacement. J Comput Biol 7:761-776.

Myers, E. 1999. A Data Set Generator for Whole Genome Shotgun Sequencing. 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), :202-210.

Nedelcu, A. M., R. W. Lee, C. Lemieux, M. W. Gray, and G. Burger. 2000. The complete mitochondrial DNA sequence of *Scenedesmus obliquus* reflects

an intermediate stage in the evolution of the green algal mitochondrial genome. Genome Res 10:819-831.

Nilsson, I., and I. Hoffmann. 2000. Cell cycle regulation by the Cdc25 phosphatase family. Prog Cell Cycle Res 4:107-114.

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

Novoselov, S. V., M. Rao, N. V. Onoshko, H. Zhi, G. V. Kryukov, Y. Xiang, D. P. Weeks, D. L. Hatfield, and V. N. Gladyshev. 2002. Selenoproteins and selenocysteine insertion system in the model plant cell system, *Chlamydomonas reinhardtii*. Embo J 21:3681-3693.

Nozaki, H., O. Misumi, and T. Kuroiwa. 2003. Phylogeny of the quadriflagellate Volvocales (Chlorophyceae) based on chloroplast multigene sequences. Mol Phylogenet Evol 29:58-66.

Nozaki, H., H. Takano, O. Misumi, K. Terasawa, M. Matsuzaki, S. Maruyama, K. Nishida, F. Yagisawa, Y. Yoshida, T. Fujiwara, S. Takio, K. Tamura, S. J. Chung, S. Nakamura, H. Kuroiwa, K. Tanaka, N. Sato, and T. Kuroiwa. 2007. A 100%-complete sequence reveals unusually simple genomic features in the hot-spring red alga *Cyanidioschyzon merolae*. BMC Biol 5:28.

Obaya, A. J., and J. M. Sedivy. 2002. Regulation of cyclin-Cdk activity in mammalian cells. Cell Mol Life Sci 59:126-142.

Oda, K., T. Kohchi, and K. Ohyama. 1992. Mitochondrial DNA of *Marchantia polymorpha* as a single circular form with no incorporation of foreign DNA. Biosci Biotechnol Biochem 56:132-135.

Ogden, C. L., M. D. Carroll, L. R. Curtin, M. A. McDowell, C. J. Tabak, and K. M. Flegal. 2006. Prevalence of overweight and obesity in the United States, 1999-2004. Jama 295:1549-1555.

Ohyama, K., H. Fukuzawa, T. Kohchi, H. Shirai, T. Sano, S. Sano, K. Umesono, Y. Shiki, M. Takeuchi, Z. Chang, S. Aota, H. Inokuchi, and H. Ozeki. 1986. Chloroplast gene organization deduced from complete sequence of liverwort *Marchantia polymorpha* chloroplast DNA. Nature 322:572 - 574.

O'Kelly, C., M. Sieracki, E. Thier, and I. Hobson. 2003. A transient bloom of *Ostreococcus* (Chlorophyta, Prasinophyceae) in West Neck Bay, Long Island, New York. Journal of Phycology 39:850-854.

Oliveira, L., and H. Huynh. 1990. Phototrophic growth of microalgae with allantoic acid or hypoxanthine serving as nitrogen source, implications for purine-N utilization. Can. J. Fish. Aquat. Sci. 47:351-356.

Page, R. D. 1996. TREEVIEW: an application to display phylogenetic trees on personal computers. Comput Appl Biosci 12:357-358.

Palenik, B., B. Brahamsha, F. W. Larimer, M. Land, L. Hauser, P. Chain, J. Lamerdin, W. Regala, E. E. Allen, J. McCarren, I. Paulsen, A. Dufresne, F. Partensky, E. A. Webb, and J. Waterbury. 2003. The genome of a motile marine *Synechococcus*. Nature 424:1037-1042.

Peeken, I. 1997. Photosynthetic pigment fingerprints as indicators of phytoplankton biomass and development in different water masses of the Southern Ocean during austral spring. Deep Sea Research Part II: Topical Studies in Oceanography 44:261-282.

Peers, G., and N. M. Price. 2006. Copper-containing plastocyanin used for electron transport by an oceanic diatom. Nature 441:341-344.

Pellegrini, M., E. M. Marcotte, M. J. Thompson, D. Eisenberg, and T. O. Yeates. 1999. Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. Proc Natl Acad Sci U S A 96:4285-4288.

Peters, T., K. Ausmeier, R. Dildrop, and U. Ruther. 2002. The mouse Fused toes (Ft) mutation is the result of a 1.6-Mb deletion including the entire Iroquois B gene cluster. Mamm Genome 13:186-188.

Peters, T., K. Ausmeier, and U. Ruther. 1999. Cloning of Fatso (*Fto*), a novel gene deleted by the Fused toes (*Ft*) mouse mutation. Mamm Genome 10:983-986.

Petersen, J., R. Teich, B. Becker, R. Cerff, and H. Brinkmann. 2006. The GapA/B gene duplication marks the origin of Streptophyta (charophytes and land plants). Mol Biol Evol 23:1109-1118.

Pfleger, C. M., and M. W. Kirschner. 2000. The KEN box: an APC recognition signal distinct from the D box targeted by Cdh1. Genes Dev 14:655-665.

Picard, A., S. Galas, G. Peaucellier, and M. Doree. 1996. Newly assembled cyclin B-cdc2 kinase is required to suppress DNA replication between meiosis I and meiosis II in starfish oocytes. Embo J 15:3590-3598.

Picard, A., E. Karsenti, M. C. Dabauvalle, and M. Doree. 1987. Release of mature starfish oocytes from interphase arrest by microinjection of human centrosomes. Nature 327:170-172.

Picard, A., and G. Peaucellier. 1998. Behavior of cyclin B and cyclin B-dependent kinase during starfish oocyte meiosis reinitiation: evidence for non-identity with MPF. Biol Cell 90:487-496.

Pohlmeyer, K., B. K. Paap, J. Soll, and N. Wedel. 1996. CP12: a small nuclear-encoded chloroplast protein provides novel insights into higher-plant GAPDH evolution. Plant Mol Biol 32:969-978.

Pombert, J. F., P. Beauchamp, C. Otis, C. Lemieux, and M. Turmel. 2006. The complete mitochondrial DNA sequence of the green alga *Oltmannsiellopsis viridis*: evolutionary trends of the mitochondrial genome in the Ulvophyceae. Curr Genet 50:137-147.

Pombert, J. F., C. Lemieux, and M. Turmel. 2006. The complete chloroplast DNA sequence of the green alga *Oltmannsiellopsis viridis* reveals a distinctive quadripartite architecture in the chloroplast genome of early diverging ulvophytes. BMC Biol 4:3.

Pombert, J. F., C. Otis, C. Lemieux, and M. Turmel. 2004. The complete mitochondrial DNA sequence of the green alga *Pseudendoclonium akinetum* (Ulvophyceae) highlights distinctive evolutionary trends in the chlorophyta and suggests a sister-group relationship between the Ulvophyceae and Chlorophyceae. Mol Biol Evol 21:922-935.

Pombert, J. F., C. Otis, C. Lemieux, and M. Turmel. 2005. The chloroplast genome sequence of the green alga *Pseudendoclonium akinetum* (Ulvophyceae)

reveals unusual structural features and new insights into the branching order of chlorophyte lineages. Mol Biol Evol 22:1903-1918.

Prentice, H. L., and R. E. Kingston. 1992. Mammalian promoter element function in the fission yeast *Schizosaccharomyces pombe*. Nucleic Acids Res 20:3383-3390.

Quesada, A., A. Galvan, R. A. Schnell, P. A. Lefebvre, and E. Fernandez. 1993. Five nitrate assimilation-related loci are clustered in *Chlamydomonas reinhardtii*. Mol Gen Genet 240:387-394.

Ral, J. P., E. Derelle, C. Ferraz, F. Wattebled, B. Farinas, F. Corellou, A. Buleon, M. C. Slomianny, D. Delvalle, C. d'Hulst, S. Rombauts, H. Moreau, and S. Ball. 2004. Starch division and partitioning. A mechanism for granule propagation and maintenance in the picophytoplanktonic green alga *Ostreococcus tauri*. Plant Physiol 136:3333-3340.

Ramesh, M. A., S. B. Malik, and J. M. Logsdon, 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.

Rao, S. K., N. C. Magnin, J. B. Reiskind, and G. Bowes. 2002. Photosynthetic and other phosphoenolpyruvate carboxylase isoforms in the single-cell, facultative C(4) system of *Hydrilla verticillata*. Plant Physiol 130:876-886.

Raven, J. A., A. M. Johnston, J. E. Kubler, R. Korb, S. G. McInroy, L. L. Handley, C. M. Scrimgeour, D. I. Walker, J. Beardall, M. N. Clayton, M. Vanderklift, S. Fredriksen, and K. H. Dunton. 2002. Seaweeds in cold seas: evolution and carbon acquisition. Ann Bot (Lond) 90:525-536.

Reinfelder, J. R., A. J. Milligan, and F. M. Morel. 2004. The role of the C4 pathway in carbon accumulation and fixation in a marine diatom. Plant Physiol 135:2106-2111.

Renaudin, J. P., J. H. Doonan, D. Freeman, J. Hashimoto, H. Hirt, D. Inze, T. Jacobs, H. Kouchi, P. Rouze, M. Sauter, A. Savoure, D. A. Sorrell, V. Sundaresan, and J. A. Murray. 1996. Plant cyclins: a unified nomenclature for plant A-, B- and D-type cyclins based on sequence organization. Plant Mol Biol 32:1003-1018.

Rice, P., I. Longden, and A. Bleasby. 2000. EMBOSS: the European Molecular Biology Open Software Suite. Trends Genet 16:276-277.

Robbens, S., B. Khadaroo, A. Camasses, E. Derelle, C. Ferraz, D. Inze, Y. Van de Peer, and H. Moreau. 2005. Genome-wide analysis of core cell cycle genes in the unicellular green alga *Ostreococcus tauri*. Mol Biol Evol 22:589-597.

Rocap, G., D. L. Distel, J. B. Waterbury, and S. W. Chisholm. 2002. Resolution of *Prochlorococcus* and *Synechococcus* ecotypes by using 16S-23S ribosomal DNA internal transcribed spacer sequences. Appl Environ Microbiol 68:1180-1191.

Rocap, G., F. W. Larimer, J. Lamerdin, S. Malfatti, P. Chain, N. A. Ahlgren, A. Arellano, M. Coleman, L. Hauser, W. R. Hess, Z. I. Johnson, M. Land, D. Lindell, A. F. Post, W. Regala, M. Shah, S. L. Shaw, C. Steglich, M. B. Sullivan, C. S. Ting, A. Tolonen, E. A. Webb, E. R. Zinser, and S. W. Chisholm. 2003. Genome divergence in two *Prochlorococcus* ecotypes reflects oceanic niche differentiation. Nature 424:1042-1047.

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

Rodríguez, F., Y. Pazos, A. Morono, J. Maneiro, and M. Zapata. 2003. Temporal variation in phytoplankton assemblages and pigment composition in a fixed station of the Ria of Pontevedra (NW Spain). Estuarine Coastal and Shelf Science 58:499-515.

Romari, K., and D. Vaulot. 2004. Composition and temporal variability of picoeukaryote communities at a coastal site of the English Channel from 18S rDNA sequences. Limnol. Oceanogr. 49:784–798.

Rouzé, P., N. Pavy, and S. Rombauts. 1999. Genome annotation: which tools do we have for it? Curr Opin Plant Biol 2:90-95.

Rubin, G. M., M. D. Yandell, J. R. Wortman, G. L. Gabor Miklos, C. R. Nelson, I. K. Hariharan, M. E. Fortini, P. W. Li, R. Apweiler, W. Fleischmann, J. M. Cherry, S. Henikoff, M. P. Skupski, S. Misra, M. Ashburner, E. Birney, M. S. Boguski, T. Brody, P. Brokstein, S. E. Celniker, S. A. Chervitz, D. Coates, A. Cravchik, A. Gabrielian, R. F. Galle, W. M. Gelbart, R. A. George, L. S. Goldstein, F. Gong, P. Guan, N. L. Harris, B. A. Hay, R. A. Hoskins, J. Li, Z. Li, R. O. Hynes, S. J. Jones, P. M. Kuehl, B. Lemaitre, J. T. Littleton, D. K. Morrison, C. Mungall, P. H. O'Farrell, O. K. Pickeral, C. Shue, L. B. Vosshall, J. Zhang, Q. Zhao, X. H. Zheng, and S. Lewis. 2000. Comparative genomics of the eukaryotes. Science 287:2204-2215.

Rusch, D. B., A. L. Halpern, G. Sutton, K. B. Heidelberg, S. Williamson, S. Yooseph, D. Wu, J. A. Eisen, J. M. Hoffman, K. Remington, K. Beeson, B. Tran, H. Smith, H. Baden-Tillson, C. Stewart, J. Thorpe, J. Freeman, C. Andrews-Pfannkoch, J. E. Venter, K. Li, S. Kravitz, J. F. Heidelberg, T. Utterback, Y. H. Rogers, L. I. Falcon, V. Souza, G. Bonilla-Rosso, L. E. Eguiarte, D. M. Karl, S. Sathyendranath, T. Platt, E. Bermingham, V. Gallardo, G. Tamayo-Castillo, M. R. Ferrari, R. L. Strausberg, K. Nealson, R. Friedman, M. Frazier, and J. C. Venter. 2007. The Sorcerer II Global Ocean Sampling Expedition: Northwest Atlantic through Eastern Tropical Pacific. PLoS Biol 5:e77.

Russell, P., and P. Nurse. 1986. cdc25+ functions as an inducer in the mitotic control of fission yeast. Cell 45:145-153.

Rutherford, K., J. Parkhill, J. Crook, T. Horsnell, P. Rice, M. A. Rajandream, and B. Barrell. 2000. ARTEMIS: sequence visualization and annotation. Bioinformatics 16:944-945.

Sage, J., A. L. Miller, P. A. Perez-Mancera, J. M. Wysocki, and T. Jacks. 2003. Acute mutation of retinoblastoma gene function is sufficient for cell cycle re-entry. Nature 424:223-228.

Sage, R. 2004. The evolution of C4 photosynthesis. New Phytologist 161:341.

Saitou, N., and M. Nei. 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol Biol Evol 4:406-425.

Salamov, A. A., and V. V. Solovyev. 2000. Ab initio gene finding in *Drosophila* genomic DNA. Genome Res 10:516-522.

Sanchez-Diaz, A., I. Gonzalez, M. Arellano, and S. Moreno. 1998. The Cdk inhibitors p25rum1 and p40SIC1 are functional homologues that play

similar roles in the regulation of the cell cycle in fission and budding yeast. J Cell Sci 111 (Pt 6):843-851.

Sanderson, M. J. 2003. Molecular data from 27 proteins do not support a Precambrian origin of land plants. American Journal of Botany 90:954-956.

Sato, S., Y. Nakamura, T. Kaneko, E. Asamizu, and S. Tabata. 1999. Complete structure of the chloroplast genome of *Arabidopsis thaliana*. DNA Res 6:283-290.

Scheibe, R., N. Wedel, S. Vetter, V. Emmerlich, and S. M. Sauermann. 2002. Co-existence of two regulatory NADP-glyceraldehyde 3-P dehydrogenase complexes in higher plant chloroplasts. Eur J Biochem 269:5617-5624.

Schiex, T., A. Moisan, and P. Rouze. 2001. EuGene: an eucaryotic gene finder that combines several sources of evidence. O. Gascuel and M.-F. Sagot, eds. Computational Biology: Selected Papers (Lecture Notes in Computer Science) 2066:111-125.

Schmidt, H. A., K. Strimmer, M. Vingron, and A. von Haeseler. 2002. TREE-PUZZLE: maximum likelihood phylogenetic analysis using quartets and parallel computing. Bioinformatics 18:502-504.

Schnarrenberger, C., A. Flechner, and W. Martin. 1995. Enzymatic Evidence for a Complete Oxidative Pentose Phosphate Pathway in Chloroplasts and an Incomplete Pathway in the Cytosol of Spinach Leaves. Plant Physiol 108:609-614.

Schwartz, S., Z. Zhang, K. A. Frazer, A. Smit, C. Riemer, J. Bouck, R. Gibbs, R. Hardison, and W. Miller. 2000. PIPMAKER—a web server for aligning two genomic DNA sequences. Genome Res 10:577-586.

Schwickart, M., J. Havlis, B. Habermann, A. Bogdanova, A. Camasses, T. Oelschlaegel, A. Shevchenko, and W. Zachariae. 2004. Swm1/Apc13 is an evolutionarily conserved subunit of the anaphase-promoting complex stabilizing the association of Cdc16 and Cdc27. Mol Cell Biol 24:3562-3576.

Schwob, E., T. Bohm, M. D. Mendenhall, and K. Nasmyth. 1994. The B-type cyclin kinase inhibitor p40SIC1 controls the G1 to S transition in *S. cerevisiae*. Cell 79:233-244.

Scott, L. J., K. L. Mohlke, L. L. Bonnycastle, C. J. Willer, Y. Li, W. L. Duren, M. R. Erdos, H. M. Stringham, P. S. Chines, A. U. Jackson, L. Prokunina-Olsson, C. J. Ding, A. J. Swift, N. Narisu, T. Hu, R. Pruim, R. Xiao, X. Y. Li, K. N. Conneely, N. L. Riebow, A. G. Sprau, M. Tong, P. P. White, K. N. Hetrick, M. W. Barnhart, C. W. Bark, J. L. Goldstein, L. Watkins, F. Xiang, J. Saramies, T. A. Buchanan, R. M. Watanabe, T. T. Valle, L. Kinnunen, G. R. Abecasis, E. W. Pugh, K. F. Doheny, R. N. Bergman, J. Tuomilehto, F. S. Collins, and M. Boehnke. 2007. A genome-wide association study of type 2 diabetes in Finns detects multiple susceptibility variants. Science 316:1341-1345.

Scuteri, A., S. Sanna, W. M. Chen, M. Uda, G. Albai, J. Strait, S. Najjar, R. Nagaraja, M. Orru, G. Usala, M. Dei, S. Lai, A. Maschio, F. Busonero, A. Mulas, G. B. Ehret, A. A. Fink, A. B. Weder, R. S. Cooper, P. Galan, A. Chakravarti, D. Schlessinger, A. Cao, E. Lakatta, and G. R. Abecasis. 2007.

Genome-Wide Association Scan Shows Genetic Variants in the *FTO* Gene Are Associated with Obesity-Related Traits. PLoS Genet 3:e115.

Shen, W. H. 2002. The plant E2F-Rb pathway and epigenetic control. Trends Plant Sci 7:505-511.

Shi, L., M. Potts, and P. J. Kennelly. 1998. The serine, threonine, and/or tyrosine-specific protein kinases and protein phosphatases of prokaryotic organisms: a family portrait. FEMS Microbiol Rev 22:229-253.

Shinozaki, K., M. Ohme, M. Tanaka, T. Wakasugi, N. Hayashida, T. Matsubayashi, N. Zaita, J. Chunwongse, J. Obokata, K. Yamaguchi-Shinozaki, C. Ohto, K. Torazawa, B. Y. Meng, M. Sugita, H. Deno, T. Kamogashira, K. Yamada, J. Kusuda, F. Takaiwa, A. Kato, N. Tohdoh, H. Shimada, and M. Sugiura. 1986. The complete nucleotide sequence of the tobacco chloroplast genome: its gene organization and expression. Embo J 5:2043-2049.

Shizuya, H., B. Birren, U. J. Kim, V. Mancino, T. Slepak, Y. Tachiiri, and M. Simon. 1992. Cloning and stable maintenance of 300-kilobase-pair fragments of human DNA in *Escherichia coli* using an F-factor-based vector. Proc Natl Acad Sci U S A 89:8794-8797.

Siaut, M., M. Heijde, M. Mangogna, A. Montsant, S. Coesel, A. Allen, A. Manfredonia, A. Falciatore, and C. Bowler. 2007. Molecular toolbox for studying diatom biology in *Phaeodactylum tricornutum*. Gene.

Silver, M. W., M. M. Gowing, and P. J. Davoll. 1986. The association of photosynthetic picoplankton and ultraplankton with pelagic detritus through the water column (0-200m). Can. Bull. Aquat. Sci. 214:311-341.

Simillion, C., K. Vandepoele, Y. Saeys, and Y. Van de Peer. 2004. Building genomic profiles for uncovering segmental homology in the twilight zone. Genome Res 14:1095-1106.

Six, C., A. Z. Worden, F. Rodriguez, H. Moreau, and F. Partensky. 2005. New insights into the nature and phylogeny of prasinophyte antenna proteins: *Ostreococcus tauri*, a case study. Mol Biol Evol 22:2217-2230.

Smith, L. M., J. Z. Sanders, R. J. Kaiser, P. Hughes, C. Dodd, C. R. Connell, C. Heiner, S. B. Kent, and L. E. Hood. 1986. Fluorescence detection in automated DNA sequence analysis. Nature 321:674-679.

Smits, V. A., and R. H. Medema. 2001. Checking out the G(2)/M transition. Biochim Biophys Acta 1519:1-12.

Sohn, J., and J. Rudolph. 2003. Catalytic and chemical competence of regulation of cdc25 phosphatase by oxidation/reduction. Biochemistry 42:10060-10070.

Sonnhammer, E. L., and R. Durbin. 1995. A dot-matrix program with dynamic threshold control suited for genomic DNA and protein sequence analysis. Gene 167:GC1-10.

Sorrell, D. A., A. Marchbank, K. McMahon, J. R. Dickinson, H. J. Rogers, and D. Francis. 2002. A WEE1 homologue from *Arabidopsis thaliana*. Planta 215:518-522.

Stals, H., and D. Inze. 2001. When plant cells decide to divide. Trends Plant Sci 6:359-364.

Stegemann, S., S. Hartmann, S. Ruf, and R. Bock. 2003. High-frequency gene transfer from the chloroplast genome to the nucleus. Proc Natl Acad Sci U S A 100:8828-8833.

Stirewalt, V. L., C. B. Michalowski, W. Löffelhardt, H. J. Bohnert, and D. B. Bryant. 1995. Nucleotide sequence of the cyanelle genome from *Cyanophora paradoxa*. Plant Mol. Biol. Rep. 13:327-332.

Stockner, J. G. 1988. Phototrophic Picoplankton: An overview from marine and freshwater ecosystems. Limnology and Oceanography 33:765-775.

Stoesser, G., W. Baker, A. van den Broek, E. Camon, M. Garcia-Pastor, C. Kanz, T. Kulikova, R. Leinonen, Q. Lin, V. Lombard, R. Lopez, N. Redaschi, P. Stoehr, M. A. Tuli, K. Tzouvara, and R. Vaughan. 2002. The EMBL Nucleotide Sequence Database. Nucleic Acids Res 30:21-26.

Sugiyama, Y., Y. Watase, M. Nagase, N. Makita, S. Yagura, A. Hirai, and M. Sugiura. 2005. The complete nucleotide sequence and multipartite organization of the tobacco mitochondrial genome: comparative analysis of mitochondrial genomes in higher plants. Mol Genet Genomics 272:603-615.

Sun, Y., B. P. Dilkes, C. Zhang, R. A. Dante, N. P. Carneiro, K. S. Lowe, R. Jung, W. J. Gordon-Kamm, and B. A. Larkins. 1999. Characterization of maize (*Zea mays L.*) Wee1 and its activity in developing endosperm. Proc Natl Acad Sci U S A 96:4180-4185.

Tamoi, M., T. Miyazaki, T. Fukamizo, and S. Shigeoka. 2005. The Calvin cycle in cyanobacteria is regulated by CP12 via the NAD(H)/NADP(H) ratio under light/dark conditions. Plant J 42:504-513.

Tarayre, S., J. M. Vinardell, A. Cebolla, A. Kondorosi, and E. Kondorosi. 2004. Two classes of the CDh1-type activators of the anaphase-promoting complex in plants: novel functional domains and distinct regulation. Plant Cell 16:422-434.

Tateno, Y., T. Imanishi, S. Miyazaki, K. Fukami-Kobayashi, N. Saitou, H. Sugawara, and T. Gojobori. 2002. DNA Data Bank of Japan (DDBJ) for genome scale research in life science. Nucleic Acids Res 30:27-30.

Tesler, G. 2002. GRIMM: genome rearrangements web server. Bioinformatics 18:492-493.

The *C. elegans* sequencing consortium. 1998. Genome sequence of the nematode C. elegans: a platform for investigating biology. Science 282:2012-2018.

The yeast sequencing consortium. 1997. The yeast genome directory. Nature 387:5.

Thompson, J. D., D. G. Higgins, and T. J. Gibson. 1994. CLUSTALW: 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.

Turmel, M., C. Lemieux, G. Burger, B. F. Lang, C. Otis, I. Plante, and M. W. Gray. 1999. The complete mitochondrial DNA sequences of *Nephroselmis olivacea* and *Pedinomonas minor*. Two radically different evolutionary patterns within green algae. Plant Cell 11:1717-1730.

Turmel, M., C. Otis, and C. Lemieux. 1999. The complete chloroplast DNA sequence of the green alga *Nephroselmis olivacea*: insights into the

architecture of ancestral chloroplast genomes. Proc Natl Acad Sci U S A 96:10248-10253.

Turmel, M., C. Otis, and C. Lemieux. 2002. The complete mitochondrial DNA sequence of *Mesostigma viride* identifies this green alga as the earliest green plant divergence and predicts a highly compact mitochondrial genome in the ancestor of all green plants. Mol Biol Evol 19:24-38.

Turmel, M., C. Otis, and C. Lemieux. 2002. The chloroplast and mitochondrial genome sequences of the charophyte *Chaetosphaeridium globosum*: insights into the timing of the events that restructured organelle DNAs within the green algal lineage that led to land plants. Proc Natl Acad Sci U S A 99:11275-11280.

Turmel, M., C. Otis, and C. Lemieux. 2003. The mitochondrial genome of *Chara vulgaris*: insights into the mitochondrial DNA architecture of the last common ancestor of green algae and land plants. Plant Cell 15:1888-1903.

Tzivion, G., and J. Avruch. 2002. 14-3-3 proteins: active cofactors in cellular regulation by serine/threonine phosphorylation. J Biol Chem 277:3061-3064.

Umen, J. G., and U. W. Goodenough. 2001. Control of cell division by a retinoblastoma protein homolog in *Chlamydomonas*. Genes Dev 15:1652-1661.

Unseld, M., J. R. Marienfeld, P. Brandt, and A. Brennicke. 1997. The mitochondrial genome of *Arabidopsis thaliana* contains 57 genes in 366,924 nucleotides. Nat Genet 15:57-61.

Uzzell, T., and K. W. Corbin. 1971. Fitting discrete probability distributions to evolutionary events. Science 172:1089-1096.

Vahrenholz, C., G. Riemen, E. Pratje, B. Dujon, and G. Michaelis. 1993. Mitochondrial DNA of *Chlamydomonas reinhardtii*: the structure of the ends of the linear 15.8-kb genome suggests mechanisms for DNA replication. Curr Genet 24:241-247.

Van de Peer, Y., and R. De Wachter. 1997. Construction of evolutionary distance trees with TREECON for Windows: accounting for variation in nucleotide substitution rate among sites. Comput Appl Biosci 13:227-230.

Van de Peer, Y., and R. De Wachter. 1994. TREECON for Windows: a software package for the construction and drawing of evolutionary trees for the Microsoft Windows environment. Comput Appl Biosci 10:569-570.

Vandepoele, K., J. Raes, L. De Veylder, P. Rouze, S. Rombauts, and D. Inze. 2002. Genome-wide analysis of core cell cycle genes in *Arabidopsis*. Plant Cell 14:903-916.

van der Hoeven, F., T. Schimmang, A. Volkmann, M. G. Mattei, B. Kyewski, and U. Ruther. 1994. Programmed cell death is affected in the novel mouse mutant Fused toes (Ft). Development 120:2601-2607.

Vee, S., L. Lafanechere, D. Fisher, J. Wehland, D. Job, and A. Picard. 2001. Evidence for a role of the (alpha)-tubulin C terminus in the regulation of cyclin B synthesis in developing oocytes. J Cell Sci 114:887-898.

Venables, W. N., and B. D. Ripley. 2002. Modern Applied Statistics with S (Springer, New York).

Venter, J. C., K. Remington, J. F. Heidelberg, A. L. Halpern, D. Rusch, J. A. Eisen, D. Wu, I. Paulsen, K. E. Nelson, W. Nelson, D. E. Fouts, S. Levy, A. H. Knap, M. W. Lomas, K. Nealson, O. White, J. Peterson, J. Hoffman, R. Parsons, H. Baden-Tillson, C. Pfannkoch, Y. H. Rogers, and H. O. Smith. 2004. Environmental genome shotgun sequencing of the Sargasso Sea. Science 304:66-74.

Venter, J. C., H. O. Smith, and L. Hood. 1996. A new strategy for genome sequencing. Nature 381:364-366.

Wakasugi, T., T. Nagai, M. Kapoor, M. Sugita, M. Ito, S. Ito, J. Tsudzuki, K. Nakashima, T. Tsudzuki, Y. Suzuki, A. Hamada, T. Ohta, A. Inamura, K. Yoshinaga, and M. Sugiura. 1997. Complete nucleotide sequence of the chloroplast genome from the green alga *Chlorella vulgaris*: the existence of genes possibly involved in chloroplast division. Proc Natl Acad Sci U S A 94:5967-5972.

Wakasugi, T., J. Tsudzuki, S. Ito, K. Nakashima, T. Tsudzuki, and M. Sugiura. 1994. Loss of all ndh genes as determined by sequencing the entire chloroplast genome of the black pine *Pinus thunbergii*. Proc Natl Acad Sci U S A 91:9794-9798.

Wedel, N., and J. Soll. 1998. Evolutionary conserved light regulation of Calvin cycle activity by NADPH-mediated reversible phosphoribulokinase/CP12/glyceraldehyde-3-phosphate dehydrogenase complex dissociation. Proc Natl Acad Sci U S A 95:9699-9704.

Wedel, N., J. Soll, and B. K. Paap. 1997. CP12 provides a new mode of light regulation of Calvin cycle activity in higher plants. Proc Natl Acad Sci U S A 94:10479-10484.

Weinberg, R. A. 1995. The retinoblastoma protein and cell cycle control. Cell 81:323-330.

Wohl, T., M. Brecht, F. Lottspeich, and H. Ammer. 1995. The use of genomic DNA probes for in-gel hybridization. Electrophoresis 16:739-741.

Wolosiuk, R. A., M. A. Ballicora, and K. Hagelin. 1993. The reductive pentose phosphate cycle for photosynthetic CO2 assimilation: enzyme modulation. Faseb J 7:622-637.

Worden, A. 2006. Picoeukaryote diversity in coastal waters of the Pacific Ocean. Aquatic Microbial Ecology 43:165-175.

Worden, A., J. Nolan, and B. Palenik. 2004. Assessing the dynamics and ecology of marine picophytoplankton: the importance of the eukaryotic component. Limnology and Oceanography 49:168-179.

Wuyts, J., G. Perriere, and Y. Van De Peer. 2004. The European ribosomal RNA database. Nucleic Acids Res 32:D101-103.

Yoon, H. S., J. D. Hackett, C. Ciniglia, G. Pinto, and D. Bhattacharya. 2004. A molecular timeline for the origin of photosynthetic eukaryotes. Mol Biol Evol 21:809-818.

Zeidner, G., C. M. Preston, E. F. Delong, R. Massana, A. F. Post, D. J. Scanlan, and O. Beja. 2003. Molecular diversity among marine picophytoplankton as revealed by psbA analyses. Environ Microbiol 5:212-216.

Zhang, K., D. S. Letham, and P. C. John. 1996. Cytokinin controls the cell cycle at mitosis by stimulating the tyrosine dephosphorylation and activation of p34cdc2-like H1 histone kinase. Planta 200:2-12.

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

Zingone, A., M. Borra, C. Brunet, G. Forlani, W. Kooistra, and G. Procaccini. 2002. Phylogenetic position of *Crustomastix stigmatica sp nov.* and *Dolichomastix tenuilepis* in relation to the mamiellales (Prasinophyceae, Chlorophyta). Journal of Phycology 38:1024-1039.

Zito, F., J. Vinh, J. L. Popot, and G. Finazzi. 2002. Chimeric fusions of subunit IV and PetL in the b6f complex of *Chlamydomonas reinhardtii*: structural implications and consequences on state transitions. J Biol Chem 277:12446-12455.

Zuckerkandl, E., and L. Pauling. 1965. Molecules as documents of evolutionary history. J Theor Biol 8:357-366.

Zur, A., and M. Brandeis. 2002. Timing of APC/C substrate degradation is determined by fzy/fzr specificity of destruction boxes. Embo J 21:4500-4510.

Appendix

Figure A1

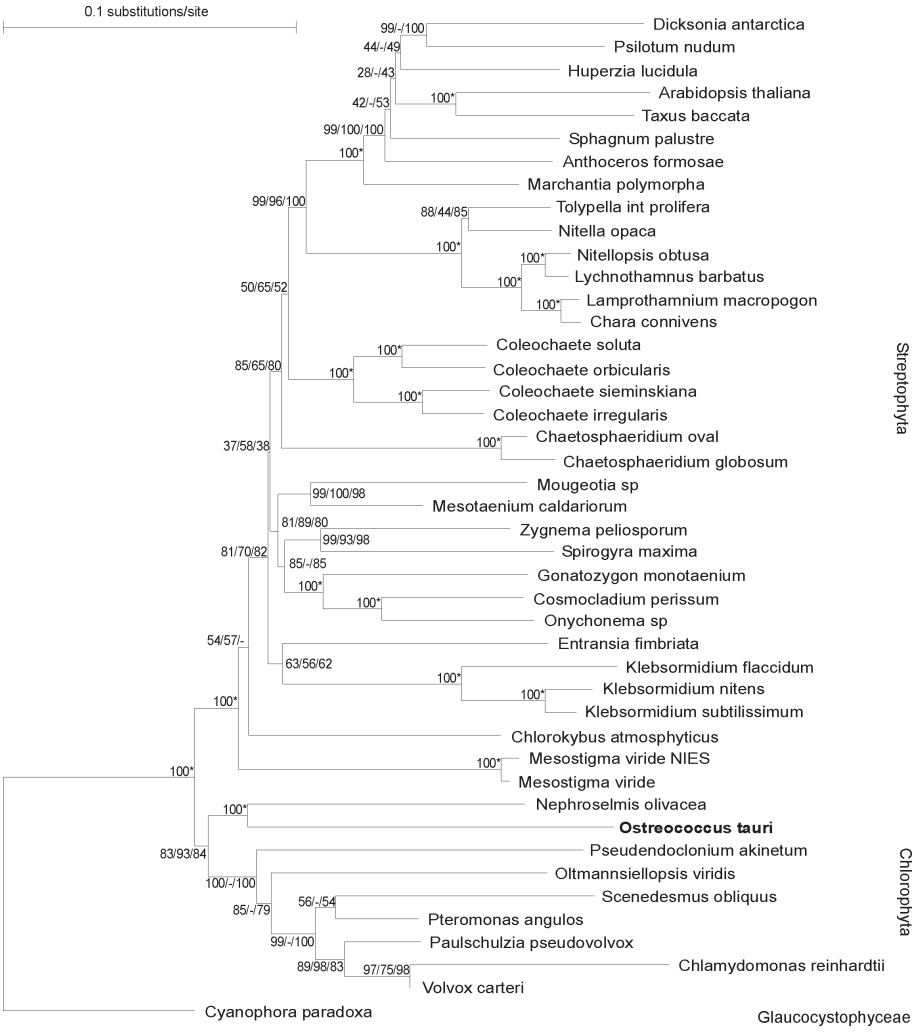

0.1 substitutions/site
Dicksonia antarctica
Psilotum nudum
Huperzia lucidula
Arabidopsis thaliana
Taxus baccata
Sphagnum palustre
Anthoceros formosae
Marchantia polymorpha
Tolypella int prolifera
Nitella opaca
Nitellopsis obtusa
Lychnothamnus barbatus
Lamprothamnium macropogon
Chara connivens
Coleochaete soluta
Coleochaete orbicularis
Coleochaete sieminskiana
Coleochaete irregularis
Chaetosphaeridium oval
Chaetosphaeridium globosum
Mougeotia sp
Mesotaenium caldariorum
Zygnema peliosporum
Spirogyra maxima
Gonatozygon monotaenium
Cosmocladium perissum
Onychonema sp
Entransia fimbriata
Klebsormidium flaccidum
Klebsormidium nitens
Klebsormidium subtilissimum
Chlorokybus atmosphyticus
Mesostigma viride NIES
Mesostigma viride
Nephroselmis olivacea
Ostreococcus tauri
Pseudendoclonium akinetum
Oltmannsiellopsis viridis
Scenedesmus obliquus
Pteromonas angulos
Paulschulzia pseudovolvox
Chlamydomonas reinhardtii
Volvox carteri
Cyanophora paradoxa
Streptophyta
Chlorophyta
Glaucocystophyceae

Figure A2

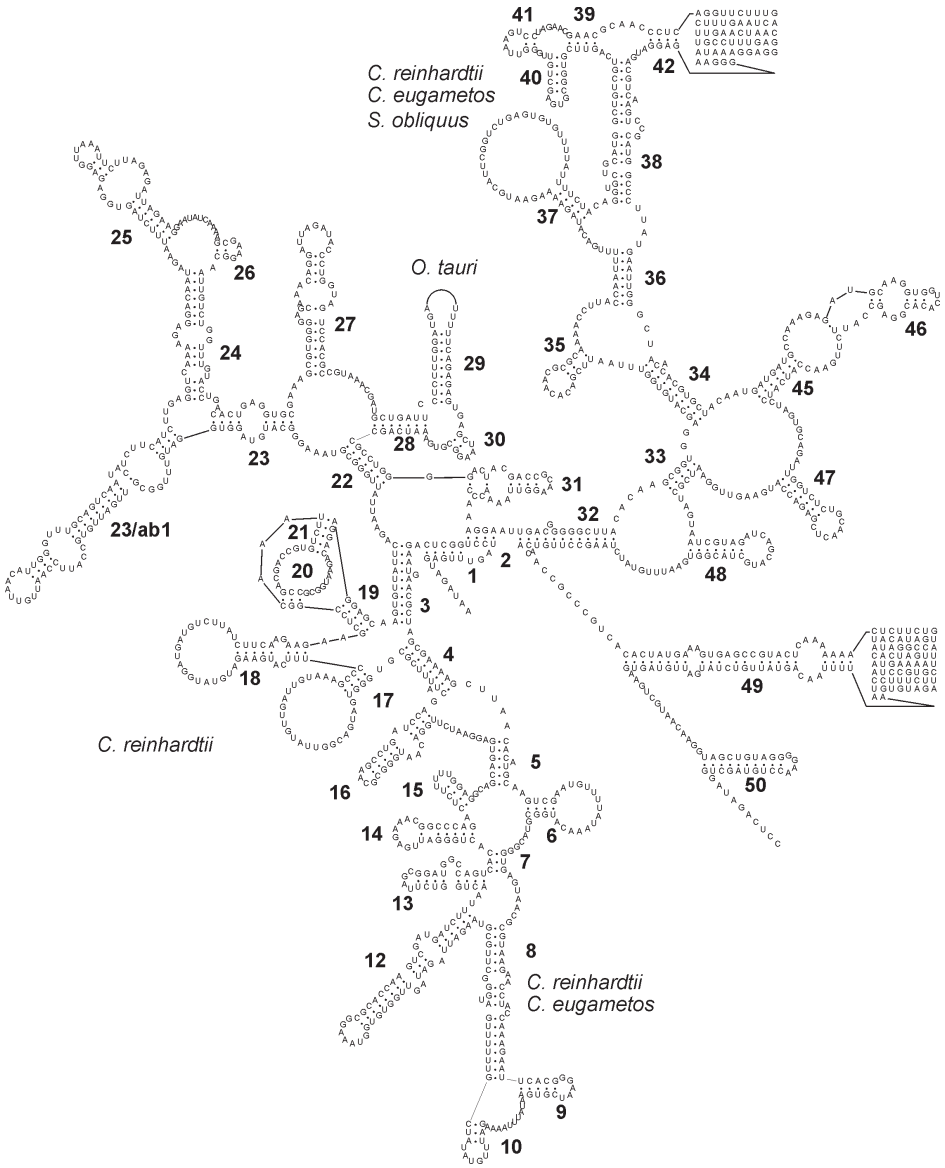

C. reinhardtii
C. eugametos
S. obliquus
O. tauri
C. reinhardtii
C. reinhardtii
C. eugametos
23/ab1

Table A3

Codon usage table of the mitochondrial genes

Codon	%	Anticodon	AA	Codon	%	Anticodon	AA	Codon	%	Anticodon	AA	Codon	%	Anticodon	AA
uuu	73.3	gaa	F	ucu	23.7	uga	S	uau	71.9	gua	Y	ugu	84.6	gca	C
uuc	26.7	gaa	F	ucc	8.9	uga	S	uac	28.1	gua	Y	ugc	15.4	gca	C
uua	39	uaa	L	uca	24.9	uga	S	uaa	81.4		*	uga	13.9		*
uug	9	uaa	L	ucg	8.9	uga	S	uag	4.7		*	ugg	100	cca	W
cuu	26.5	uag	L	ccu	28.2	ugg	P	cau	57.5	gug	H	cgu	28	acg	R
cuc	13.9	gag	L	ccc	6.6	ugg	P	cac	42.5	gug	H	cgc	7.1	acg	R
cua	5.3	uag	L	cca	51.9	ugg	P	caa	82.1	uug	Q	cga	23	acg	R
cug	6.3	uag	L	ccg	13.3	ugg	P	cag	17.9	uug	Q	cgg	1.2	acg	R
auu	74.2	gau	I	acu	21.3	ggu	T	aau	63	guu	N	agu	29	gcu	S
auc	25.1	gau	I	acc	12.9	ggu	T	aac	37	guu	N	agc	4.6	gcu	S
aua	0.7	gau	I	aca	53.1	ggu	T	aaa	84.7	uuu	K	aga	35.1	ucu	R
aug	100	cau	M	acg	12.7	ggu	T	aag	15.3	uuu	K	agg	5.6	ucu	R
guu	36.3	uac	V	gcu	31.2	ugc	A	gau	72.3	guc	D	ggu	32	gcc + ucc	G
guc	19.4	uac	V	gcc	13.6	ugc	A	gac	27.7	guc	D	ggc	7.1	gcc + ucc	G
gua	21.3	uac	V	gca	40.7	ugc	A	gaa	79.8	uuc	E	gga	40.4	ucc	G
gug	23	uac	V	gcg	14.5	ugc	A	gag	20.2	uuc	E	ggg	20.5	ucc	G

Table A4

Codon usage table of the chloroplast genes

Codon	%	Anticodon	AA	Codon	%	Anticodon	AA	Codon	%	Anticodon	AA	Codon	%	Anticodon	AA
uuu	50.1	gaa	F	ucu	23.4	uga	S	uau	50	gua	Y	ugu	85.9	gca	C
uuc	49.9	gaa	F	ucc	4.6	uga	S	uac	50	gua	Y	ugc	14.1	gca	C
uua	39.6	uaa	L	uca	38	uga	S	uaa	95		*	uga	0		*
uug	3.6	uaa	L	ucg	8.6	uga	S	uag	5		*	ugg	100	cca	W
cuu	45.7	uag	L	ccu	26.3	ugg	P	cau	27	gug	H	cgu	76.7	acg	R
cuc	8.7	gag	L	ccc	4.2	ugg	P	cac	73	gug	H	cgc	9.5	acg	R
cua	1.1	uag	L	cca	54.6	ugg	P	caa	78.7	uug	Q	cga	11	acg	R
cug	1.3	uag	L	ccg	14.9	ugg	P	cag	21.3	uug	Q	cgg	0.8	acg	R
auu	76.5	gau	I	acu	39.9	ggu	T	aau	47.2	guu	N	agu	21.3	gcu	S
auc	23.1	gau	I	acc	4.3	ggu	T	aac	52.8	guu	N	agc	4.1	gcu	S
aua	0.4	gau	I	aca	43.8	ggu	T	aaa	71.9	uuu	K	aga	1.7	ucu	R
aug	100	cau	M	acg	12	ggu	T	aag	28.1	uuu	K	agg	0.3	ucu	R
guu	34.8	uac	V	gcu	39	ugc	A	gau	68	guc	D	ggu	40.9	ucc	G
guc	4.4	uac	V	gcc	6	ugc	A	gac	32	guc	D	ggc	3.5	ucc	G
gua	52.1	uac	V	gca	45.2	ugc	A	gaa	69.6	uuc	E	gga	43.3	ucc	G
gug	8.7	uac	V	gcg	9.8	ugc	A	gag	30.4	uuc	E	ggg	12.3	ucc	G

Publications and author contribution

Khadaroo, B., Robbens, S., Ferraz, C., Derelle, E., Eychenie, S., Cooke, R., Peaucellier, G., Delseny, M., Demaille, J., Van de Peer, Y., Picard, A., Moreau, H. (2004) The First Green Lineage cdc25 Dual-Specificity Phosphatase. Cell Cycle 3, 513-8.

As a second author I did all the bioinformatics analyses: screening the nuclear genome for the presence or absence of the Cdc25 gene, followed by making phylogenetic trees for the in silico validation part. I also wrote the text and made the figures regarding the bioinformatics work I did.

Robbens, S., Khadaroo, B., Camasses, A., Derelle, E., Ferraz, C., Inzé, D., Van de Peer, Y., Moreau, H. (2005) Genome-wide analysis of core cell cycle genes in the unicellular green alga *Ostreococcus tauri*. Mol. Biol. Evol. 22, 589-97.

Derelle, E., Ferraz, C., Rombauts, S., Rouzé, P., Worden, A.Z., Robbens, S., Partensky, F., Degroeve, S., Echeynie, S., Cooke, R., Saeys, Y., Wuyts, J., Jabbari, K., Bowler, C., Panaud, O., Piegu, B., Ball, S., Ral, J.P., Bouget, F.-Y., Piganeau, G., De Baets, B., Picard, A., Delseny, M., Demaille, J., Van de Peer, Y., Moreau, H. (2006) Genome analysis of the smallest free-living eukaryote *Ostreococcus tauri* unveils many unique features. Proc. Natl. Acad. Sci. USA 103, 11647-52.

As the third author on this paper, I was involved in the initial creation of a manual annotated gene dataset, which was later used as a training and validation set for and after the automatic annotation. I also performed the taxon distribution analyses: blast searches combined with phylogenetic trees. Concerning the manuscript: I helped writing the paper and I made most of the figures.

Robbens, S., Derelle, E., Ferraz, C., Wuyts, J., Moreau, H., Van de Peer, Y. (2007) The Complete Chloroplast and Mitochondrial DNA Sequence of *Ostreococcus tauri*: Organelle Genomes of the Smallest Eukaryote are Examples of Compaction. Mol. Biol. Evol. 24, 956-68

Robbens, S., Petersen, J., Brinkmann, H., Rouzé, P., Van de Peer, Y. (2007) Unique regulation of the Calvin cycle in the ultrasmall green alga *Ostreococcus*. J. Mol. Evol. 64, 601-4.

Palenik, B., Grimwood, J., Aerts, A., Rouzé, P., Salamov, A., Putnam, N., Zhou, K., Jorgensen, R., Derelle, E., Rombauts, S., Otillar, R., Merchant, S., Podell, S., Gaasterland, T., Manuell, A., Napoli, C., Gendler, K., Vallon, O., Peyretaillade, E., Jancek, S., Piganeau, G., Heijde, M., Lohr, M., Jabbari, K., Bowler, C., Werner, G., Robbens, S., Pazour, G., Dubchak, I., Ren, Q., Delwiche, C., Paulsen, I., De Boever, P., Schmutz, J., Rokhsar, D., Van de Peer, Y., Moreau, H., Grigoriev, I. (2007) The Tiny Eukaryote *Ostreococcus* Provides Genomic Insights Into The Paradox Of Plankton Speciation. Proc. Natl. Acad. Sci. USA 104, 7705-10.

As one of the many authors, I performed some initial analyses: looking at the synteny present or absent in both genomes and manually annotating certain genes. I also helped making the first figure.

Robbens, S., Rouzé, P., Cock, M.J., Spring, J., Worden, A.Z., and Van de Peer, Y. (2007) The *FTO* gene, implicated in human obesity, is found only in vertebrates and marine algae. J. Mol. Evol. (in press).

Vandenbroucke, K., Robbens, S., Vandepoele, K., Inzé, D., Van de Peer, Y., and Van Breusegem, F. (2007) H_2O_2-Induced Gene Expression across Kingdoms: A Comparative Analysis. Mol. Biol. Evol. (accepted).

Scientific activity

Oral presentation

"Unique findings in the green alga *Ostreococcus tauri*" - ESF-EMBO Symposium on Comparative Genomics of Eukaryotic Microorganisms: Eukaryotic Genome Evolution.
Sant Feliu de Guixols, Spain, 20-25 October 2007

Awarded Price for Best Oral Presentation

Poster presentation

"Annotation of the green alga *Ostreococcus tauri*" - Belgian Bioinformatics Conference
Brussels, 23 April 2004

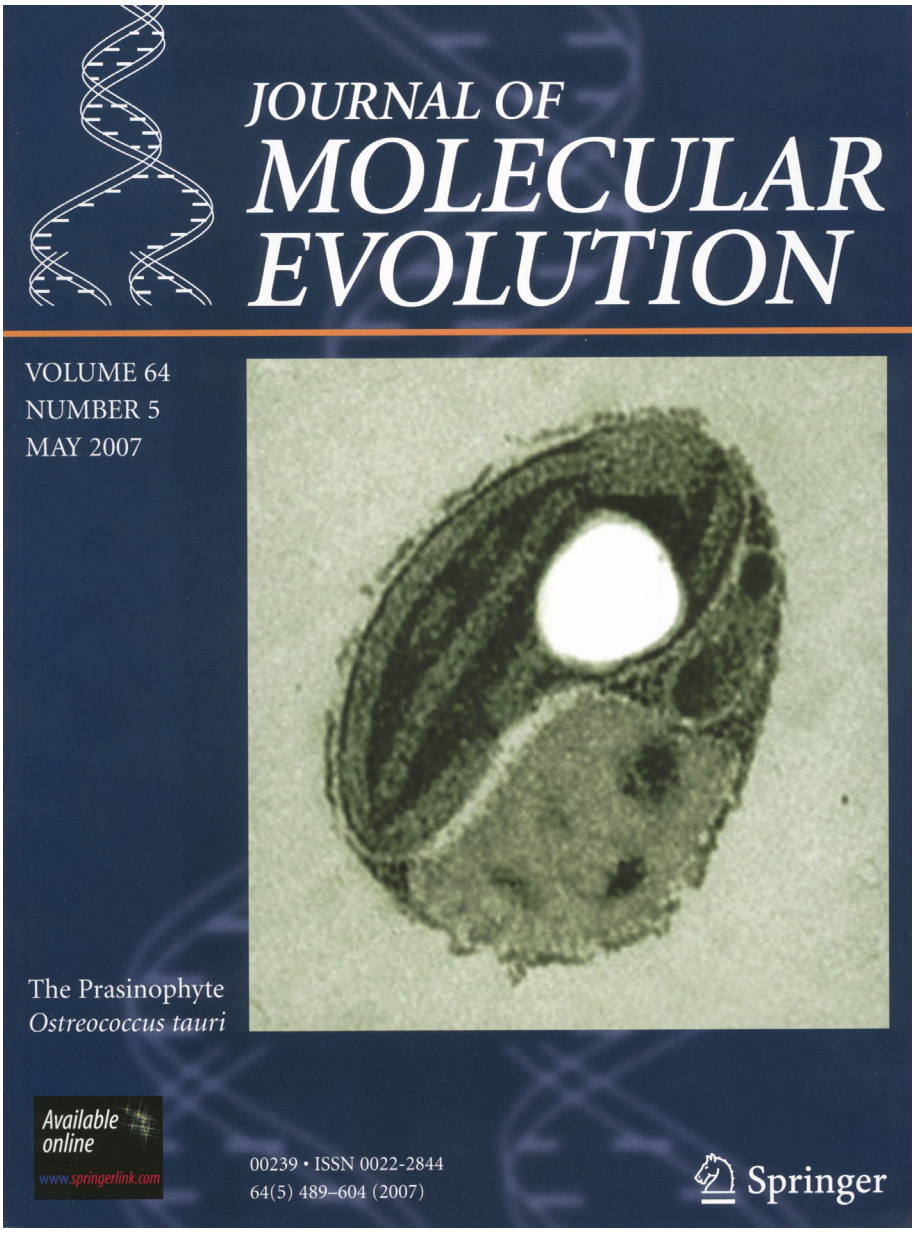
"Genome analysis of the world's smallest free-living eukaryote *Ostrecococcus tauri* unveils unique genome heterogeneity" - Molecular Biology and Evolution Conference
Auckland, New Zealand, 19-23 June 2005

"The ultrasmall green alga *Ostreococcus* unveils a unique regulation of the Calvin cycle" - Gent - Lille Workshop on Computational Biology
Lille, France, 20 June 2006

Award


ESF-EMBO Symposium on Comparative Genomics of Eukaryotic Microorganisms: Eukaryotic Genome Evolution
Sant Feliu de Guixols, Spain, 20-25 October 2007
Contributed Presentation Prize
is hereby granted to:
Steven Robbens
for outstanding contributed presentation
Awarded: October 25, 2007
Professor Ed Louis
Professor Teun Boekhout
EMBO
EUROPEAN SCIENCE FOUNDATION
SETTING SCIENCE AGENDAS FOR EUROPE

Cover


JOURNAL OF
MOLECULAR
EVOLUTION
VOLUME 64
NUMBER 5
MAY 2007
The Prasinophyte
Ostreococcus tauri
Available online
www.springerlink.com
00239 • ISSN 0022-2844
64(5) 489–604 (2007)
Springer