“Transcriptional and Epigenetic regulation in the marine diatom Phaeodactylum tricornutum”

Florian Maumus

Thèse de Doctorat

“Transcriptional and Epigenetic regulation in the marine diatom *Phaeodactylum tricornutum*”

Présentée par:

Florian Maumus

Soutenance le 6 juillet 2009 devant les membres du jury:

Prof. Martine Boccara

Dr. Chris Bowler

Dr. Pascale Lesage

Prof. Olivier Panaud

Jury présidé par Prof. Pierre Capy

Thesis director: Chris Bowler
CNRS UMR 8186
Département de Biologie
Ecole Normale Supérieure
46 rue d'Ulm, Paris, France

External supervisors: Vincent Colot
CNRS UMR 8186
Département de Biologie
Ecole Normale Supérieure
46 rue d'Ulm, Paris, France

David Moreira
CNRS UMR 8079
Unité d'Ecologie, Systématique et Evolution
Université Paris-Sud, bâtiment 360
91405 Orsay Cedex, France.

I would like to dedicate this work to my parents Chantal and Olivier, my sister Laure, and my little princess Diana for their love, comprehension, and support. I also think about my dear friend Bertrand who was the most brilliant person I ever met and was very interested in physics and biology but accidentally lost the ability to go to the University ten years ago. Bertrand, I know you would have had unlimited pioneering ideas about all the topics I will talk about.

Contents

Summary

The unicellular chlorophyll *c*-containing algal class Bacillariophyceae (diatoms) is among the most successful and diversified groups of photosynthetic eukaryotes, with possibly over 100,000 extant species (Round et al., 1990) widespread in all kinds of open water masses. The contribution of diatom photosynthesis to marine primary productivity has been estimated to be around 40% (Nelson et al., 1995; Raven and Waite, 2004). Diatoms have a peculiar genetic makeup in that they are likely to have emerged following a secondary endosymbiotic process between a photosynthetic eukaryote, most probably red algal-like, and a heterotrophic eukaryote (Falkowski et al., 2004). They are traditionally divided into two orders: the centric diatoms which are radially symmetrical and are thought to have arisen around 180 Million years ago (Mya), followed by the pennate diatoms around 90 Mya which are bilaterally symmetrical. The complete nuclear, mitochondrial, and plastid genome sequences of the centric diatom *Thalassiosira pseudonana* (32 Mb) and the pennate diatom *Phaeodactylum tricornutum* (27 Mb) have recently become available (Armbrust et al., 2004 and Bowler et al., 2008).

My PhD focalized on the study of various aspects of the regulation of gene expression in diatoms as well as on the study of genome evolution and dynamics in these species. Gene expression is regulated at the transcriptional, post-transcriptional, and epigenetic (or pre-transcriptional) levels. In the framework of my PhD, I performed an in silico search in diatom genomes for transcription factors (TFs), which are master control proteins involved in transcriptional regulation, in order to get a panorama of the TF complement in these species and to identify lineage-specific peculiarities. We found, for example, that Heat Shock Factors (HSFs) have been amplified dramatically during the course of diatom evolution. Analysis of

the abundance of TFs in different *P. tricornutum* and *T. pseudonana*-derived EST libraries enabled us to identify some specificities of their expression.

The evolution of eukaryotic genomes is impacted by the direct or secondary effects of transposable elements (TEs), which are mobile DNA sequences that inhabit the genomes of most organisms. In order to evaluate genome dynamics in diatoms, the search for transposable elements in diatoms enabled to establish that a specific class of TEs, the Copia-like retrotransposons, is the most abundant in diatom genomes and has been significantly amplified in the *P. tricornutum* genome with respect to *T. pseudonana*, constituting 5.8 and 1% of the respective genomes. Phylogenetic analysis enabled me to demonstrate that diatom genomes harbor two classes of diatom-specific Long Terminal Repeat retrotransposons (LTR-RTs), as well as another class which is widespread among eukaryotes. Analysis of their abundance in various *P. tricornutum*-derived EST libraries has shown that two of these elements are activated in response to stress such as nitrate starvation. This activation is accompanied by DNA hypomethylation and the analysis of insertion profiles of different *P. tricornutum* ecotypes from around the world as well as other clues suggest that TEs play an important role in the generation of variability in diatom genomes. My interest in TEs further led me to try to characterize them in the genome of the brown alga *Ectocarpus siliculosus*.

In silico searches for genes encoding proteins putatively capable of introducing or stabilizing epigenetic modifications such as histone modifiers and DNA methyltransferase has shown the presence of a large set of such proteins in the *P. tricornutum* genome, as well as their particularities. The presence of particular histone modifications in the *P. tricornutum* proteome has also been assessed by western blot experiments. In addition, I adapted a chromatin immunoprecipitation protocol for *P. tricornutum* which was used successfully to observe that histones within nucleosomes associated with TEs were marked with specific modifications.

Other experiments enabled to establish that most TEs are marked by DNA methylation in *P. tricornutum*. An experiment aiming to obtain genome-wide DNA methylation pattern in this species has been designed and launched and will allow to whether some genes are also methylated.

Finally, small RNAs (sRNAs) constitute an additional mode of regulation of gene expression, acting at the interface between transcriptional regulation, post-transcriptional regulation, and epigenetic phenomena. I have studied sRNA-based mechanisms in diatoms by performing an in silico search that showed that proteins putatively involved in the generation of sRNAs are only poorly conserved in diatoms. I was also able to determine experimentally a link between sRNA and DNA methylation in *P. tricornutum*.

Résumé en français

Les océans couvrent plus de 70% de la surface de la Terre (planète bleue) et la productivité primaire nette (PPN) marine est équivalente à celle terrestre. Alors qu'il ne représente que 1% de la biomasse totale d'organismes photosynthétiques de la planète, le phytoplancton est responsable d'environ 45% de la PPN globale. Le terme phytoplancton décrit un assemblage polyphylétique comprenant des eucaryotes et procaryotes photosynthétiques dérivant avec les courants. Dans les océans contemporains, les diatomées constituent un groupe d'eucaryotes unicellulaires autotrophes très abondant, responsable de 40% de la PPN marine. Les diatomées appartiennent à la lignée des straménopiles qui sont issus d'un évènement d'endosymbiose entre une algue rouge et un hôte hétérotrophe. Elles sont classifiées en deux groupes majeurs : les centriques qui son apparues il y a environ 200 millions d'années (Ma), et les pennées qui ont évolué il y a environ 90 Ma. Deux génomes de diatomées ont récemment été séquencés : celui de la diatomée centrique *Thalassiosira pseudonana* (32 Mb), et celui de la diatomée pennée *Phaeodactylum tricornutum* (27 Mb).

Mon sujet de doctorat s'est focalisé sur l'étude de différents aspects de la régulation de l'expression génique ainsi que sur la dynamique et l'évolution de ces génomes. L'expression des gènes est régulée à différents niveaux: trancriptionel, post-transcriptionel, et épigénétique. Dans le cadre de mon doctorat, une étude de la régulation transcriptionelle chez les diatomées a été effectuée et comprend l'identification et l'analyse in silico des facteurs de transcription (FT). Cela a permis par exemple d'établir qu'une classe spécifique de FT, les Heat Shock Factors, sont particulièrement abondants chez les diatomées par rapport aux autres eucaryotes. L'analyse de la représentation des FT identifiés dans différentes librairies d'EST élaborées à partir de cultures ayant subi divers stress a permis de détecter certaines spécificités d'expression.

L'évolution des génomes eucaryotes est largement impactée par les effets directs et secondaires des éléments transposables (ET) qui sont des éléments génétiques mobiles se trouvant dans le génome de la plupart des organismes. Dans le but d'étudier la dynamique des génomes de diatomées, la recherche de différents types d'ET a permis d'établir qu'une certaine classe, les rétrotransposons de type Copia, est la plus abondante dans ces génomes et constitue un part significativement plus importante du génome de *P. tricornutum* (5,8%) par rapport à *T. pseudonana* (1%). D'autre part, des analyses phylogénitiques ont montré que les rétrotransposons de type copia forment deux classes distinctes et éloignées de la lignée Copia. L'analyse de leurs niveaux d'expression a montré que la transcription de deux éléments s'active en réponse à des stress spécifiques comme la limitation en nitrate dans le milieu de culture. Cette activation est accompagnée par un hypométhylation de l'ADN et l'analyse de profils d'insertions chez différents écotypes de *P. tricornutum* ainsi que l'étude d'autres phénomènes suggèrent que les rétrotransposons de type Copia ont joué un rôle important dans l'évolution des diatomées. Mon grand intérêt pour les ET m'a ensuite amené à chercher à les caractériser dans d'autres génomes récemment séquencés tels celui de l'algue brune *Ectocarpus siliculosus.*

La recherche in silico de différents gènes codant des protéines capables d'introduire ou de stabiliser des états épigénétiques telle que la modification des histones et la méthylation de l'ADN a montré leur présence chez *P. tricornutum* ainsi que leurs particularités. La présence de certaines modifications d'histones spécifiques d'une conformation compacte ou ouverte de la chromatine dans le proteome de *P. tricornutum* a été montrée. De plus, la mise au point de la technique d'immunoprécipitation de la chromatine chez *P. tricornutum* a permis d'établir que les nucléosomes enrobés d'éléments transposables étaient marqués par des modifications spécifiques.

D’autres expériences ont permis d’établir que l’ADN de différents types d’éléments transposables est marqué par la méthylation de cytosines chez *P. tricornutum*. Une expérience permettant l’analyse du profil de méthylation à l’échelle de génome en utilisant une puce à ADN a été lancée et permettra de découvrir si certains gènes portent aussi des traces de méthylation.

Enfin, les ARN interférents constituent un troisième mode de régulation de l’expression se situant à l’interface de la régulation transcriptionelle, post-transcriptionelle et épigénétique. Les mécanismes d’interférences chez les diatomées ont été étudié par la recherche in silico d’enzymes clés impliquées dans ce processus ainsi qu’en établissant expérimentalement un lien direct avec la méthylation de l’ADN.

Aims of the thesis

Diatoms are chlorophyll c-containing algae that have divergent evolutionary histories compared with green and red algae, and higher plants. Whole-genome analyses of the diatoms *P. tricornutum* and *T. pseudonana* gene repertoires indicate that they have combined genes and pathways from a range of different sources. Beyond DNA, crucial information complementary to genomic data relies on deciphering the *modo*, *tempo*, and networks of gene expression. Towards a better comprehension of the regulation of gene expression in *P. tricornutum* and *T. pseudonana*, various cDNA libraries have generated from cells grown under different conditions of stress such as iron or silica starvation. An additional key step towards a better understanding of diatom biology is the identification and study of master regulatory genes and mechanisms that govern gene expression. At the beginning of my PhD, these questions had never been addressed in diatoms or in any stramenopile. Therefore, I focused my research on the study of different mechanisms regulating gene expression at the transcriptional, post-transcriptional, and epigenetic levels. Another important and little documented issue in diatoms regards genome dynamics, and the generation of genetic variability. Among several factors and processes, transposable elements (TEs) are mobile genetic sequences that play a key role in shaping the genomes of most eukaryotes. Therefore, I dedicated much of my PhD to the identification and study of such elements in diatom genomes.

Aknowledgements

My PhD has been a very fascinating and exciting epic because of the different topics addressed and findings made. But the strongest feeling that comes at the end is that I lived an extraordinarily rich human story. I received so many inputs and support from the different people I met during this period that I would like to thank some of those who participated to the real good memories of these last years. I will carry this love in my back pack for the rest of my life, together will all the ideas, discussions and expertise that they shared with me.

I first of all want to thank Chris Bowler. Chris is a very busy boss. He's also quite strict and pragmatic as a manager. Although we were maybe not prepared to work together, I consider my PhD a full success and this is much thanks to him, to the way we interacted together, and to the opportunities he offered to me. He was always able to find the time to listen to me and advise me. Although always up to date and critic, he left me fairly autonomous and always accepted to consider my ideas and proposals in terms of experiments and collaborations. This relative freedom has also been the chance for me to face many training situations that complement the scientific knowledge I acquired, and make me stronger for the future and aware of scientific life drawbacks. Chris is originally a plant biologist and I started working with him on plant photomorphogenesis. I've seen him growing into a diatomist and then broaden his fields of interest and go towards microbial oceanography. My big chance has been to follow his steps and learn little to little, from tomatoes to diatoms to microbiology, evolution, ecology, and genomics. So thank you Chris for all these inputs you bring to the lab and for your successful steps in these fields that I would probably never have explored otherwise.

The second person to whom I really want to dedicate a few lines is Kamel Jabbari. Kamel was a postdoc on comparative genomics when I arrived in Chris' lab and left about two years ago. I've been discussing more about science with Kamel in the last four years than

with anyone else. Kamel and I have been arguing for hours about the evolutionary aspects of my thesis and much more, and he has been an excellent external referee all along my PhD, as well as a friend. Thank you, Kamel, for sharing with me your knowledge and your experience of life.

I am really grateful to Angela Falciatore who is the head of a diatom lab we closely collaborate with in Naples; Italy. Angela is a brilliant diatomist and one of the most enthusiastic scientists I ever met. Even when she looks bored, she's making fun out of it. Angela was co-supervising with Chris my stage on photoperception in diatoms during the second year of master and she welcomed me at that time in her lab with great benevolence and encouraged me all along my PhD. It has also been of great interest to collaborate with her and her team.

I also want to thank many ex- PhD students and postdocs from Chris' lab that were there when I arrived and slowly left to all the parts of the world. Andy Allen, Assaf Vardi, Anton Montsant, Uma Maheswari, and Marc Heijde. Together with Kamel, I shared much more than science with these exceptional persons, such as memorable petanque games and parties. It was also a great honor to collaborate with all of them, especially Marc Heijde who dedicated much of his time to introduce me to all the molecular biology material and methods in the lab as well as all the drawers and cupboards.

Finally, I really want to thank Alessandra de Martino for her friendly support and her precious advices, Agnès Meichenin for her kindness and all her help with cultivating and transforming diatoms, Joseph-Anne Timons for his daily good mood, Gerald Zabulon for his "tip of the day" and his famous punch, Jean-Claude Thomas for his wise speeches and his love and dedication to science, Sophie de Peindray as a model of organization, Chantal Guidi-Rontani for her inexhaustible and interesting arguments and opinions, and last but not least, I'd like to thank Edda Rayko for the huge support and as an exemplary hard worker.

Chapter I:

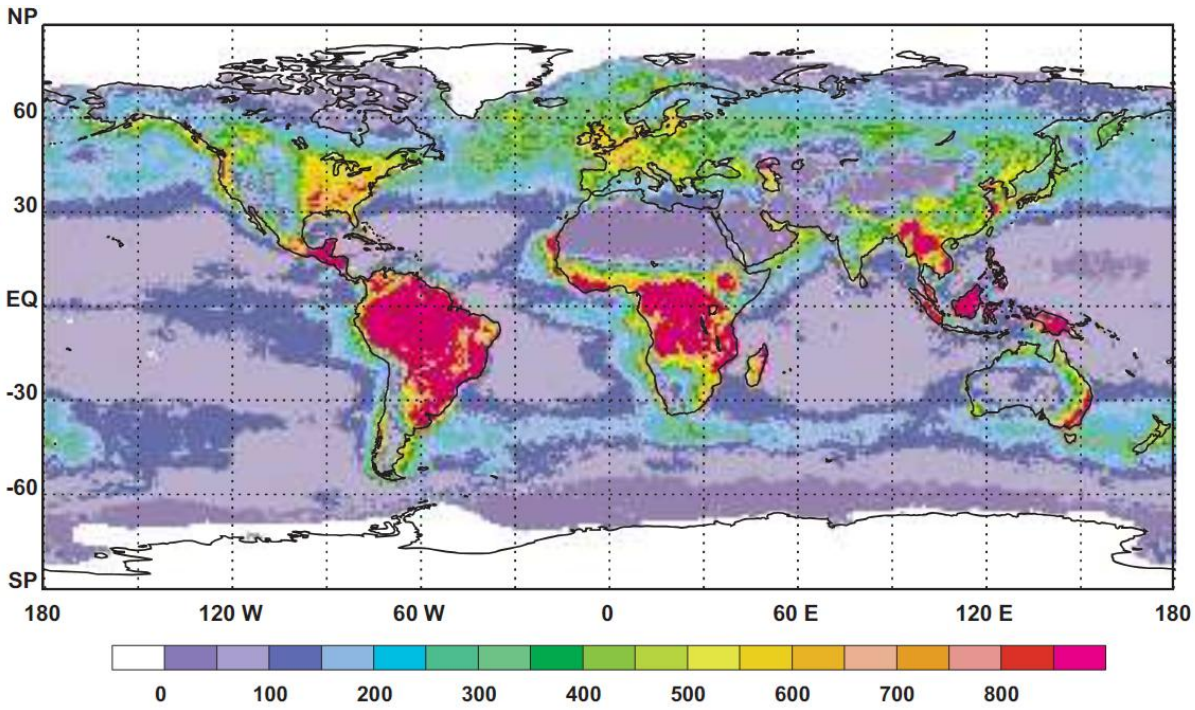
Introduction

1.1 Diatoms, marine genomics, and the marine ecosystem

1.1.1 First glance :

The contemporary oceans cover about 70% of the Earth's surface. Each year, around 50 billion tons of carbon in the form of CO2 are fixed into organic material by the microorganisms populating the upper sunlit portion of the ocean, and each year a similar amount of organic carbon is transferred into marine ecosystems by sinking and grazing (Figure 1.1). Thus, although accounting for only about 1% of the Earth's photosynthetic biomass these organisms, commonly referred to as "phytoplankton", are responsible for about 45% of our planet's annual net primary productivity (NPP) (Field, et al.,1998), while the contribution from macroalgae is estimated to be around 1% (Smith, 1981). The term "phytoplankton" describes a diverse, polyphyletic group of mostly single-celled photosynthetic organisms that drift with the currents in marine and fresh waters including cyanobacteria and eukaryotes. In contemporary marine ecosystems, diatoms constitute one of the most successful groups of phytoplankton, responsible for about 40% of marine net primary productivity. In corollary, diatoms are important producers of the oxygen we breathe and form the base of the food chain in marine ecosystems upon which the higher trophic levels depend.

Figure 1.1: Global annual NPP (in grams of C per square meter per year) for the biosphere (from Field et al., 1998).



1.1.2 Diatom biology

Diatoms (class Bacillariophyceae**)** are a group of autotrophic single-celled eukaryotic algae found in all kinds of humid environments and open water masses. One of the most characteristic features of diatoms is their ability to generate a hard outer silica-based cell wall called the frustule which consists of two asymmetrical halves assembled like the two overlapping halves of a Petri dish (Figure 1.2), hence the Greek etymology of "diatom": *dia* "through" and *temein* "to cut"; i.e., "cut in half". The larger of the two halves is called the epitheca and the inner one is called the hypotheca. The frustule has been proposed to constitute a physically strong and chemically inert protective covering because silica cannot be attacked enzymatically.

When a diatom divides to produce two daughter cells, each cell keeps one of the two halves and grows a smaller half within it. Because the frustule cannot grow once it has been laid down, the average size of diatom cells in a population decreases after each division cycle (Figure 1.3). Once such cells reach a certain minimum size, regeneration of the original size typically occurs via sexual reproduction in which male and female gametes combine to produce a diploid auxospore. This expands in size to generate a cell larger than either parent, which then returns to size-diminishing mitotic divisions. Another consequence of size reduction, particularly for pennate diatoms, is that smaller cells may be mistaken for other species because valve outline and proportions may differ from that of larger cells (Figure 1.4).

Frustules are often finely ornamented with pores, processes, spines, and other distinguishing features, and resemble an elegant pattern of lace drawn at the nanometer scale (Figure 1.5). They come in many fascinating and beautiful shapes, whose intricate architecture is faithfully reproduced from generation to generation, implicating a strict genetic control of the process. Actually, the silicon-based pattern design is used by scientists to distinguish one diatom species from another and forms the basis of diatom classification

(Round et al., 1990). This astonishing level of precision has rendered diatoms popular with microscopists as early as the nineteenth century because they were used as guides to improve optical microscope resolution. Nowadays, understanding the biochemical process of frustule formation in diatoms is appealing for nanotechnologists because these architectures far exceed the capabilities of human engineering and are accomplished under mild physiological conditions.

Figure 1.2 : Schematic overview of the general structural features of a pennate diatom (from Falciatore and Bowler, 2002)

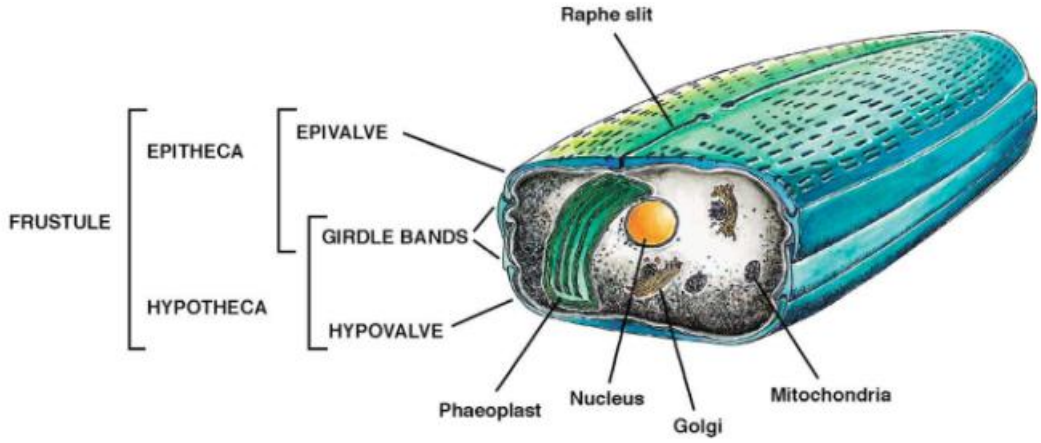


Figure 1.3: Schematic overview of mitotic cell cycle and hypovalve formation in a pennate diatom. Following mitosis and cytokinesis, a specialized vesicle known as the silica deposition vesicle (SDV) forms between the nucleus and the plasma membrane, at the position where the new hypovalve will be generated. The SDV elongates into a tube and spreads out perpendicularly to eventually form a huge vesicle along one side of the cell. A new valve is formed within the SDV by the transport of silica, proteins, and polysaccharides into it, and once complete, it is exocytosed from the cell. The two daughter cells can then separate and grow unidirectionally along the cell division axis by the biogenesis of gridle bands, which are also formed within the SDVs. (from Falciatore and Bowler, 2002)

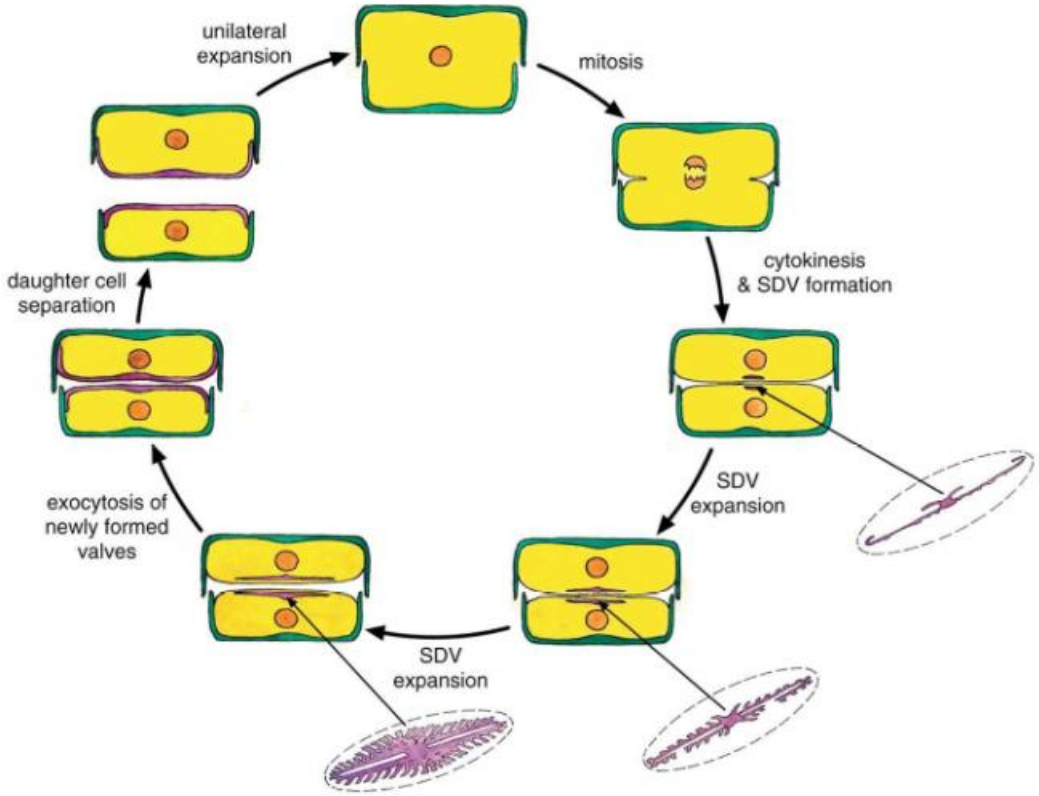
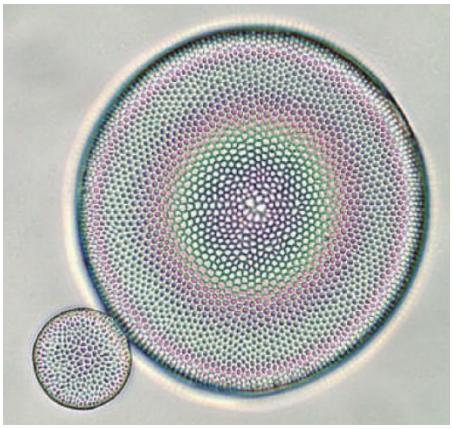


Figure 1.4: (A) Example of cell size reduction in *Coscinodiscus* (from Armbrust, 2009). (B) Example of change in valve outline and shape with size reduction in *Navicula reinhardtii* (from M.G. Kelly, H. Bennion, E.J. Cox, B. Goldsmith, J. Jamieson, S. Juggins D.G. Mann & R.J. Telford. 2005. Common freshwater diatoms of Britain and Ireland: an interactive key. Environment Agency, Bristol, UK.)

(A)



(B)

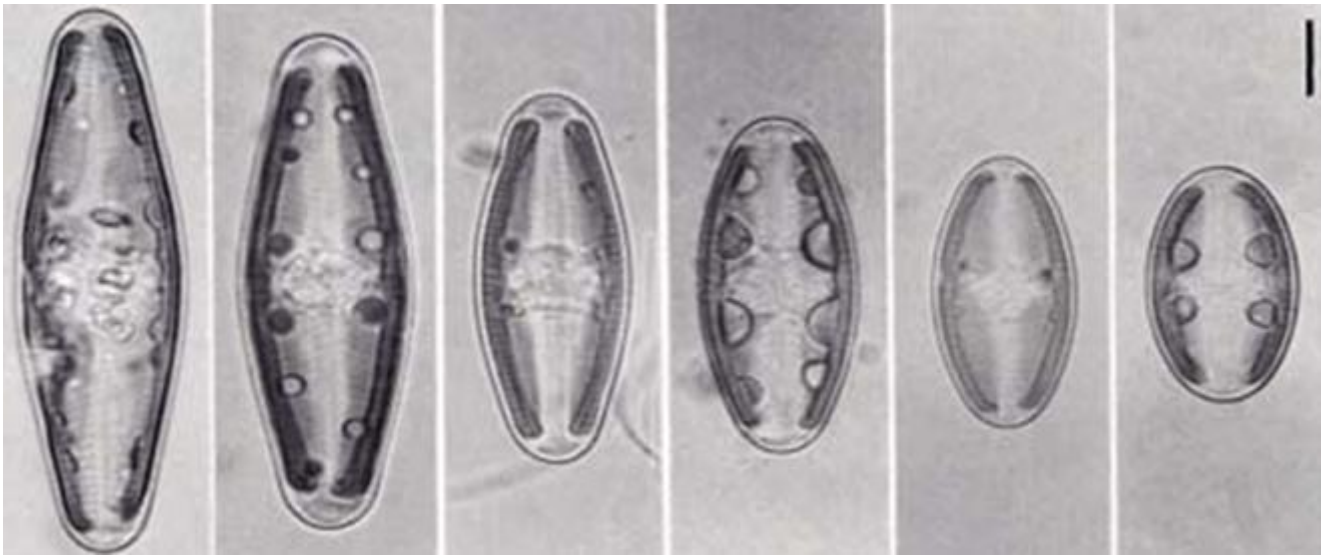
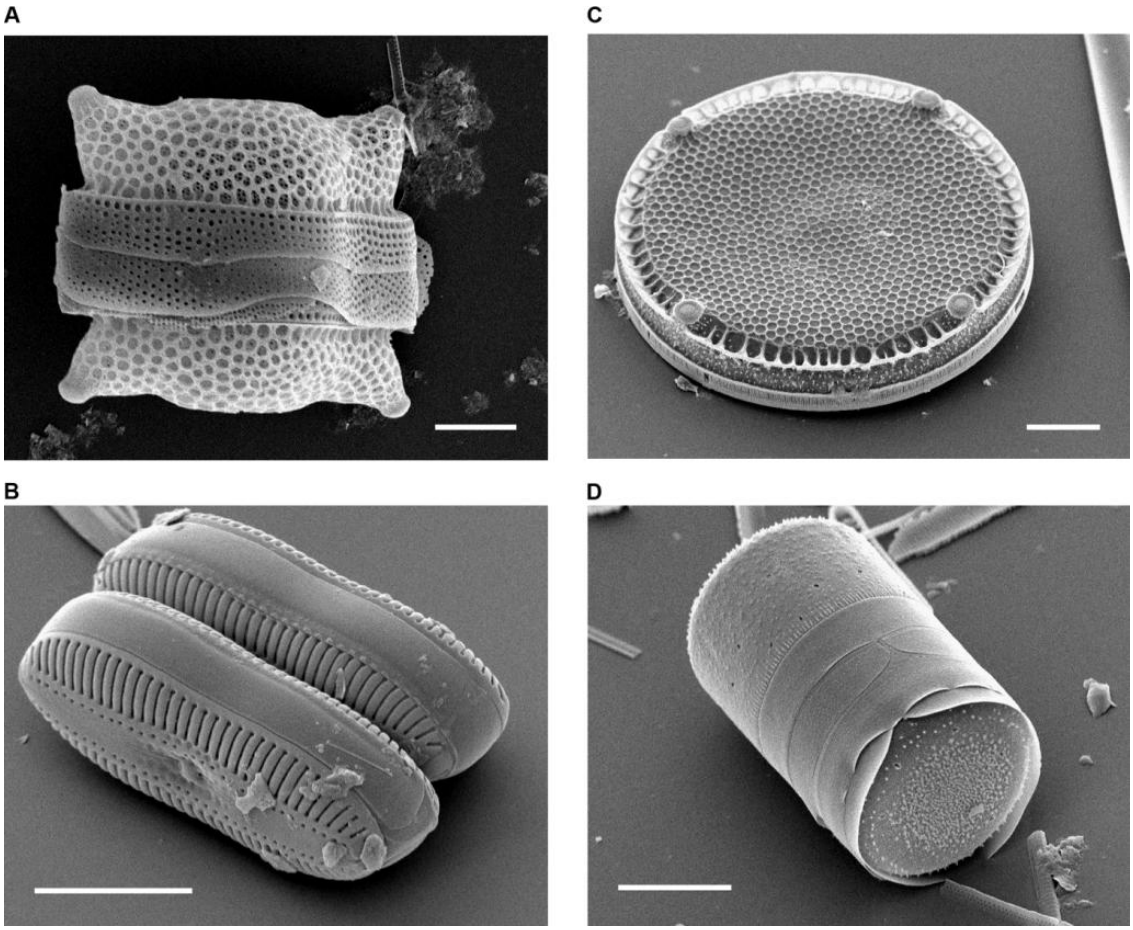


Figure 1.5: Scanning electron micrographs of diatoms (from Bradbury, 2004). (A) *Biddulphia reticulata*. The whole shell or frustule of a centric diatom showing valves and girdle bands (size bar = 10 micrometres). (B) *Diploneis* sp. This picture shows two whole pennate diatom frustules in which raphes or slits, valves, and girdle bands can be seen (size bar = 10 micrometres). (C) *Eupodiscus radiatus*. View of a single valve of a centric diatom (size bar = 20 micrometres) (D*) Melosira varians*. The frustule of a centric diatom, showing both valves and some girdle bands (size bar = 10 micrometres)



1.1.3 Classification and Habitats

The diatom group is among the most diversified group of photosynthetic eukaryotes with more than 200 living genera and approximately 100,000 extant species (Round et al. 1990). Diatoms are traditionally divided into two orders according to cell symmetry: the centrics which are radially symmetrical and the pennates which are bilaterally symmetrical. Further, some pennate diatoms known as raphid diatoms have a slit (raphe) in the cell wall for movement via the secretion of polysaccharides; the araphid pennates lack this slit and are non-motile. Also, an additional subgrouping of centrics is now recognized, the radial centrics and bi/multi polar centrics. The diatom taxonomy thus distinguishes between the four classes (Kooistra, 2007) (Figure 1.6). It is generally agreed that the pennate diatoms evolved from the centric forms and that the raphid pennates evolved from the araphid pennates due to their order of appearance in the fossil record. Indeed, the siliceous frustules preserve remarkably well, and consequently diatoms have a detailed fossil record.

The first reliable diatom fossil is of a centric from the early Jurassic about 185 Million years ago (Mya) (Sims et al., 2006). However, it is believed that diatoms originated further back in time than their fossil record indicates because of the incompleteness of sampling. The lack of earlier fossil records may also be because the first diatoms lacked robust silica frustules. After the early Jurassic period, diatoms are scarce until the Lower Cretaceous (145-110 Mya) where highly diversified centric diatoms occur in many deposits. The first pennate diatoms, the araphids, appear in fossil records from the Late Cretaceous (87–65 Mya) and the raphid pennate diatoms appear at about 50 Mya (Kooistra and Medlin, 1996). The order in which the four diatom groups appeared is in accordance with that inferred with molecular phylogenies, although the molecular clocks place the origin of diatoms at 250 Mya (Sorhannus et al., 2007).

Nowadays, although being the youngest, the pinnate diatoms are by far the most diversified. They are major components of both benthic and pelagic habitats, having either a planktonic existence or being associated with surfaces. On reefs, pennate diatoms can dominate the communities of protists on surfaces such as macroalgae and rocks. The appearance of the raphe was probably a determining feature that permitted subsequent rapid diversification of pennate diatoms into a range of habitats and although it likely evolved to glide on surfaces in benthic diatoms, many raphid pennates now colonize planktonic environments.

Planktonic diatoms live in the photic zone of the ocean, between the surface layer and the nutrient-rich deep chlorophyll maximum (DCM), and generally prefer well-mixed nutrient-rich waters. Species composition of the surface and DCM layer communities are typically quite distinct (Venrick, 1998). When nutrients become scarce diatoms can sink to great depths where it is thought that they lie dormant until conditions become favorable again for growth. Diatoms have also colonized sea-ice ecosystems where other autotrophs are especially scarce (Figure 1.7) and thus ice diatoms constitute an important component of polar food webs.

Figure 1.6: Phylogenetic relationships among Bacillariophyceae (adapted from a figure courtesy of Dr. Wiebe Kooistra from the Stazione Zoologicale Anton Dorn in Naples, Italy). Phylogenies inferred from SSU rRNA-gene regions of diatoms and their stramenopile relatives reveal that radial centrics are the most ancestral and probably paraphyletic. They gave rise to multipolar centrics, which are also paraphyletic. One of its lineages gave rise to the pennates. Within the latter, araphid pennates are paraphyletic and raphid pennates monophyletic.

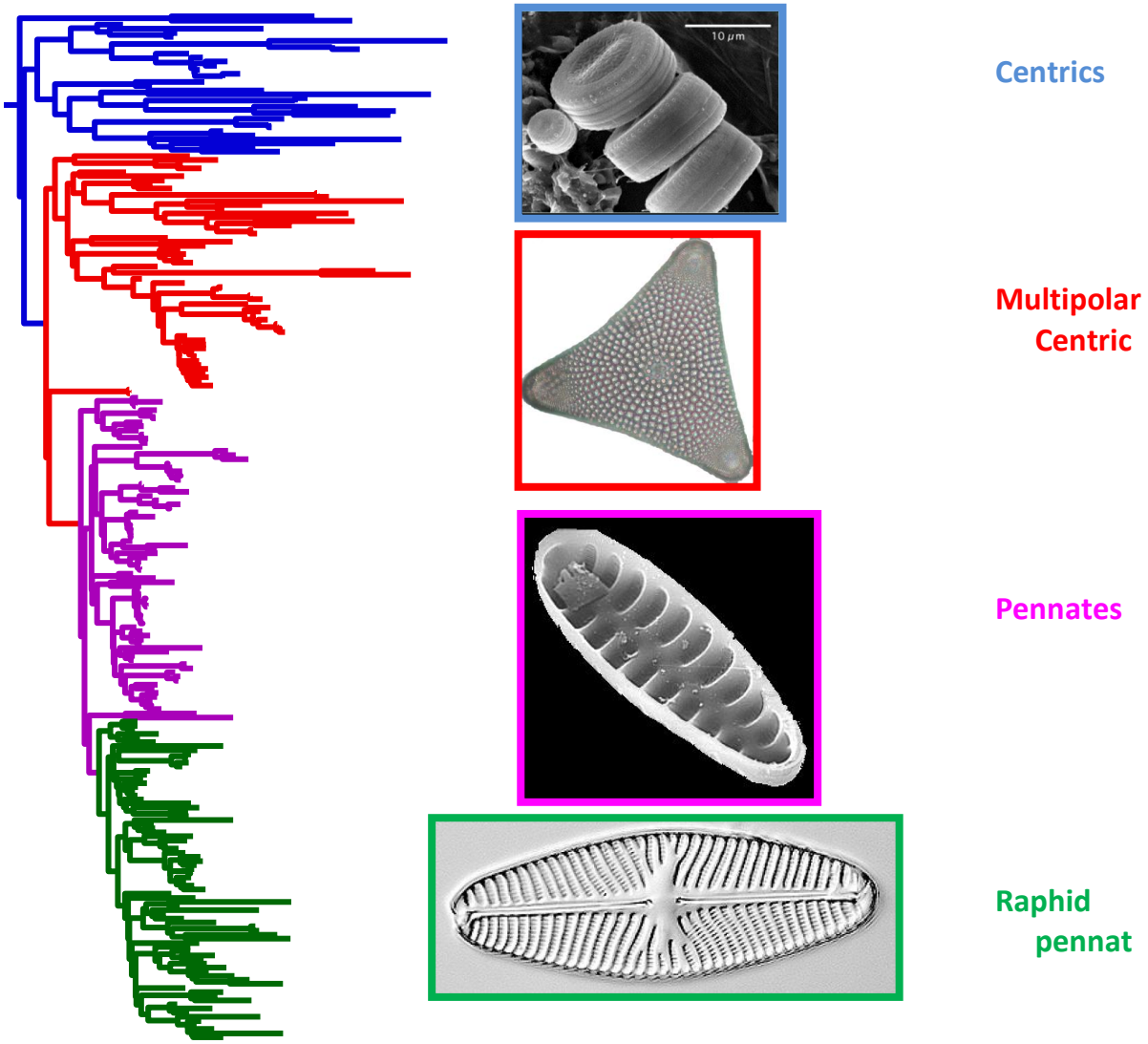
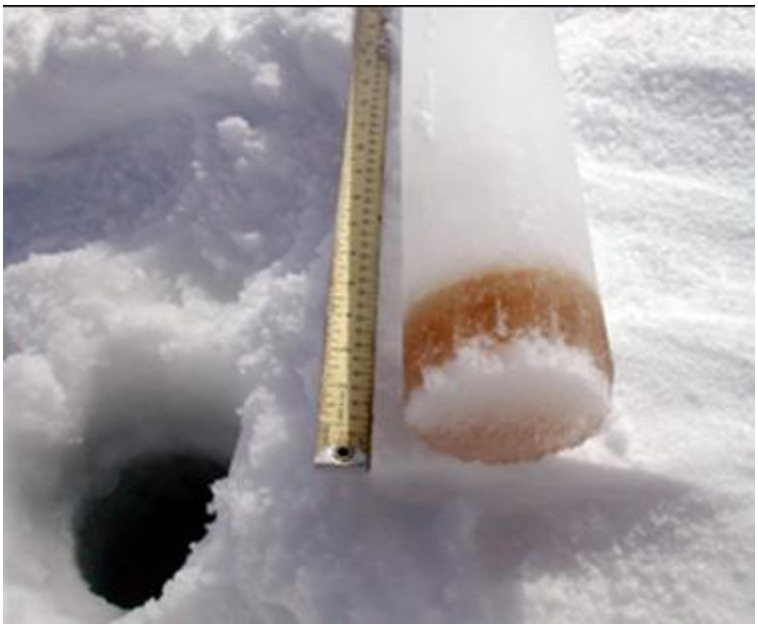


Figure 1.7: Polar diatom community. The orange ice is filled with diatoms. The diatoms live in the pore spaces between ice crystals and in salty brine water that is within the ice. Some zooplankton in the ocean water are able to eat diatoms from the bottom of the ice. (from The National Oceanic and Atmospheric Administration (NOAA)



1.1.4 Role of diatoms in biogeochemical cycles

Until the Mesozoic, the dominant phytoplankton were organic-walled organisms (cyanobacteria, green algae, acritarchs, and dinoflagellates). Eukaryotic phytoplankton with mineralized skeletons became dominant during the Mesozoic and account for most contemporary marine primary productivity. These include the coccolithophorids (with calcium carbonate cell walls), silicoflagellates and diatoms (with silica-based cell walls). As a result of gravitational settling, the accelerated sinking rates of such heavy cells contributed enormously to the biogeochemical pump, decreasing CO_2 and increasing O_2 in the atmosphere, and resulting in carbon sequestration in sediments, creating many of today's oil and gas reserves and a significant amount of oxygen. The downward flux of dead mineralized phytoplankton is also essential to transfer organic material and ions to the ecosystem living beneath the photic zone. The abundance of diatoms in modern marine ecosystems is such that they constitute a key component of the biological carbon pump that transports carbon to the ocean interior, largely contributing to the long term sequestration of atmospheric CO_2 and supporting much of the marine food chain.

Because they use silica to synthesize their frustules, diatoms are also the most significant consumers of silicic acid dissolved in the ocean. Every atom of silicon entering the ocean has been estimated to be incorporated into 40 successive diatom frustules before sinking to the sea floor (Tréguer et al., 1995). Deposits of diatoms formed over geological time as ancient diatoms died and settled to the bottom have built enormous layers of diatom skeletons known as diatomite, or diatomaceous earth. These deposits are mined for use in toothpastes, paints, filtering agents, and abrasives, as well as explosives.

Nitrogen is an essential nutrient for living organisms in being a constituent of nucleic and amino acids. Most marine eukaryotes and cyanobacteria are able to incorporate inorganic (NO3-, NO2-, NH4+) and organic (urea, amino acids) forms of nitrogen. Diatoms tend to

dominate other microalgae in nitrate-rich waters (e.g., upwelling environments), but they are rare in nitrogen-poor open ocean ecosystems where recycled nitrogen (e.g., NH4+) and nitrogen fixation drive production. Although eukaryotic phytoplankton do not fix nitrogen gas, some diatom species can elaborate symbiotic relations with nitrogen-fixing cyanobacteria (Foster and Zehr, 2006).

The oxidizing conditions of the modern oceans result in very low concentrations of dissolved iron and the cellular demand from phytoplankton is often in excess of iron availability. This paradox is though to reflect the fact that cellular machineries depending on iron evolved in the iron-replete Proterozoic (2.5 to 0.5 Mya) reducing oceans. Indeed, the oxygen released over time since the appearance of photosynthesis in cyanobacteria gradually filled chemical sinks of unoxidized iron before oxygen levels began to increase in the atmosphere. Nowadays, iron is supplied to phytoplankton by aeolian transport of dust (Cassar et al., 2007) and possibly by the upwelling of deep water from hydrothermal vents. It is now well accepted that iron depletion is a key limitation of phytoplankton growth and the cause of so-called "photosynthesis deserts" (e.g., polar Southern Ocean) which are defined as high nutrient low chlorophyll (HNLC) regions. Conversely, large-scale open ocean iron fertilization experiments have been shown to trigger phytoplankton blooms, especially of pennate diatoms (Boyd et al., 2000).

1.1.5 Evolutionary history

The vast majority of characterized eukaryotes can now be assigned to one of five or six supergroups (Keeling et al., 2005) which comprise the Opisthokonta and Amoeboza (often united in the Unikonts), Plantae, Excavata, Chromalveolata, and Rhizaria (Figure 1.8). These supergroups have been established on the basis of molecular and ultrastructural data. Molecular analysis involves mainly the analysis of distances in concatenated multi-gene

datasets and the conservation of gene fusions, while ultrastructural data involves characterization of different visible characters such as the shape of the mitochondrial inner membrane (cristae), the presence of a plastid, or the number of flagella.

The Unikonts, which includes animal and fungal kingdoms and related forms, have a single basal flagellum on reproductive cells and flat mitochondrial cristae (most eukaryotes have tubular ones). The unikonts have a triple-gene fusion encoding enzymes for synthesis of the pyrimidine nucleotides (carbamoyl phosphate synthase, dihydroorotase, aspartate carbamoyltransferase) that is lacking in bacteria and bikonts. This must have involved a double fusion, a rare pair of events, which provides the first really compelling support for Amoebozoa being sisters to opisthokonts rather than to bikonts (Stechmann et al., 2003).

Excavata is a major assemblage of heterotrophic single-celled organisms which contains a variety of free-living and symbiotic forms. Many excavates such as diplomonads and parabasalids (e.g., *Giardia* and *Trichomonas*) lack classical mitochondria, although most retain a mitochondrial organelle in greatly modified form. Others have mitochondria with tubular, discoidal, or in some cases, laminar cristae. Most excavates have two, four, or more flagella.

The Rhizaria unites a heterogeneous group of flagellates and amoebae including cercomonads, foraminifera, and former members of the polyphyletic radiolaria (Cavalier-Smith, 2002). Cercozoa and foraminifera appear to share a unique insertion in their ubiquitin gene that consolidates the Rhizaria supergroup (Archibald et al., 2003).

The group Plantae arose through endosymbiosis whereby a non-photosynthetic single-celled eukaryote (host) engulfed and retained a free-living photosynthetic cyanobacterium to form a primary symbiotic oxygenic eukaryote more than 1,500 Mya (Figure 1.9A). Over time, the prokaryote was reduced to a double membrane-bound plastid and was vertically transmitted to subsequent generations. Part of this process involved the transfer of hundreds

of genes from the cyanobacterium/plastid to the eukaryotic host nucleus via a process known as endosymbiotic gene transfer (EGT) concomitant with reduction of the enslaved cyanobacterium genome to become the plastid genome. Recent analysis also suggests the presence of some chlamydial (obligate intracellular bacteria) genes in both plants and red algae, but not in cyanobacteria, and therefore that a chlamydial endosymbiont may also have imprinted during the early stages of the primary endosymbiosis (Becker et al., 2008).

The photosynthetic Plantae ancestor diversified into three lineages: green algae which include land plants and have chlorophyll a and b pigments, rhodophytes (red algae), and glaucophytes which both have chlorophyll a and phycobilin pigments (Reyes-Prieto et al., 2007). The plastids of glaucophytes retain a cyanobacterial-type peptidoglycan layer that differentiates them from red algae. In the ocean, the dominant unicellular algae derived from a primary endosymbiosis are prasinophytes, which belong to the green lineage.

Soon after the split of rhodophytes and green algae, it is hypothesized that a member of the rhodophyte lineage was engulfed by an unknown heterotroph secondary host giving rise to the pigmented ancestor of the Chromalveolata (Cavalier-Smith, 1999) about 1,300 Mya (Yoon et al. 2004) (Figure 1.9B). This supergroup of secondary endosymbionts was proposed as a parsimonious explanation for the presence of plastids of red algal origin in photosynthetic members of both the Alveolata and Chromista. The Alveolata group is well supported and includes ciliates, dinoflagellates, and apicomplexa. Chromista unifies diverse microbial lineages with red algal plastids (and their non-photosynthetic descendents) including cryptophytes, haptophytes, and heterokonts (which include diatoms).

Over evolutionary time, extant Chromalveolata have lost most of the rhodophyte symbiont structures but the process of secondary endosymbiosis has left its evolutionary signature with the plastid and its unique membrane topography. Secondary plastids are often surrounded by four membranes (in haptophytes, heterokonts, and cryptophytes). The two

additional membranes are thought to correspond to the plasma membrane of the engulfed alga and the phagosomal membrane of the host cell. The endosymbiotic acquisition of a eukaryotic cell is represented most clearly in the cryptophytes where the remnant nucleus of the red algal symbiont (the nucleomorph) is present between the inner two and outer two plastid membranes. The retention of the plastid can be attributed to the advantagous autotrophic lifestyle that it confers (many plastid-less cells living today engulf algae for this purpose in a process known as kleptoplastidy).

Besides plastid membrane topology, other evidence also supports the chromalveolate hypothesis which posits that the alveolate group and the less well resolved chromist group acquired secondary red plastids by a single common endosymbiosis. Notably, all photosynthetic chromalveolates contain chlorophyll c, which is absent from other algae (Cavalier-Smith, 1999). Second, two plastid-targeted proteins (glyceraldehyde-3-phosphate dehydrogenase and fructose-1,6-bisphosphate aldolase) have unusual but common features that are unique to chromalveolates (Harper and Keeling, 2003; Patron et al., 2004). However, the chromalveolate hypothesis is hotly debated and may not appear to be monophyletic as originally proposed. Specifically, the inclusion of haptophytes and cryptophytes in Chromalveolata is thus far only weakly supported by phylogenetic analyses using host nucleus-encoded genes (Harper et al., 2005). Further, recent phylogenomic data generated using expressed sequence tags (ESTs) from two species belonging to the supergroup Rhizaria suggest a very robust relationship between Rhizaria and two main clades of the chromalveolate supergroup: stramenopiles and alveolates; while cryptophytes and haptophytes cluster together in a little-supported clade (Burki et al. 2007). The name SAR (Stramenopile, Alveolate, Rhizaria) was proposed to accommodate this new super assemblage of eukaryotes.

At least two scenarios are conceivable to explain the evolution of chlorophyll-c containing plastids. First, a single engulfment of red algae might have occurred at a very early stage of chromalveolate evolution and the resulting plastid was secondarily lost in certain lineages, such as ciliates and Rhizaria. Second, it is possible that stramenopiles, alveolates, or haptophytes and cryptophytes have acquired their secondary plastid through an independent endosymbiosis event. Phylogenomic data incorporating EST data from cryptophytes and haptophytes also support the monophyly of haptophytes and cryptophytes and the association of Rhizaria with alveolates and stramenopiles (Hackett et al., 2007).

In separate, more recent endosymbioses, green algae were independently engulfed by the common ancestor of the chlorarachniophyte amoebae (Rhizaria) and of the euglenids (Excavata), giving rise to two distinct lines of green secondary plastids (Rogers et al., 2007). It has also been postulated that the ancestral dinoflagellate acquired its plastid from a haptophyte through a tertiary endosymbiosis event with plastid replacement (Yoon et al., 2002). Consequently, oxygenic photosynthesis that originated in cyanobacteria about 2,800 Mya has been transferred to a wide diversity of eukaryotes through successive and mixed endosymbiosis events over evolutionary times such that contemporary eukaryotic phytoplankton includes members of all eukaryotic supergroups except the Unikonts (Figure 8).

Figure 1.8: Phylogenetic relationships among the eukaryotic supergroups (adapted from Baldauf, 2003). Dashed parables cover supergroup-level classification. Blue dots indicate phytoplanktonic lineages.

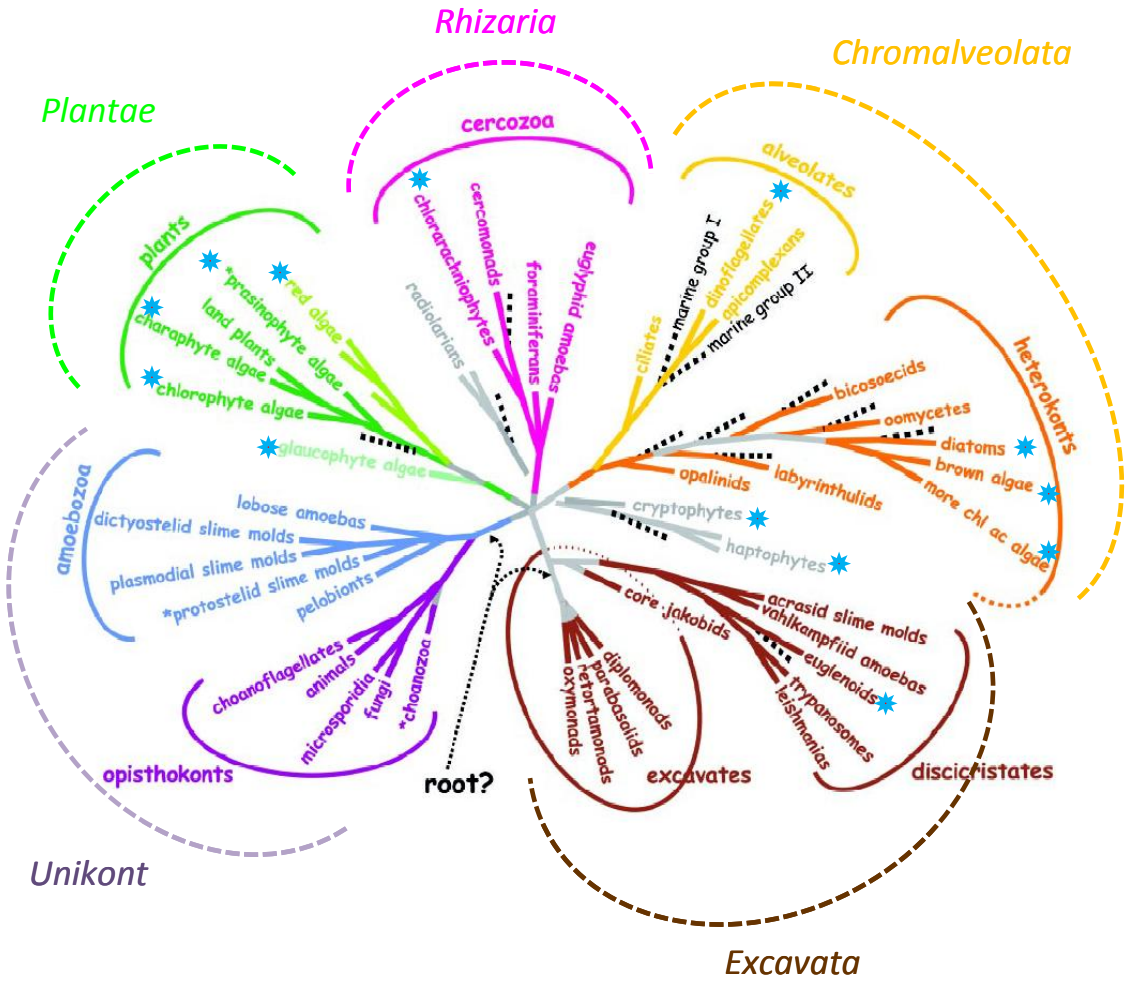
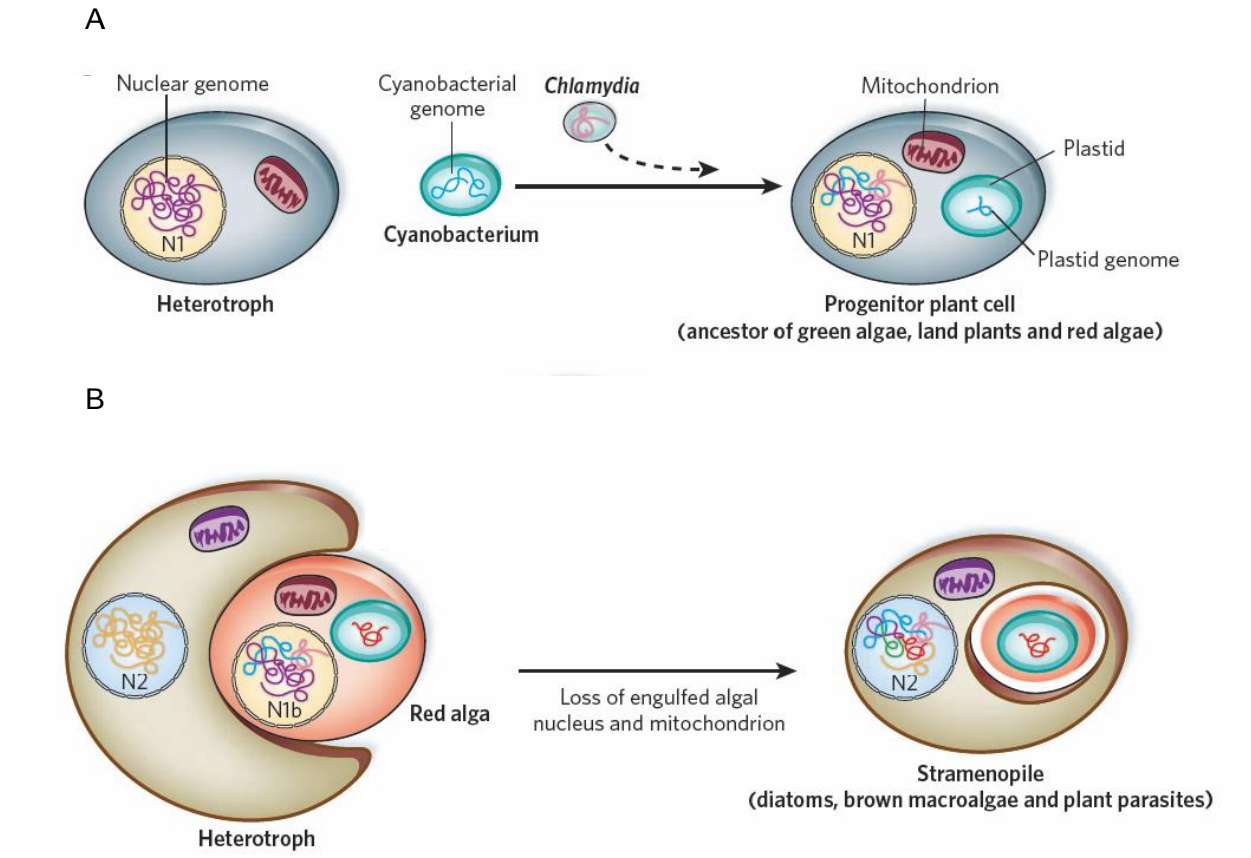


Figure 1.9: Representation of the origin of diatom plastids through sequential primary (A) and secondary (B) endosymbioses (adapted from Armbrust, 2009). (A) During primary endosymbiosis, a large proportion of the engulfed cyanobacterial genome is transferred to the host nucleus (N1), with few of the original genes retained within the plastid genome. The potential for invasion of the host by a chlamydial parasite is indicated with a dashed arrow, and the ensuing transfer of chlamydial genes to the host nucleus is indicated in pink. The progenitor plant cell subsequently diverged into red and green algae and land plants. (B) During secondary endosymbiosis, a different heterotroph engulfs a eukaryotic red alga. The algal mitochondrion and nucleus are lost, and crucial algal nuclear and plastid genes (indicated in blue, purple and pink) are transferred to the heterotrophic host nucleus, N2.



1.1.6 Diatom and (algal) genomics

The key role of eukaryotic phytoplankton in diverse biogeochemical cycles and the importance of understanding their evolutionary origins has driven the need to acquire genomic information. The emerging field of marine algal genomics began with the publication of the 551-kilobase remnant nucleomorph genome of the cryptophyte *Guillardia theta* (Douglas et al., 2001). Since then, several whole genome sequences from green algae, red algae, haptophytes, and stramenopiles have become available as well as the sequence of the nucleomorph from a chlorarachniophyte (Rhizaria) *Bigelowiella natans* (Table 1.1). The genomes of several non aquatic chromalveolates have also been published such as from the plant pathogens *Phytophthora sojae* and *P. ramorum* (oomycetes), the ciliates *Tetrahymena thermophila* and *Paramecium tetraurelia*, and the human parasite *Plasmodium falciparum* (apicomplexa). These sequences have revealed a wealth of information about biology and evolutionary origins of these organisms.

Two diatom genomes have been published: from the centric *Thalassiosira pseudonana* (32.4 Mb) and the raphid pennate *Phaeodactylum tricornutum* (27.4 Mb) (Armbrust et al., 2004; Bowler et al., 2008) (Figure 1.10). The *P. tricornutum* genome includes fewer genes (10,402 opposed to 11,776) than *T. pseudonana* and no major synteny (gene order) could be detected between the two genomes. *T. pseudonana* genes show an average of ~1.52 introns per gene as opposed to 0.79 in *P. tricornutum*, suggesting recent widespread intron gain in the centric diatom. Also, the molecular divergence between centrics and pennates is about halfway between the *Homo sapiens*/*Takifugu rubripes* and the *H. sapiens*/*Ciona intestenalis* divergence, indicating rapid evolutionary rates within the Bacillariophyceae compared to other eukaryotic groups (for example, the fish/mammal divergence probably occurred earlier than 550 Mya) (Bowler et al., 2008).

A major finding about diatom biology was the discovery that the genomes of both *T. pseudonana* and *P. tricornutum* contain genes encoding a full urea cycle. This was unanticipated because this pathway has not been previously described in a eukaryotic autotroph. Although animals excrete urea to remove excess nitrogen, diatoms possess an active urease and can grow with urea as the unique source of nitrogen. The urea cycle appears to be fully integrated into diatom metabolism and can lead to the production of metabolites such as proline and polyamines, and the energy-storage molecule creatine-phosphate (Armbrust et al., 2004; Allen et al., 2006).

The genomic era also brought much information about the dynamics of the EGT process. For example, although the *G. theta* and *B. natans* nucleomorph genomes were fully autonomous eukaryotic genomes about 1,300 Mya, they have now been reduced to only a few hundred genes through EGT and gene loss over evolutionary time. Both genomes today contain densely packed genes with numerous overlaps and miniaturized introns (Douglas et al., 2001; Gilson et al., 2006). EGT from genome to genome can be illustrated by the fact that the *T. pseudonana* genome encodes six proteins which are most closely related to genes encoded by the *G. theta* nucleomorph genome. Four of these genes are also found in red algal plastid genomes, thus demonstrating successive EGT from red algal plastid to red algal nucleus (nucleomorph) to heterokont host nucleus (Armbrust et al., 2004). Almost half of the *T. pseudonana* proteins were found to be similarly distant to their closest homolog in the green plant *Arabidopsis thaliana*, the red alga *Cyanidioschyzon merolae* and the animal *Mus musculus*, indicative of the ancient divergence between Plantae, Unikonts, and the unknown secondary host that gave rise to stramenopiles. A total of 806 *T. pseudonana* proteins find matches only in the *M. musculus* proteome, suggesting that these derive from the secondary host.

In addition to EGT, horizontal gene transfer (HGT) can occur independently of an endosymbiotic event and is common among bacteria, archaea, and even between bacteria and archeae (Gogarten and Townsend, 2005; Keeling, 2008). It is though to occur through transformation, transduction, and conjugation processes. The genomes of some bacteria can consist of more than 20% of horizontally acquired genes (Nelson et al. 1999). Although the erroneous report of substantial HGT in the human genome (Stanhope et al., 2001) has probably cooled the field, HGT has also been reported to occur between bacteria and eukaryotes. Moderate HGT from bacteria to some fungi has been documented with ten potential cases in the *Saccharomyces cerevisiae* genome such as an aryl- and alkyl-sulfatase gene of bacterial origin that allows utilization of sulfate from several organic sources (Hall et al., 2005). In the genome of the human parasite *Entamoeba histolytica* (Unikont), the authors found 96 potential cases of bacterial genes acquired through relatively recent HGT (Loftus et al., 2005). 148 potential bacterial genes were also detected in a cDNA library built from a set of rumen ciliates which were proposed to have facilitated the Ciliate colonization of the rumen. These genes provided evidence for the role of HGT in adaptation to new niches because they are absent from the genome of the free-living ciliate *T. thermophila*.

Hence, most HGT observed to date in eukaryotes involves unicellular organisms, although only limited genome-scale analyses have been performed until now. Furthermore, all these examples of HGT engage the integration of genes of bacterial origin. Recently, the multicellular bdelloid rotifer *Adineta vaga* (Unikont) was found to contain genes from bacterial as well as from fungal origin concentrated in telomeric regions rich in transposable elements (TEs). Bdelloid rotifers apparently lack sexual reproduction and can withstand extreme dessication. Their ability to withstand cellular disruption and DNA fragmentation may facilitate the HGT process (Gladyshev et al., 2008). The genome of the green alga *Ostreococcus tauri* has also been hypothesized to have been the target of large scale HGT. In

this genome, two chromosomes differ structurally from the other eighteen and are enriched in TEs. On one of them, chromosome 19, most genes have no similarity with green lineage genes but resemble genes from various origins, mainly bacteria but also fungi and animals (Derelle et al., 2006).

The recent publication of the *P. tricornutum* genome overwhelmed the scale that HGT had been described previously in eukaryotes (Bowler et al., 2008). Indeed, among the 10,402 putative proteins encoded in this ~27 Mb genome, at least 587 appear to be most closely related to bacterial genes, accounting for more than 5% of the *P. tricornutum* proteome. 320 (56%) of these are also found in the *T. pseudonana* genome, attesting their ancient incorporation in the diatom lineage. 59 of these are also found in the *P. sojae* genome, suggesting that most HGT occurred after the divergence of diatoms and other heterokonts. The acquisition of genes from bacterial origin in diatoms has recently been shown to have a crucial role in diatom biology. Indeed, Marchetti and collaborators published in late 2008 the identification of a *Pseudo-nitzschia multiseries* expressed sequence tag (EST) with similarity to a cyanobacterial ferritin-like gene which is used for iron storage. This protein had not been reported previously in any heterokont and the authors also found bacterial-like ferritins in several other pennate diatoms, including *P. tricornutum*, but not in centrics. In *P. multiseries*, the encoded protein was shown to be fully functional and to be up-regulated in response to iron starvation. In corollary with the fact that pennate diatoms can undergo many more cell divisions than centrics under iron limitation, the protein is thought to contribute to the success of pennate diatoms in low-iron oceanic regions (Marchetti et al., 2008).

Figure 1.10: The two diatom species whose genomes have been sequenced and published: (A) The centric diatom *Thalassiosira pseudonana*, (B) The three different morphotypes of the raphid pennate diatom *Phaeodactylum tricornutum* (1: fusiform, 2: oval, 3: triradiate).

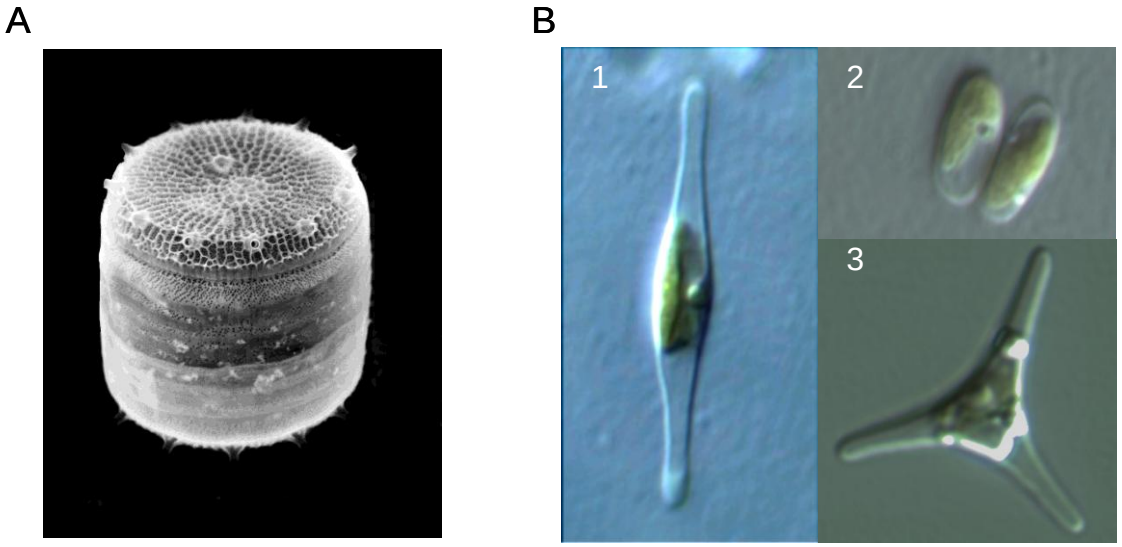


Table 1.1:

Species	Supergroup	order	class	size (Mb)	reference
Guillardia theta nucleomorph	Chromalmveolata	Cryptophyta	Cryptomonadaceae	0,5	Douglas et al., 2001
Thalassiosira pseudonana	Chromalmveolata	Stramenopile	Bacillariophyta (centric diatom)	32,4	Armbrust et al., 2004
Phaeodactylum tricornutum	Chromalmveolata	Stramenopile	Bacillariophyta (pennate diatom)	27,4	Bowler et al., 2008
Phytophthora sojae	Chromalmveolata	Stramenopile	Oomycete	95	Tyler et al., 2006
Tetrahymena thermophila macronucleus	Chromalmveolata	Alveolata	Ciliophora	104	Eisen et al., 2006
Plasmodium falciparum macronucleus	Chromalmveolata	Alveolata	Apicomplexa	23	Gardner et al., 2002
Paramecium	Chromalmveolata	Alveolata	Ciliophora	72	Aury et al., 2006
Ostreococcus tauri	Plantae	Chlorophyta	Prasinophyceae (green algae)	12,6	Derelle et al., 2006
Ostreococcus lucimarinus	Plantae	Chlorophyta	Prasinophyceae (green algae)	13,2	Palenik et al., 2007
Chlamydomonas reinhardtii	Plantae	Chlorophyta	Chlorophyceae	121	Merchant et al., 2007
Cyanidioschyzon merolae	Plantae	Rhodophyta	Cyanidiales (red algae)	16,5	Matsuzaki et al., 2004
Bigelowiella natans nucleomorph	Rhizaria	Cercozoa	Chlorarachniophyceae	0,4	Gilson et al., 2006
Fragilariopsis cylindrus	Chromalmveolata	Stramenopile	Bacillariophyta (pennate diatom)		unpublished
Pseudo-nitzschia multiseries	Chromalmveolata	Stramenopile	Bacillariophyta (pennate diatom)		unpublished
Aureococcus anophagefferens	Chromalmveolata	Stramenopile	Pelagophyceae	56,7	unpublished
Ectocarpus siliculosus	Chromalmveolata	Stramenopile	Phaeophyceae		unpublished
Emiliania huxleyi	Chromalmveolata	Haptophyceae	Isochrysidales	167	unpublished
Micromonas pusilla strain NOUM17	Plantae	Chlorophyta	Prasinophyceae (green algae)	21	unpublished
Micromonas pusilla strain CCMP1545	Plantae	Chlorophyta	Prasinophyceae (green algae)	22	unpublished

1.1.7 The diatom *Phaeodactylum tricornutum*

P. tricornutum has been used in laboratory-based studies of diatom physiology for several decades and is now becoming a model for genetic studies in diatoms. Although not considered to be of great ecological significance, it has been found in several locations around the world, typically in coastal areas with wide fluctutations in salinity. Various ecotypes have been sampled around the world and our laboratory grows and studies more than ten different accessions of *P. tricornutum* (De Martino et al., 2007). Unlike other diatoms *P. tricornutum* can exist in different morphotypes (fusiform, triradiate, and oval, see Figure 10), and changes in cell shape can be stimulated by environmental conditions. This feature can be used to explore the molecular basis of cell shape control and morphogenesis. Furthermore *P. tricornutum* can grow in the absence of silicon, and the biogenesis of silicified frustules is facultative, thereby providing opportunities for experimental exploration of silicon-based nanofabrication in diatoms.

A complete molecular toolbox applied to *P. tricornutum* has been set up during the last decade, largely involving achievements from Chris Bowler's team. Nuclear genetic transformation by biolistic bombardement is used to transform *P. tricornutum* (Falciatore et al., 1999), and a specially constructed Gateway-based system is now available for *P. tricornutum* transformation (Siaut et al., 2007).

1.2 Transposable Elements

1.2.1 Introduction

Transposable elements (TEs) are mobile genetic sequences that inhabit the genomes of most organisms. Within their sequence TEs contain all the genetic information required for their movement in a genome. They can insert into genes or regulatory sequences and therefore potentially impact genes and gene expression. TEs can also trigger chromosomal rearrangements and their accumulation significantly contributes to the physical size of a genome. For example, TEs constitute 42% of the human genome (Lander et al., 2001) and 75% of the maize genome (SanMiguel et al., 1998). As a result, TEs can contribute to genetic innovation by acting as key players in genome and species evolution. Identification of TEs and large-scale analyses have shown that they are scattered throughout the genomes of most organisms.

TEs were first discovered in the early 1950s by Barbara McClintock (1950, 1951) for which she received the Nobel Prize in Physiology or Medicine in 1983. She observed the existence of various "mutable loci" in different maize breeding crosses such as a gene controlling grain color (Figure 1.11). She interpreted her results in these words: "It has been concluded that the changed phenotypic expressions of such loci are related to changes in a chromatin element other than that composing the genes themselves and that mutable loci arise when such chromatin is inserted adjacent to the genes that are showing the variegated expression. The events occurring to this inserted chromatin are reflected in a changed expression of the neighboring genes, or sometimes in a loss of these genes. It is the inserted material that is undergoing the mutational events". Her conclusions implied that activation and repression of gene expression can occur and lead to phenotypic variations. This idea was not well received by the scientific community at that time and she stopped publishing about

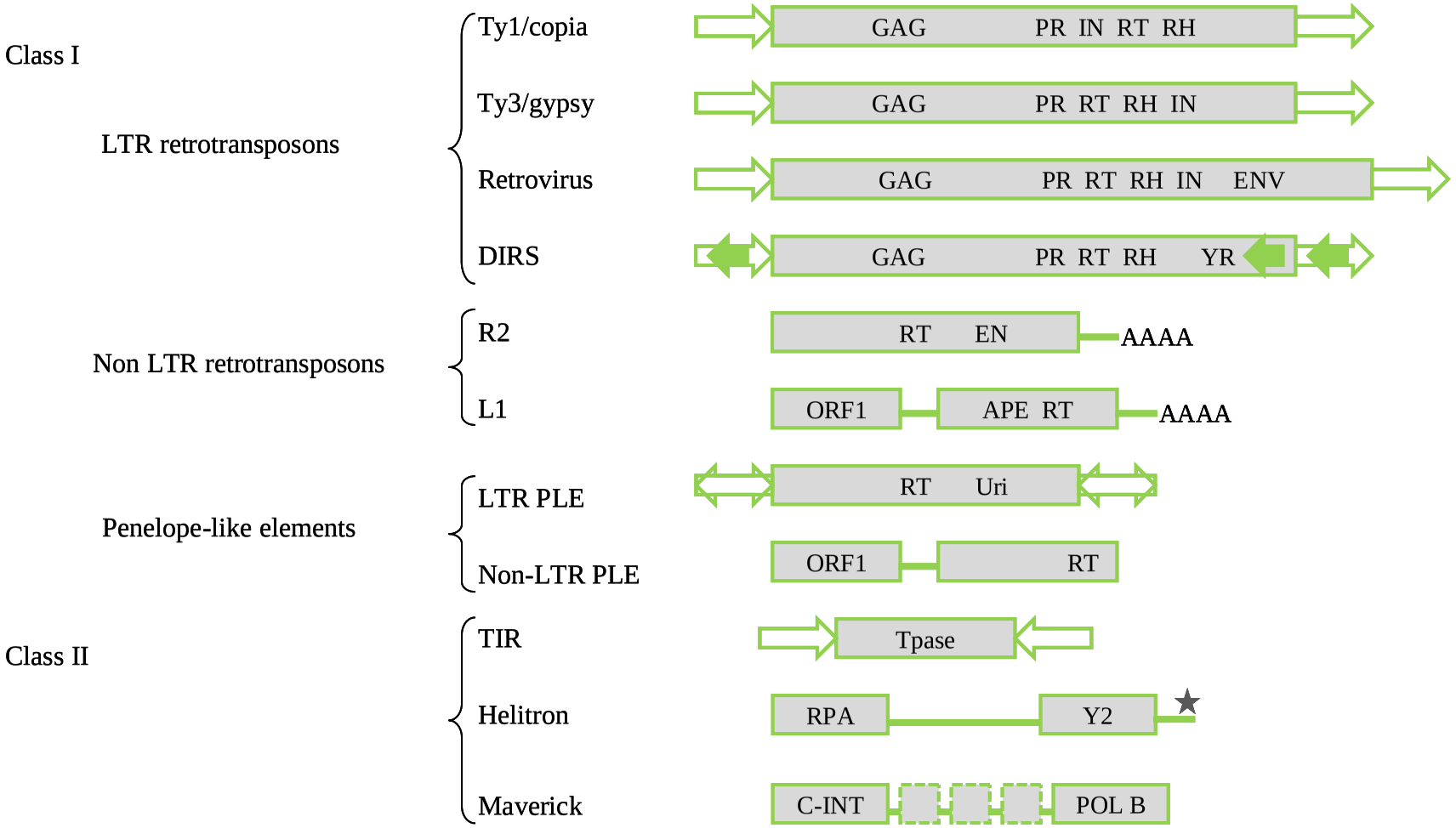
"mutable loci". The importance of her work was highlighted after François Jacob and Jacques Monod described in 1961 the regulation of the Lac operon in *Escherichia coli* and in the same year she wrote a paper entitled "Some parallels between gene control systems in maize and in bacteria" (McClintock, 1961).

Figure 1.11: Illustration of the effect of transposable element mobilization on maize kernel color.

1.2.2 Diversity

There is a large variety of TEs and these are primarily distinguished based on their transposition mechanism. Retrotransposons (class I) transpose via a copy/paste-type mechanism: they use host machinery for transcription followed by reverse transcription of the RNA intermediate into DNA by a TE-encoded reverse transcriptase (RT). DNA transposons (class II) act as DNA intermediate through a cut/paste mechanism and multiply using the host cell's replication machinery. Class I and class II TEs with composite structures are continually being discovered, illustrating the enormous flexibility of these elements (Figure 1.12).

Figure 1.12 (next page): representation of the structure of different types of transposable elements. Abbreviations: group-specific antigen (GAG), protease (PR), integrase (IN), reverse transcriptase (RT), ribonuclease H (RH), envelope (ENV), tyrosine recombinase (YR), endonuclease (EN), apurinic–apyrimidinic endonucleases (APE), Uri endonuclease (Uri), transposase (Tpase), replication protein A (RPA), Y2-type tyrosine recombinase (Y2), c-int-type integrase (C-INT), DNA polymerase B (POLB).


Class I
LTR retrotransposons
Ty1/copia
GAG
PR IN RT RH
Ty3/gypsy
GAG
PR RT RH IN
Retrovirus
GAG
PR RT RH IN ENV
DIRS
GAG
PR RT RH
YR
Non LTR retrotransposons
R2
RT
EN
AAAA
L1
ORF1
APE RT
AAAA
Penelope-like elements
LTR PLE
RT
Uri
Non-LTR PLE
ORF1
RT
Class II
TIR
Tpase
Helitron
RPA
Y2
Maverick
C-INT
POL B

1.2.2.1 Class I elements

As retrotransposons are transcribed from a genomic copy, each replication cycle produces a new copy. Consequently, retrotransposons are often the major contributors to the repetitive fraction of large genomes (Kumar and Bennetzen, 1999). Class I retrotransposons can be divided into those that have long terminal repeats (LTR) at their extremities (LTR retrotransposons) and those that do not (non-LTR retrotransposons).

LTR retrotransposons

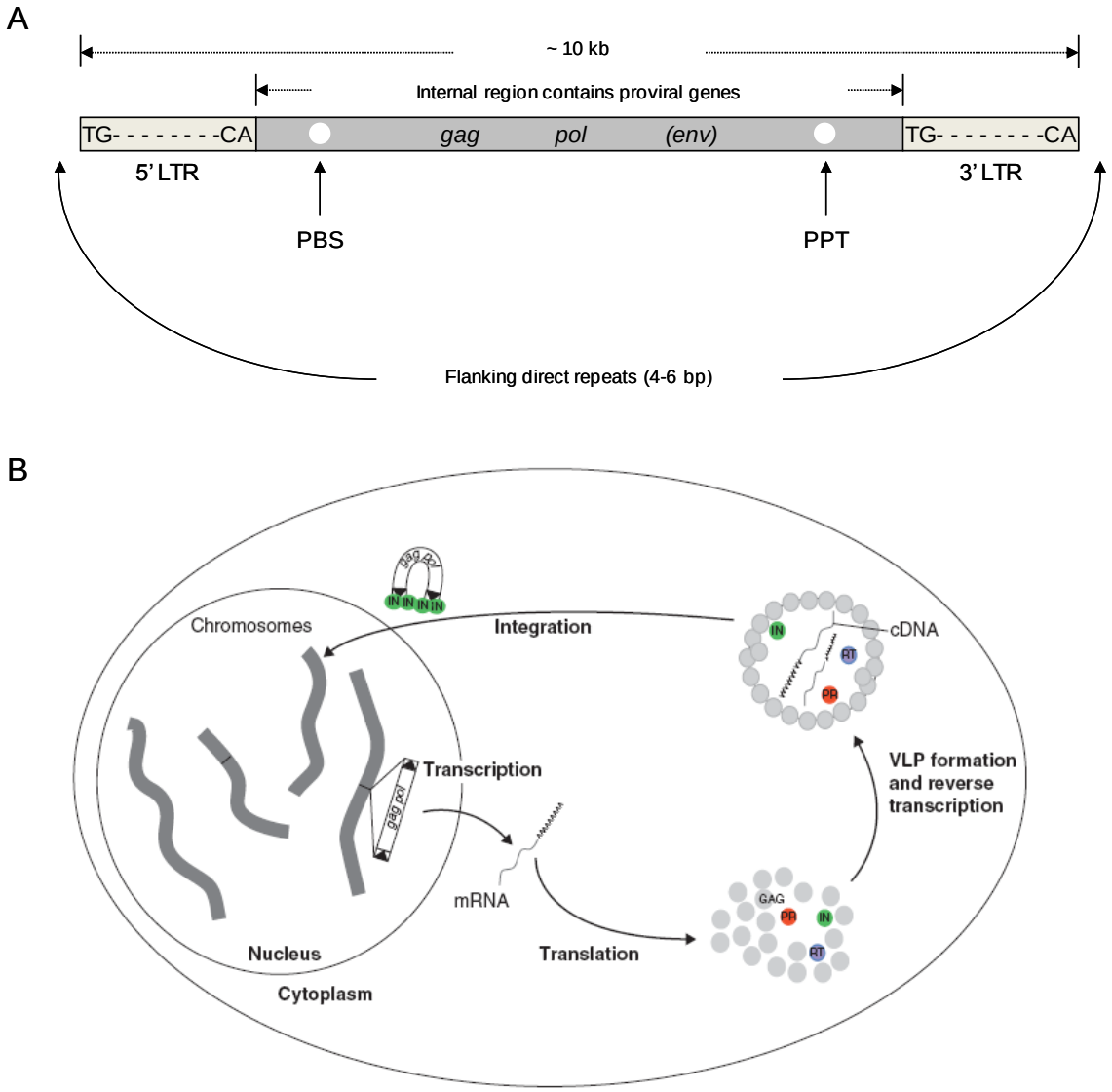
The abundance of LTR retrotransposons (LTR-RT) is usually low in fungi, highly variable in animals, and high in plants. For example, the 75% increase in maize genome size in the last 5 Mya is a result of the proliferation of 11 families of these elements (SanMiguel at el., 1998). On the other hand, LTR-RTs represent less than 8% of the human genome, and no element appears to have been active in this lineage for the last 50 Mya (Langer et al., 2001).

LTRs are the direct sequence repeats that flank the internal region encoding both structural and enzymatic proteins. The *gag* (group specific antigen) gene encodes structural proteins that form the virus-like particle (VLP), inside which reverse transcription takes place. The *pol* (polyprotein) gene encodes several enzymatic functions, including a protease (PR) that cleaves the Pol polyprotein, a reverse transcriptase (RT) that copies the retrotransposon RNA into cDNA, a ribonuclease (RN), and an integrase (IN) that integrates the cDNA into the genome (Figure 1.13A). When Pol proteins are packaged within the VLP, first strand DNA synthesis by reverse transcription is primed by the annealing of the 3' end of a tRNA to a primer binding site (PBS) near the left LTR, while second strand DNA synthesis is primed from a polypurine tract (PPT) near the right LTR. The LTR retrotransposon life cycle can be separated into different phases; transcription, RNA encapsidation and VLP assembly, reverse transcription, nuclear entry, target site regognition, and integration (Figure 1.13B). LTR

retrotransposons have similar structures to those of retroviruses, such as LTR, *gag*, and *pol* genes and, in some cases, an incomplete and nonfunctional *env* (envelop) gene. The retroviral *env* gene codes for a glycoprotein whose polymerized form encloses the retrovirus and is able to cross plasma membranes and resist external conditions. It thus enables retroviruses to spread between individuals in a population. Studies of reverse transcriptase similarities in retrotransposons and retroviruses suggest that on one hand retroviruses may have evolved from LTR retrotransposons by acquiring a functional *env* gene, and that on the other hand the acquisition of a functional *env* gene by LTR retrotransposons could have given rise to other retroviruses (Xiong and Eickbush, 1990; Lerat and Capy, 1999).

Based on the phylogeny of their RT domain, LTR retrotransposons can be divided into three major lineages that are historically referred to as the Ty1/copia group, the Ty3/gypsy group, and the Bel group. The Ty1/copia and Ty3/gypsy groups have extremely broad distributions in animals, plants and fungi, while the Bel group has to date only been reported in animals. Although Ty3/gypsy and Bel elements conserve a *pol* gene organization similar to those of most retroviruses (PR, RT, RH, IN), the IN domain is displaced in the Ty1/copia group (PR, IN, RT, RH), which appears to be the ancestral arrangement. Finally, the DIRS group of LTR retrotransposons is composed of elements very similar to the ones discussed above. However, instead of the integrase domain, these code for a domain with sequence similarity to tyrosine recombinases (YR) for insertion (Goodwin and Poulter, 2001). In addition, they contain LTRs as well as additional LTR segments called internal complementary repeats (ICR) with variable orientation.

Figure 1.13: Schematic representation of (A) the structure of LTR retrotransposons and retroviruses, and (B) of the retrotransposition cycle of an LTR retrotransposon (from Havecker et al., 2004). Abbreviations: primer binding site (PBS), polypurine tract (PPT), virus-like particle (VLP).



Non-LTR retrotransposons

Non-LTR retrotransposons, also called LINEs (Long Interspersed Nucleotide Elements), lack LTRs. Instead LINEs most frequently end with a poly(A) tail at their 3' end. There is a large diversity of non-LTR retrotransposons (Malik et al., 1999) and new lineages continue to be identified. These lineages have somewhat different coding capacities but generally there appear to be two major structures. The first class encodes a single ORF with a centrally located RT domain. The most extensively studied members of this class are the R2 elements in which N-terminal domains show little similarity to each other or to known proteins except that some elements appear to encode DNA-binding motifs (Christensen et al., 2006). C-terminal to the RT domain is another conserved domain that appears to be the endonuclease for the element (EN). The second major class of non-LTR-RT usually encodes two ORFs. The most extensively studied members of this group are the mammalian LINE1 (L1) elements. The first ORF may have functional similarity to the *gag* gene of retroviruses since conserved zinc-finger domains are found in many lineages, and the protein has been shown to bind RNA (Martin and Bushman, 2001). The second ORF encodes the RT domain as well as an endonuclease domain at the N-terminal end. This endonuclease has been termed APE because of sequence similarity to apurinic–apyrimidinic endonucleases involved in DNA repair. L1 elements are highly abundant non-LTR-RTs in mammals with about 800,000 copies in the human genome, representing 17% of the genome (Lander et al., 2001).

The mechanism of retrotransposition for the non-LTR retrotransposons has only been determined in detail for the R2 element. R2 elements insert in a sequence-specific manner in the 28S rRNA genes in at least five animal phyla. The single ORF of R2 is translated into one protein which contains both reverse transcriptase and endonuclease domains. R2 protein subunits bind either the 3' or 5' end of the R2 transcript to generate the complex used for retrotransposition. The RNA bound end determines whether the subunit binds upstream or

downstream of the insertion site on the target DNA. The protein subunit bound upstream of the insertion site cleaves the lower DNA strand and uses the released 3' end to prime reverse transcription of the RNA. This use of the target site to prime reverse transcription has been termed target-primed reverse transcription (TPRT). The protein subunit bound downstream of the insertion site cleaves the upper DNA strand and uses the released 3' end to prime second-strand DNA synthesis. RNA still bound to the first DNA strand is displaced during this synthesis. While the precise details of protein binding to the RNA transcript and the DNA target site may only be relevant for the R2 elements, the general steps of the reaction, including the use of the first and second strand cleavages to prime the two DNA strands is likely to be common for many non-LTR elements (see review by Ostertag and Kazazian, 2001).

Penelope-like elements

The last group of retrotransposons to be identified, the Penelope-like elements (PLE), was discovered only recently in *Drosophila virilis* (Evgen'ev et al., 1997). PLE encodes an RT that is more closely related to telomerase than to the RT from LTR-RT or non-LTR-RT. A second domain has sequence similarity to the Uri endonucleases of bacterial mobile group I introns, as well as UvrC bacterial DNA-repair endonucleases (Lyozin et al., 2001). PLEs have the most diverse structures of any class of retrotransposon. Some elements contain apparent LTRs that may either be in direct or inverted orientations, some contain a first ORF, and some lack the Uri domain. Remarkably, some PLE elements retain a functional spliced intron, which is unexpected for an element that replicates by reverse transcription. The possible phylogenetic relationship with telomerase is particularly intriguing because lineages of PLE retrotransposons have been identified in bdelloid rotifers, fungi and plants that do not encode the Uri domain (Gladyshev and Arkhipova, 2007). These elements are found near or at the telomeres of the host organisms. While the mechanism of integration is not established for

PLE elements, their frequent 5' truncations and the possibility that some elements use the end of a chromosome to prime reverse transcription suggest that these elements utilize a TPRT-like mechanism of retrotransposition.

1.2.2.2 Class II elements

Class II TEs (also called DNA transposons), like class I, are ancient and found in almost all eukaryotic lineages. Class II elements are also found in prokaryotes in simple forms called insertion sequences (IS). Class II contains two subclasses, which are distinguished by the number of DNA strands that are cut during transposition, but neither moves via an RNA intermediate.

Subclass I elements

Subclass 1 comprises classical 'cut-and-paste' TEs which are characterized by a transposase encoded by autonomous copies and, with few exceptions, by the presence of terminal inverted repeats (TIRs) of variable length. To date, nine superfamilies of cut-and-paste DNA transposons are recognized and the transposase of most of them contain a characteristic [D, D, E] catalytic triad. Alliances with prokaryotic insertion sequence families can be drawn for six of the ten eukaryotic superfamilies, suggesting that the divergence of most superfamilies may predate the split of eukaryotes and prokaryotes (Feschotte and Pritham, 2007). These TEs can increase their numbers by transposing during chromosome replication from a position that has already been replicated to another that the replication fork has not yet passed. Transposition is mediated by the self-encoded transposase that recognizes the TIRs and cuts both strands at each end. Subsequently, the transposase binds the TIRs and a sequence of DNA that makes up the target site. Some transposases require a specific sequence as their target site; others can insert the transposon anywhere in the genome. After the transposase cuts DNA at the target

site in a "sticky ends" manner, it ligates the transposon and a DNA polymerase fills in the resulting gaps from the sticky ends, thus creating a target-site duplication of 4-6 bp flanking the element.

Subclass II elements

Subclass II comprises DNA TEs that undergo a transposition process that entails replication without double-stranded cleavage. These TEs transpose by replication involving the displacement of only one strand. Their placement within class II reflects the common lack of an RNA intermediate, but not necessarily a common ancestry.

Elements belonging to the Helitron group appear to replicate via a rolling-circle mechanism, with only one strand cut (Kapitonov and Jurka, 2001). Helitron ends are defined by TC or CTRR motifs (where R is a purine) and a short hairpin structure lying a few nucleotides before the 3′ end, although this does not seem to be a true diagnostic feature. Autonomous Helitrons encode a Y2-type tyrosine recombinase such as that found in the bacterial IS91 rolling-circle transposons, with a helicase domain and replication initiator activity. They can also encode a single-strand binding protein with similarity to Replication Protein A (RPA) or other proteins. Helitrons have been best characterized in maize, in which they carry gene fragments that have been captured from the host genome (Morgante et al., 2005). Although Helitrons have been described mainly in plants, they also exist in animals and fungi, constituting for example 2% of the *Caenorhabditis elegans* genome (Kapitonov and Jurka, 2001).

TEs of the order Maverick (also known as Polintons) are large, up to 10–20 kb, and are bordered by long TIRs. They encode up to 11 proteins, but these vary in number and order. Some show limited homology to proteins of various DNA viruses. All Mavericks encode an INT (c-int type) that is related to those found in some class I TEs, but they do not

contain RT, suggesting that they undergo replicative transposition without RNA intermediates. This is proposed to proceed via excision of a single strand followed by extrachromosomal replication, followed by integration at a new site. DNA polymerase B is also well conserved among the Mavericks but not a general feature. So far, Mavericks have been found sporadically in diverse eukaryotes, but not in plants (see Pritham et al., 2007).

Non-autonomous copies

Many autonomous TEs give rise to non-autonomous derivatives. A TE is considered autonomous if it encodes the domains that are necessary for its transposition, without implying that the element is either functional or active. A family can contain various elements that have been rendered defective by point mutations, small insertions or deletions. These may have retained sufficiently high DNA identity for family membership. Such elements that are defective owing to their limited mutations were probably functional in the recent past (Sabot and Schulman, 2006). In contrast, non-autonomous TEs define a group of elements that have lost some or all domains and structures found in autonomous elements. Usually, non-autonomous elements have a highly degenerate coding region, or even completely lack coding capacity but still share sequence conservation and/or specific characteristics with autonomous relatives from which they borrow machinery for transposition. SINEs, which rely on LINEs for transposition functions such as reverse transcription, constitute an extreme case of non-autonomous elements since they do not derive from TEs. Instead, they originate from accidental retrotransposition of various polymerase III (Pol III) transcripts such as transfer RNAs and contain an internal Pol III promoter allowing them to be expressed. The Pol III promoter defines SINE superfamilies and reveals their origin: tRNA, 7SL RNA and 5S RNA. SINE internal regions (50–200 bp) are family-specific and of are variable origin, and the 3′ region is usually made of short repeats (1-5 bp) (Kramerov and Vassetzky, 2005). The best

known SINE is the Alu element, which is present in at least 500,000 copies in the human genome (Rowold and Herrara, 2000).

1.2.3 Impact of TEs on genome evolution

The impact of TEs on genome organization, gene regulation and genome size, and species evolution in general is nowadays broadly accepted and TEs are considered major drivers of genome/species evolution. I will mention below some of the TE-driven evolutionary processes demonstrating that TEs do act both on species evolution at the population level and over evolutionary time.

1.2.3.1 Domestication

The most striking beneficial contribution of TEs is illustrated by an evolutionary process referred to as "molecular domestication", by which a TE-derived coding sequence gives rise to a functional host gene. The protein products of domesticated TEs are phylogenetically linked to transposon-encoded proteins and assume biological roles in vivo. These exist as single copies in the genome, are devoid of the molecular hallmarks of transposition such as flanking repeats, and have lost their capacity to mediate transposition. Nowadays, a large number of such cases have been reported (see Sinzelle et al., 2009).

Jawed vertebrates' V(D)J recombination during lymphocyte development is certainly one of the most spectacular examples. In this process, preexisting V (variable), D (diversity), and J (joining) gene segments are rearranged to generate a large repertoire of T cell surface receptors and immunoglobulins. The recombination event involves cis-acting sequences known as recombination signal sequences (RSSs) that flank each receptor gene segment and two proteins encoded by the recombination-activating genes RAG1 and RAG2. Mechanistically, the V(D)J recombination reaction shares significant similarities with the

excision step of the cut-and-paste transposition process by which subclass I DNA transposons are excised from the donor-site DNA via double-strand breaks (DSB) (Agrawal et al., 1998). In addition, the C-terminal domain of RAG1 including the [D, D, E] catalytic triad, the structure of the RSSs, as well as the characteristic TSDs, strongly support that RAG1 and the RSSs originate from a formerly active transposon (Kapitonov and Jurka, 2005).

1.2.3.2 The generation of genetic variability in response to stress

TEs can essentially move into any locus in a genome. Their potentially deleterious effects are huge respect to rare and fortuitous insertions that can bring selective advantages. Thus, TEs could have highly adverse effects on a population in a short period of time if their expression and transposition were not monitored. In contrast, the idea that transposable elements might activate because they benefit their host was proposed by McClintock (1984) who suggested that TE activity could be a response to challenges to the genome. In this model, the deleterious effects of transposition could be minimized by maintaining active elements in a quiescent state during normal growth and development. In life-threatening situations, elements could be activated, thereby increasing the mutation rate and restructuring the genome. Populations able to diversify their genomes rapidly in this manner are thought to be more likely to survive changing conditions and to produce progeny. In this regard, TE movements in response to stress can be interpreted as a strategy of long term genomic adaptation (see Capy et al., 2000). TE activation in response to stressful conditions is well known and documented in plants (see Wessler, 1996) and a other cases have been reported in yeast (Servant et al., 2008;) and animals (Vieira and Biemont, 1996). In general, a single family responds to a specific stress and has evolved ad hoc regulatory sequences.

In the year 2000, Kalendar et al. published an in situ real life example of local environment genome shaping by LTR retrotransposons. The authors examined the role of the

BARE-1 LTR-RT in genome diversification of wild barley from the Evolution Canyon microsite, Lower Nahal Oren, Mount Carmel, Israel (Nevo, 1997). This 400 meter wide erosion gorge dating from about 2 Mya presents north- and south-facing slopes with common geologies and macroclimates but microclimates sharply differing in solar irradiation and aridity. The data suggest a linkage between BARE-1 numbers and the ecogeography of the Evolution Canyon microsite: more BARE-1 copies are found in plants sampled on the drier south-facing slope within the canyon. This is especially relevant considering the presence within the BARE-1 promoter of abscisic acid-response elements typical for water stress-induced genes (Manninen and Schulman, 1993). The data therefore suggest that BARE-1 propagation may be beneficial to barley populations in arid environments by generating genetic variability.

New TE insertions can subsequently cause differential expression of genes in the vicinity of the integration site. TEs may insert into cis-regulatory sequences and alter the regulation of transcription. Second, because TEs can carry their own regulatory sequences, they can convey these in the genome by moving around. For example, new insertions of the yeast adenine starvation responsive element *Ty1* into the vicinity of a reporter gene was recently shown to confer adenine starvation regulation to it. The authors also provided evidence that similar coactivation of genes adjacent to *Ty1* sequences occurs naturally in the yeast genome, indicating that *Ty1* insertions can mediate transcriptional control of yeast gene expression (Servant et al., 2008).

1.2.3.3 TE-mediated recombination

Genome size is clearly the result of a balance between amplification and loss of DNA. In addition to adding physical size to a genome, TEs can also be responsible for various kinds of

chromosomal rearrangements that can lead to DNA loss, inversions, macrotranspositions and exon shuffling through genetic recombination.

Genetic recombination is the process by which a strand of DNA is broken and then joined to a different DNA molecule. In eukaryotes recombination commonly occurs during double-strand break (DSB) repair. During meiosis, this process causes the shuffling of genes in the gametes and potentially brings evolutionary advantages to the offspring as it leads to the selection of one out of two alleles (one out of many at the population/multi-generational scale) at a given genetic position. Crossover usually occurs when matching regions on matching chromosomes break and then reconnect to the other chromosome. However, crossovers can also occur between similar sequences in mismatched alignments. These processes, called unbalanced recombination, can result in a deletion on one chromosome and duplication on the other (among other possible outcomes). Because TEs from the same family can share a high level of sequence similarity, they potentially constitute a major source of unbalanced recombination.

In the case of class I transposons, the frequency of recombination may increase as transposases can themselves cause DSBs. In addition, transposase-induced rearrangements can connect a pair of 5' and 3' termini from two different transposon copies and engage a transposition reaction in a process called alternative transposition (see Gray, 2000). Depending on the orientation of the termini used for the reaction and the respective chromosomal location of the elements engaged, this process can generate various outcomes including macrotranspositions, inversions, duplications and deletions (Zhang and Peterson, 2004; Huang and Dooner, 2008). In addition, local hopping is a preference of many DNA transposons to transpose into a linked chromosomal location (Guimond et al., 2003) which may also augment their propensity to trigger local rearrangements as recombination frequency is negatively correlated with genetic distances. Furthermore, various LTR retrotransposons, in

particular Ty3/gypsy-like elements with chromodomains (also called chromovirus) have an integration bias towards TE-rich (or euchromatin) regions (Gao et al., 2008). Although many of the TE-mediated chromosomal rearrangements may be deleterious in nature, at the population scale some may occasionally bring a selective advantage to the individuals carrying them such as gene creation by exon shuffling/fusion. For example, Zhang and collaborators (2006) recently reported that alternative transposition induced by the Ac DNA transposon in maize can join the coding and regulatory sequences of two linked paralogous genes and generate a series of chimeric genes, some of which are functional and specify a novel orange pericarp phenotype.

1.2.3.4 Gene duplication

Gene duplication is another process influencing species evolution. The new copy of a gene is often free from selective pressure and mutates faster than the original copy because mutations in it have no deleterious effects to the host organism. It may thus acquire new functions and/or regulatory sequences. As seen above, duplications can occur through genetic recombination involved during the repair of DSBs in the case of misaligned homologous chromosomes (TE-mediated or not) and through alternative transposition. In addition, both class I and class II TEs can more directly mediate a third mode of gene duplication. In the case of class I elements, gene duplication can occur by readthrough of transcription to cellular genes flanking the 3' terminus of a retroelement, reverse transcription of mature mRNA, and subsequent insertion into genomic DNA (Moran et al., 1999; Elrouby and Bureau, 2001). This mechanism is called transduction and is thought to be at the origin of the assembly of Class I TEs by successive gathering and fusion of exons from an ancestral host genome. The mechanism of gene duplication by Class II elements, known as transduplication, is unknown but is clearly different from transduction because duplicated fragments retain introns and are

found between the elements termini (Juretic et al. 2005). In the view of the above discussion, class I elements likely acquire gene(s) during DSB repair. The genetic segment formed may be disabled to transpose autonomously but can constitute a non autonomous element (see above). Different classes of DNA transposons have been documented to gather pieces of genes and to pepper throughout a genome such as the subclass I Mutator-like elements (MULE) (Chandler et al., 1986), CACTA (Kawasaki and Nitasaka 2004) and the subclass II Helitrons. As example of the latter, analysis of several insertions in the maize genome were defined by Helitron-like termini and contained different sets of gene-derived fragments (leaving the original copy intact), supporting the idea that transduplication may have a role in exon shuffling and the formation of new genes (Morgante et al., 2005).

1.3 References:

Agrawal A, Eastman QM, Schatz DG: Transposition mediated by RAG1 and RAG2 and its implications for the evolution of the immune system. *Nature* 1998, 394:744-751.

Allen AE, Vardi A, Bowler C: An ecological and evolutionary context for integrated nitrogen metabolism and related signaling pathways in marine diatoms. *Curr Opin Plant Biol* 2006, 9:264-273.

Archibald JM, Longet D, Pawlowski J, Keeling PJ: A novel polyubiquitin structure in Cercozoa and Foraminifera: evidence for a new eukaryotic supergroup. *Mol Biol Evol* 2003, 20:62-66.

Armbrust EV, Berges JA, Bowler C, Green BR, Martinez D, Putnam NH, Zhou S, Allen AE, Apt KE, Bechner M, et al: The genome of the diatom Thalassiosira pseudonana: ecology, evolution, and metabolism. *Science* 2004, 306:79-86.

Becker B, Hoef-Emden K, Melkonian M: Chlamydial genes shed light on the evolution of photoautotrophic eukaryotes. *BMC Evol Biol* 2008, 8:203.

Bowler C, Allen AE, Badger JH, Grimwood J, Jabbari K, Kuo A, Maheswari U, Martens C, Maumus F, Otillar RP, et al: The Phaeodactylum genome reveals the evolutionary history of diatom genomes. *Nature* 2008, 456:239-244.

Boyd PW, Watson AJ, Law CS, Abraham ER, Trull T, Murdoch R, Bakker DC, Bowie AR, Buesseler KO, Chang H, et al: A mesoscale phytoplankton bloom in the polar Southern Ocean stimulated by iron fertilization. *Nature* 2000, 407:695-702.

Burki F, Shalchian-Tabrizi K, Minge M, Skjaeveland A, Nikolaev SI, Jakobsen KS, Pawlowski J: Phylogenomics reshuffles the eukaryotic supergroups. *PLoS ONE* 2007, 2:e790.

Capy P, Gasperi G, Biemont C, Bazin C: Stress and transposable elements: co-evolution or useful parasites? *Heredity* 2000, 85 (Pt 2):101-106.

Cassar N, Bender ML, Barnett BA, Fan S, Moxim WJ, Levy H, 2nd, Tilbrook B: The Southern Ocean biological response to aeolian iron deposition. *Science* 2007, 317:1067-1070.

Cavalier-Smith T: Principles of protein and lipid targeting in secondary symbiogenesis: euglenoid, dinoflagellate, and sporozoan plastid origins and the eukaryote family tree. *J Eukaryot Microbiol* 1999, 46:347-366.

Cavalier-Smith T: The phagotrophic origin of eukaryotes and phylogenetic classification of Protozoa. *Int J Syst Evol Microbiol* 2002, 52:297-354.

Chandler VL, Walbot V: DNA modification of a maize transposable element correlates with loss of activity. *Proc Natl Acad Sci U S A* 1986, 83:1767-1771.

Christensen SM, Ye J, Eickbush TH: RNA from the 5' end of the R2 retrotransposon controls R2 protein binding to and cleavage of its DNA target site. *Proc Natl Acad Sci U S A* 2006, 103:17602-17607.

De Martino A, Meichenin A, Shi J, Pan KH, Bowler C: Genetic and phenotypic characterization of Phaeodactylum tricornutum (Bacillariophyceae) accessions. *Journal of Phycology* 2007, 43:992-1009.

Derelle E, Ferraz C, Rombauts S, Rouze P, Worden AZ, Robbens S, Partensky F, Degroeve S, Echeynie S, Cooke R, et al: Genome analysis of the smallest free-living eukaryote Ostreococcus tauri unveils many unique features. *Proc Natl Acad Sci U S A* 2006, 103:11647-11652.

Douglas S, Zauner S, Fraunholz M, Beaton M, Penny S, Deng LT, Wu X, Reith M, Cavalier-Smith T, Maier UG: The highly reduced genome of an enslaved algal nucleus. *Nature* 2001, 410:1091-1096.

Evgen'ev MB, Zelentsova H, Shostak N, Kozitsina M, Barskyi V, Lankenau DH, Corces VG: Penelope, a new family of transposable elements and its possible role in hybrid dysgenesis in Drosophila virilis. *Proc Natl Acad Sci U S A* 1997, 94:196-201.

Falciatore A, Casotti R, Leblanc C, Abrescia C, Bowler C: Transformation of Nonselectable Reporter Genes in Marine Diatoms. *Mar Biotechnol (NY)* 1999, 1:239-251.

Feschotte C, Pritham EJ: DNA transposons and the evolution of eukaryotic genomes. *Annu Rev Genet* 2007, 41:331-368.

Field CB, Behrenfeld MJ, Randerson JT, Falkowski P: Primary production of the biosphere: integrating terrestrial and oceanic components. *Science* 1998, 281:237-240.

Foster RA, Zehr JP: Characterization of diatom-cyanobacteria symbioses on the basis of nifH, hetR and 16S rRNA sequences. *Environ Microbiol* 2006, 8:1913-1925.

Gao X, Hou Y, Ebina H, Levin HL, Voytas DF: Chromodomains direct integration of retrotransposons to heterochromatin. *Genome Res* 2008, 18:359-369.

Gilson PR, Su V, Slamovits CH, Reith ME, Keeling PJ, McFadden GI: Complete nucleotide sequence of the chlorarachniophyte nucleomorph: nature's smallest nucleus. *Proc Natl Acad Sci U S A* 2006, 103:9566-9571.

Gladyshev EA, Arkhipova IR: Telomere-associated endonuclease-deficient Penelope-like retroelements in diverse eukaryotes. *Proc Natl Acad Sci U S A* 2007, 104:9352-9357.

Gogarten JP, Townsend JP: Horizontal gene transfer, genome innovation and evolution. *Nat Rev Microbiol* 2005, 3:679-687.

Goodwin TJ, Poulter RT: The DIRS1 group of retrotransposons. *Mol Biol Evol* 2001, 18:2067-2082.

Gray YH: It takes two transposons to tango: transposable-element-mediated chromosomal rearrangements. *Trends Genet* 2000, 16:461-468.

Guimond N, Bideshi DK, Pinkerton AC, Atkinson PW, O'Brochta DA: Patterns of Hermes transposition in Drosophila melanogaster. *Mol Genet Genomics* 2003, 268:779-790.

Hackett JD, Yoon HS, Li S, Reyes-Prieto A, Rummele SE, Bhattacharya D: Phylogenomic analysis supports the monophyly of cryptophytes and haptophytes and the association of rhizaria with chromalveolates. *Mol Biol Evol* 2007, 24:1702-1713.

Hall C, Brachat S, Dietrich FS: Contribution of horizontal gene transfer to the evolution of Saccharomyces cerevisiae. *Eukaryot Cell* 2005, 4:1102-1115.

Harper JT, Keeling PJ: Nucleus-encoded, plastid-targeted glyceraldehyde-3-phosphate dehydrogenase (GAPDH) indicates a single origin for chromalveolate plastids. *Mol Biol Evol* 2003, 20:1730-1735.

Harper JT, Waanders E, Keeling PJ: On the monophyly of chromalveolates using a six-protein phylogeny of eukaryotes. *Int J Syst Evol Microbiol* 2005, 55:487-496.

Havecker ER, Gao X, Voytas DF: The diversity of LTR retrotransposons. *Genome Biol* 2004, 5:225.

Huang JT, Dooner HK: Macrotransposition and other complex chromosomal restructuring in maize by closely linked transposons in direct orientation. *Plant Cell* 2008, 20:2019-2032.

Juretic N, Hoen DR, Huynh ML, Harrison PM, Bureau TE: The evolutionary fate of MULE-mediated duplications of host gene fragments in rice. *Genome Res* 2005, 15:1292-1297.

Kalendar R, Tanskanen J, Immonen S, Nevo E, Schulman AH: Genome evolution of wild barley (Hordeum spontaneum) by BARE-1 retrotransposon dynamics in response to sharp microclimatic divergence. *Proc Natl Acad Sci U S A* 2000, 97:6603-6607.

Kapitonov VV, Jurka J: Rolling-circle transposons in eukaryotes. *Proc Natl Acad Sci U S A* 2001, 98:8714-8719.

Kapitonov VV, Jurka J: RAG1 core and V(D)J recombination signal sequences were derived from Transib transposons. *PLoS Biol* 2005, 3:e181.

Kawasaki S, Nitasaka E: Characterization of Tpn1 family in the Japanese morning glory: En/Spm-related transposable elements capturing host genes. *Plant Cell Physiol* 2004, 45:933-944.

Keeling PJ, Burger G, Durnford DG, Lang BF, Lee RW, Pearlman RE, Roger AJ, Gray MW: The tree of eukaryotes. *Trends Ecol Evol* 2005, 20:670-676.

Keeling PJ, Palmer JD: Horizontal gene transfer in eukaryotic evolution. *Nat Rev Genet* 2008, 9:605-618.

Kooistra WH, Gersonde R, Medlin LK: *Evolution of primary producers in the sea.* Academic Press 2007 edn.

Kooistra WH, Medlin LK: Evolution of the diatoms (Bacillariophyta). IV. A reconstruction of their age from small subunit rRNA coding regions and the fossil record. *Mol Phylogenet Evol* 1996, 6:391-407.

Kramerov DA, Vassetzky NS: Short retroposons in eukaryotic genomes. *Int Rev Cytol* 2005, 247:165-221.

Kumar A, Bennetzen JL: Plant retrotransposons. *Annu Rev Genet* 1999, 33:479-532.

Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W, et al: Initial sequencing and analysis of the human genome. *Nature* 2001, 409:860-921.

Lerat E, Capy P: Retrotransposons and retroviruses: analysis of the envelope gene. *Mol Biol Evol* 1999, 16:1198-1207.

Loftus B, Anderson I, Davies R, Alsmark UC, Samuelson J, Amedeo P, Roncaglia P, Berriman M, Hirt RP, Mann BJ, et al: The genome of the protist parasite Entamoeba histolytica. *Nature* 2005, 433:865-868.

Lyozin GT, Makarova KS, Velikodvorskaja VV, Zelentsova HS, Khechumian RR, Kidwell MG, Koonin EV, Evgen'ev MB: The structure and evolution of Penelope in the virilis species group of Drosophila: an ancient lineage of retroelements. *J Mol Evol* 2001, 52:445-456.

Maheswari U, Mock T, Armbrust EV, Bowler C: Update of the Diatom EST Database: a new tool for digital transcriptomics. *Nucleic Acids Res* 2008.

Malik HS, Burke WD, Eickbush TH: The age and evolution of non-LTR retrotransposable elements. *Mol Biol Evol* 1999, 16:793-805.

Manninen I, Schulman AH: BARE-1, a copia-like retroelement in barley (Hordeum vulgare L.). *Plant Mol Biol* 1993, 22:829-846.

Marchetti A, Parker MS, Moccia LP, Lin EO, Arrieta AL, Ribalet F, Murphy ME, Maldonado MT, Armbrust EV: Ferritin is used for iron storage in bloom-forming marine pennate diatoms. *Nature* 2009, Jan 22;457(7228):467-70

Martin SL, Bushman FD: Nucleic acid chaperone activity of the ORF1 protein from the mouse LINE-1 retrotransposon. *Mol Cell Biol* 2001, 21:467-475.

McClintock B: The origin and behavior of mutable loci in maize. *Proc Natl Acad Sci U S A* 1950, 36:344-355.

McClintock B: Chromosome organization and genic expression. *Cold Spring Harb Symp Quant Biol* 1951, 16:13-47.

McClintock B: Some parallels between gene control systems in maize and in bacteria. *American Naturalist* 1961, 95:265-277.

McClintock B: The significance of responses of the genome to challenge. *Science* 1984, 226:792-801.

Moran JV, DeBerardinis RJ, Kazazian HH, Jr.: Exon shuffling by L1 retrotransposition. *Science* 1999, 283:1530-1534.

Morgante M, Brunner S, Pea G, Fengler K, Zuccolo A, Rafalski A: Gene duplication and exon shuffling by helitron-like transposons generate intraspecies diversity in maize. *Nat Genet* 2005, 37:997-1002.

Nelson DM, Treguer P, Brzezinski MA, Leynaert A, Queguiner B: Production and dissolution of biogenic silica in the ocean – revised global estimates, comparison with regional data and relationship to biogenic sedimentation. *Global Biogeochemical Cycles* 1995, 9:359-372.

Nelson KE, Clayton RA, Gill SR, Gwinn ML, Dodson RJ, Haft DH, Hickey EK, Peterson JD, Nelson WC, Ketchum KA, et al: Evidence for lateral gene transfer between Archaea and bacteria from genome sequence of Thermotoga maritima. *Nature* 1999, 399:323-329.

Nevo E: Evolution in action across phylogeny caused by microclimatic stresses at "Evolution Canyon". *Theor Popul Biol* 1997, 52:231-243.

Ostertag EM, Kazazian HH, Jr.: Biology of mammalian L1 retrotransposons. *Annu Rev Genet* 2001, 35:501-538.

Patron NJ, Rogers MB, Keeling PJ: Gene replacement of fructose-1,6-bisphosphate aldolase supports the hypothesis of a single photosynthetic ancestor of chromalveolates. *Eukaryot Cell* 2004, 3:1169-1175.

Pritham EJ, Putliwala T, Feschotte C: Mavericks, a novel class of giant transposable elements widespread in eukaryotes and related to DNA viruses. *Gene* 2007, 390:3-17.

Reyes-Prieto A, Weber AP, Bhattacharya D: The origin and establishment of the plastid in algae and plants. *Annu Rev Genet* 2007, 41:147-168.

Rogers MB, Gilson PR, Su V, McFadden GI, Keeling PJ: The complete chloroplast genome of the chlorarachniophyte Bigelowiella natans: evidence for independent origins of chlorarachniophyte and euglenid secondary endosymbionts. *Mol Biol Evol* 2007, 24:54-62.

Rowold DJ, Herrera RJ: Alu elements and the human genome. *Genetica* 2000, 108:57-72.

Sabot F, Schulman AH: Parasitism and the retrotransposon life cycle in plants: a hitchhiker's guide to the genome. *Heredity* 2006, 97:381-388.

SanMiguel P, Gaut BS, Tikhonov A, Nakajima Y, Bennetzen JL: The paleontology of intergene retrotransposons of maize. *Nat Genet* 1998, 20:43-45.

Servant G, Pennetier C, Lesage P: Remodeling yeast gene transcription by activating the Ty1 long terminal repeat retrotransposon under severe adenine deficiency. *Mol Cell Biol* 2008, 28:5543-5554.

Siaut M, Heijde M, Mangogna M, Montsant A, Coesel S, Allen A, Manfredonia A, Falciatore A, Bowler C: Molecular toolbox for studying diatom biology in Phaeodactylum tricornutum. *Gene* 2007, 406:23-35.

Sims PA, Mann DG, Medlin LK: Evolution of the diatoms: insights from fossil, biological and molecular data. *Phycologia* 2006, 45:361-402.

Sinzelle L, Izsvak Z, Ivics Z: Molecular domestication of transposable elements: From detrimental parasites to useful host genes. *Cell Mol Life Sci* 2009.

Smith SV: Marine Macrophytes as a Global Carbon Sink. *Science* 1981, 211:838-840.

Sorhannus U: A nuclear-encoded small-subunit ribosomal RNA timescale for diatom evolution. *Marine Micropaleontology* 2007, 65:1-12.

Stanhope MJ, Lupas A, Italia MJ, Koretke KK, Volker C, Brown JR: Phylogenetic analyses do not support horizontal gene transfers from bacteria to vertebrates. *Nature* 2001, 411:940-944.

Stechmann A, Cavalier-Smith T: Rooting the eukaryote tree by using a derived gene fusion. *Science* 2002, 297:89-91.

Stechmann A, Cavalier-Smith T: The root of the eukaryote tree pinpointed. *Curr Biol* 2003, 13:R665-666.

Treguer P, Nelson DM, Van Bennekom AJ, Demaster DJ, Leynaert A, Queguiner B: The Silica Balance in the World Ocean: A Reestimate. *Science* 1995, 268:375-379.

Venrick EL: Spring in the California current: the distribution of phytoplankton species, April 1993 and April 1995. *Marine Ecology-Progress Series* 1998, 167:73-88.

Vieira C, Biemont C: Geographical variation in insertion site number of retrotransposon 412 in Drosophila simulans. *J Mol Evol* 1996, 42:443-451.

Wessler SR: Turned on by stress. Plant retrotransposons. *Curr Biol* 1996, 6:959-961.

Xiong Y, Eickbush TH: Origin and evolution of retroelements based upon their reverse transcriptase sequences. *EMBO J* 1990, 9:3353-3362.

Yoon HS, Hackett JD, Bhattacharya D: A single origin of the peridinin- and fucoxanthin-containing plastids in dinoflagellates through tertiary endosymbiosis. *Proc Natl Acad Sci U S A* 2002, 99:11724-11729.

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

Zhang J, Peterson T: Transposition of reversed Ac element ends generates chromosome rearrangements in maize. *Genetics* 2004, 167:1929-1937.

Zhang J, Zhang F, Peterson T: Transposition of reversed Ac element ends generates novel chimeric genes in maize. *PLoS Genet* 2006, 2:e164.

Chapter II:

Transcription factors in diatom genomes

2.1 Introduction

In all organisms, the regulation and coordination of growth, development, and cell cycle progression, as well as the physiological and metabolic adaptation to a variable environment, depends on the regulation of gene expression. A major step in the modulation of gene expression is the regulation of transcription, which controls when a gene is expressed and how much RNA is generated from a DNA template. Transcription factors (TFs) are key transcriptional regulators (TRs) that can bind to specific DNA sequences called *cis*-acting elements found in the promoters of target genes. They thereby modulate transcriptional timing and levels by influencing positively or negatively the efficiency of the transcriptional machinery to bind and transcribe a target gene. Transcriptional rates are also governed by other transcriptional regulators (TRs) that may not directly bind cis-regulatory elements but act by interacting with the transcriptional machinery or by introducing chromatin modifications rendering a gene more or less prone to transcriptional activity. Most known TFs can be grouped into different families according to their DNA binding domains. Other TRs can be classified according to their enzymatic activity or to their affinity for particular epitopes such as modified histones or 5-methyl cytosine.

With the completion of genome sequences, the entire set of genes encoding known TRs can be identified, described, and compared between species and lineages. These genome-wide analyses by computational methods are important tasks that provide information about TF families and their peculiarities within individual species or lineages and are of particular interest from an evolutionary point of view to help understanding how developmental and physiological diversification has occurred over time. For example, over 1500 proteins or about 5% of the *Arabidopsis thaliana* proteome consists of TFs, about 45% of which constitute plant-specific families. The MYB family has been substantially amplified in Arabidopsis with about 190 copies compared with 38 in *Drosophila melanogaster*, 21 in

Caenorhabditis elegans, and 23 in *Saccharomyces cerevisiae*. This observation suggests that MYB TFs have evolved plant-specific regulatory functions (Riechmann et al., 2000). Because complex organisms contain large numbers of transcriptional regulators, it has been suggested that organismal complexity correlates with increased absolute numbers and proportions of TRs in a proteome (Levine and Tjian, 2003). This is consistent with the observation that the number of MYBs has gradually increased during the evolution of the Viridiplantae lineage, with about 10 copies in the single-celled green algae *Chlamydomonas reinhardtii* and *Ostreococcus tauri* and the red alga *Cyanidioschyzon merolae*, about 60 copies in the moss *Physcomitrella patens*, and over two hundred copies in the tree *Populus trichocarpa* (Riano-Pachon et al., 2008).

The contemporary ocean covers more than 70% of the surface of the Earth and the phytoplankton within the photic zone are responsible for about 45% of our planet's annual net primary productivity (NPP) (Field, et al.,1998). Diatoms (Bacillariophyceae) are unicellular chlorophyll c-containing photosynthetic eukaryotes that populate all kinds of humid environments and constitute one of the most successful groups of phytoplankton, being responsible for about 40% of marine NPP. They can be divided into two major subclasses: the centrics that are radially symmetrical and appeared about 200 Mya, and the pennates that show bilateral symmetry and evolved around 90 Mya. The complete genome sequences of a centric (Armbrust et al., 2004) and a pennate (Bowler et al., 2008) diatom have recently become available. Analysis and comparison of both genomes has unveiled the complex evolutionary history of these organisms. They are members of the eukaryotic group known as Chromista which unifies diverse lineages with red algal plastids including cryptophytes, haptophytes, and stramenopiles, all of which are believed to be derived from a secondary endosymbiotic event involving a red alga and a eukaryotic heretotroph. Stramenopiles are a diverse group that includes the diatoms as well as non-photosynthetic oomycetes such as

Phytophthora, and a range of chlorophyll c-containing unicellular and multicellular brown algae such as *Aureococcus anophagefferens* (Pelagophyceae) and *Ectocarpus siliculosus* (Phaeophyceae).

Genome-wide analysis of TF complements are still lacking for diatoms and for stamenopiles in general. However, the identification of the TF complement in specific species and its comparison with the TF repertoire of other organisms is important in order to evaluate their role in lineage-specific evolution (e.g., lineage-specific expansion of particular TF families or gene fusions). Being the most abundant eukaryotic phytoplankton in contemporary oceans, diatoms may have adapted transcription-based regulatory pathways that contribute to their ecological success. In this work, we have performed a genome-wide computational identification and analysis of the TF complement in the two sequenced diatom genomes, and compared it to those of other stramenopiles, specifically *A. anophagefferens*, *E. siliculosus*, *Phytophthora sojae* and *Phytophthora ramorum*, as well as the haptophyte *Emiliania huxleyi*, and a range of other eukaryotes. Sequence annotation and analysis were mainly contributed by Edda Rayko. We investigated in detail the main families of TFs found in diatom genomes and analyzed their abundance in different cDNA libraries derived from *P. tricornutum* and *T. pseudonana* cells grown in a range of different conditions.

2.2 Results and Discussion

2.2.1 Transcription factor content in stramenopiles

We identified a total of 204 and 247 TFs in the *P. tricornutum* and *T. pseudonana* genomes, respectively, belonging to 16 distinct families of TFs. These diatoms therefore dedicate ~2% of their proteomes to genes encoding TFs (Table 2.1). We also identified 147, 283, 298, 301, and 406 TFs in the *A. anophagefferens*, *E. siliculosus*, *P. ramorum*, *P. sojae*, and *E. huxleyi* genomes, respectively, accounting for about 1.3, 1.5, 1.8, 1.5, and 1% of their predicted proteomes (Table 2.1). Since all eukaryotic TF families were historically identified and characterized in plants, fungi or animals, these numbers are likely to be lower estimates and stramenopiles may have evolved some lineage-specific TFs that await identification. For comparison, "higher" or multicellular organisms such as *D. melanogaster* and *A. thaliana* dedicate 4.5% and 5.9% of their respective proteomes to TFs (Adams et al., 2000; Riechman et al., 2000). In contrast, about 2.1% and 0.9% of the proteome of unicellular eukaryotes such as *C. reinhardtii* and *S. cerevisisae*, respectively, can be attributed to TFs. Among the diatom TFs, 119 clear orthologous pairs were identified between the two diatoms.

2.2.2 Transcription factor complement among stramenopiles

One striking characteristic when comparing the diatom TF complement with other eukaryotes is the high abundance of heat shock factors (HSFs) in both the centric and pennate diatom genomes. By contrast, *S. cerevisiae* and *D. melanogaster* harbor only a single HSF (Sorger and Pelham, 1988). There are one, two and three HSFs in *O. tauri*, *C. reinhardtii* and *C. merolae*, respectively (different databases). Higher plants contain somewhat higher numbers of HSF genes, with *A. thaliana* and *Oryza sativa* possessing 21 (Nover et al., 2001) and 26 (Riechmann et al., 2000) distinct HSF genes, respectively. We identified as many as 70 and

94 putative HSF genes in the genomes of *P. tricornutum* and *T. pseudonana*, respectively, thus constituting ~34 and ~38% of the respective TF complements (Table 2.2). HSF expansion was not observed in other stramenopile genomes. Several diatom HSFs have orthologs in both species and many others have close homologs (Figure 2.1) suggesting that HSF expansion occurred in a diatom common ancestor and has been continuing after the separation of centrics and pennates. Such disproportionate expansion of a particular family of TFs is comparable to what has been observed for other TFs in other non-plant eukaryotic genomes, such as the nuclear hormone receptors (NHRs) which make up ~38% of the *C. elegans* TF complement (Riechmann et al., 2000).

Interestingly, we notice that Myb and C2H2-type zinc finger TFs are the most abundant families in all the other stramenopile genome investigated, as well as in the *E. huxleyi* genome (Table 2.2). We also observed a disproportionate contribution of AP2 factors in the TF complement of the haptophyte *E. huxleyi* (~20%) with respect to the general trend found in stramenopiles (Table 2.2) (note that this is a preliminary estimate since genome assembly is not yet fully completed for this species). Although the AP2 family was previously described as being plant-specific because it is absent from the genomes of opisthokonts and diplomonads (Riano-Pachon et al., 2008), we also found AP2-containing proteins in all stramenopile genomes examined. In order to get more insights into the nature of these key transcriptional regulators, we undertook a detailed analysis of some important TF families in diatoms.

Figure 2.1: Phylogenetic relationships between the DBD domain of diatom HSFs and other eukaryotes. *P. tricornutum* and *T. pseudonana* sequences are displayed with red and blue background, respectively. (A), (B), (C), and (D) show phylogenetic relationships between group 1, group 2, group 3, and group 4 diatom HSFs. The black dots indicate branches supported by bootstrap scores above 60% (over 1,000 iterations).

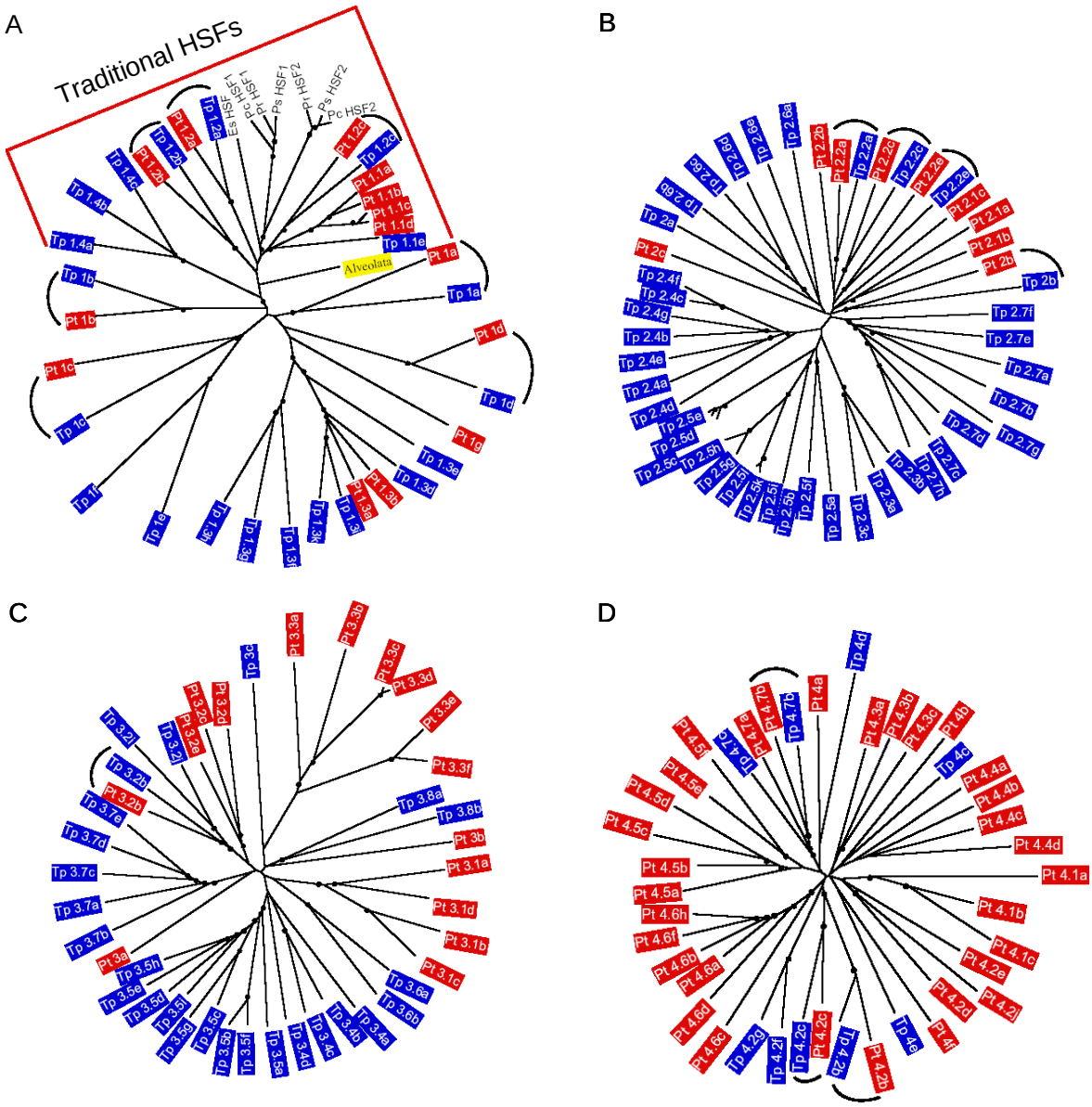


Table 2.1:

Family	Species						
	P. tricornutum	*T. pseudonana*	*A. anophagefferens*	*E. siliculosus*	*P.ramorum*	*P.sojae*	*E. huxleyi*
HSF	70	94	22	25	16	19	9
Myb	34	37	46	51	95	72	119
bzip	23	23	5	17	22	23	5
Zn_finger, C2H2-type	19	17	28	82	68	94	86
Zn_finger, CCCH-type	15	19	18	29	25	30	43
Zn-finger, TAZ-type	5	6	4	7	6	5	0
Zn_finger, GATA-type	0	0	0	0	1	2	5
bHLH	7	5	0	4	27	23	0
Fungal TRF	1	4	0	26	0	0	20
CCAAT-binding	1	0	0	9	2	1	1
Homeobox	4	9	0		7	6	20
AP2-EREBP	3	12	6	8	1	1	84
MADS	0	0	0		1	1	1
Nin-like	0	0	0	0	15	14	0
TAF9	1	1	1	1	1	1	2
E2F-DP	5	3	2	3	4	3	1
ABI3/VP1	0	0	0	0	0	0	0
WRKY	0	0	0	0	0	0	0
Zn_finger, DOF-type	0	0	0	0	0	0	0
CBF/NF-Y/archaeal	5	6	6	0	4	3	4
CXC/tesmin	5	5	3	4	3	3	2
Sigma-70 r2/r3/r4	8	8	6	17	0	0	4
Total	204	247	147	283	298	301	406

Table 2.2:

Family	Species						
	P. tricornutum	*T. pseudonana*	*A. anophagefferens*	*E. siliculosus*	*P.ramorum*	*P.sojae*	*E. huxleyi*
HSF	34,3	38,1	15,0	8,8	5,4	-	2,2
Myb	16,7	15,0	31,3	18,0	31,9	25,5	29,3
bzip	11,3	9,3	3,4	6,0	7,4	8,2	1,2
Zn_finger, C2H2-type	9,3	6,9	19,0	29,0	22,8	33,3	21,2
Zn_finger, CCCH-type	7,4	7,7	12,2	10,2	8,4	10,6	10,6
Zn-finger, TAZ-type	2,5	2,4	2,7	2,5	2,0	1,8	-
Zn_finger, GATA-type	-	-	-	-	0,3	0,7	1,2
bHLH	3,4	2,0	-	1,4	9,1	8,2	-
Fungal TRF	0,5	1,6	-	9,2	-	-	4,9
CCAAT-binding	0,5	-	-	3,2	0,7	0,4	0,2
Homeobox	2,0	3,6	-	-	2,3	2,1	4,9
AP2-EREBP	1,5	4,9	4,1	2,8	0,3	0,4	20,7
MADS	-	-	-	-	0,3	0,4	0,2
Nin-like	-	-	-	-	5,0	5,0	-
TAF9	0,5	0,4	0,7	0,4	0,3	0,4	0,5
E2F-DP	2,5	1,2	1,4	1,1	1,3	1,1	0,2
ABI3/VP1	-	-	-	-	-	-	-
WRKY	-	-	-	-	-	-	-
Zn_finger, DOF-type	-	-	-	-	-	-	-
CBF/NF-Y/archaeal	2,5	2,4	4,1	-	1,3	1,1	1,0
CXC/tesmin	2,5	2,0	2,0	1,4	1,0	1,1	0,5
Sigma-70 r2/r3/r4	3,9	3,2	4,1	6,0	-	-	1,0

2.2.3 Heat Shock Factors (HSFs)

Heat shock factors (HSFs) activate the expression of heat shock genes in response to different stresses (Morimoto et al., 1992). They bind specifically to heat shock promoter elements, which are palindromic sequences rich with repetitive purine and pyrimidine motifs. In vertebrates, heat shock factor 1 (HSF1) is a key regulator of heat shock protein transcription. In non stressful conditions, HSF1 is inhibited by association with heat shock proteins (HSPs). Protein misfolding in response to cellular stresses such as increased temperature leads to dissociation of HSPs from HSF-1, allowing its trimerization and activation as a transcription factor following translocation to the nucleus (Shamovsky and Nudler, 2008). While HSFs from different species share relatively little homology, the overall domain organization is conserved. HSFs have a modular structure: an N-terminal DNA binding domain (DBD) which is the most conserved, followed by two hydrophobic repeat domains (HR-A/B) and by a loosely defined regulatory domain which contains residues for post-transcriptional modifications. An additional hydrophobic repeat, HR-C, is situated further downstream and followed by a transactivation domain (Morimoto, 1998; Scharf et al., 1998).

We analyzed the DBD of all diatom HSFs in comparison with the DBDs of HSFs found in different species. Phylogenetic analysis identified four major groups of *P. tricornutum* and *T. pseudonana* HSFs (Figure 2.1). All diatom HSFs studied were named according to their phylogenetic position in this tree. Diatom HSF subgroups 1.1, 1.2 and 1.4 (7 from *P. tricornutum* and 7 from *T. pseudonana* in total), as well as some other stramenopile HSFs (7 in total) were found to be most closely related to HSFs from other species, including plants, fungi, and animals, and especially from alveolata HSFs (Figure 2.1). The remaining 63 and 87 *P. tricornutum* and *T. pseudonana* HSF DBDs, respectively, appear to diverge considerably, with only few orthologous pairs identified (Bowler et al., 2008).

We noticed that in all HSF DBDs studied, eight aromatic amino acids (numbered from A1 to A8) are regularly spaced over the entire DBD. In almost all HSF DBDs, the A1 and A3 to A8 residues correspond to Phe, whereas A2 corresponds mostly to Trp. The distance betweew diagnostic aromatic residues are, with some minor exceptions, well conserved among all HSF DBDs analyzed and correspond generally to 6-8 residues.

It is known that the activation of HSFs in response to different stress conditions involves the formation of homotrimers. Trimerization depends on several arrays of evolutionarily conserved, hydrophobic heptad repeats located next to the DBD (for reviews, see Wu, 1995; Morimoto, 1998; Schoffl et al., 1998). The heptad pattern of hydrophobic residues (HR-A/B region) suggests a coiled-coil structure similar to that reported for leucine-zipper-type protein interaction domains (Peteranderl et al., 1999). Surprisingly, among stramenopile HSFs we observed heptad repeats only in traditional HSFs. Among heptad repeat-containing diatom HSFs, two (Pt_1.2b and Tp_1.2b) contain only one heptad repeat, and two other HSFs, Pt_1.2a and Tp_1.2a, have two heptad repeats which are separated from a putative HR-B domain. The other traditional stramenopile HSFs have a compact HR-A/B domain with 4 (rarely 3) heptad repeats, and appear to be similar to non-plant HSFs and plant class B HSFs. It should also be noted that we have not found classical heptad repeats in any of the three *C. merolae* HSFs, even though there are some hydrophobic residues in the region homologous to heptad repeats of green algae. Two possible explanations for the absence of heptad repeats in the majority of stramenopile and *C. merolae* HSFs can be proposed: (*i*) conversion from a monomeric form of HSF to a trimer is triggered by other HSF elements, (*ii*) trimerization of HSFs is not necessary for their DNA-binding activity.

An important functional aspect of HSFs concerns the presence of activator domains (ADs) assumed to make (or to break) contacts with components of the general transcription

complex. In most HSFs from plants, vertebrates and insects, ADs were observed to be located near the C-terminus and to be enriched in hydrophobic and acidic residues (Wisniewski et al., 1996). The excess of hydrophobic and acidic residues suggests that interacting targets of such putative ADs are hydrophobic and basic, such as histones (Erkine and Gross, 2003). Many mammalian and yeast transcription factor ADs contain short peptide motifs with a characteristic composition of aromatic, hydrophobic, and acidic amino acid residues (AHA motifs, reviewed in Triezenberg, 1995).

In this work, C-terminal domains of all diatom HSFs were inspected for the presence of potential ADs. Analysis of traditional stramenopile HSFs showed that 13 of them have one or two putative AHA motifs. Some non-traditional diatom HSFs also contain putative AHA motifs. In these motifs, the proportion of aromatic/hydrophobic and acidic residues is very high and can reach 100%. Interestingly, several diatom putative AHA motifs are highly similar to those from other organisms. For example, the putative AHA motif of Tp_1.1e, ADLLVEDF, is almost identical to the AHA1 motif from rice HSFA2 (AELLVEDF, Kotak et al., 2004). The C-terminal domain of Tp_1.2a has no evident AHA sequences, however it contains a motif VETFMMEVI similar to the motif IDTFLMEML which represents the part of the AD of the mammalian retinoid receptors (Leng et al., 1995). We also found that Pt_1.2c and Tp_1.2c contain the C-terminal domain sequences TKNIDQLTA and TKNIDDLTS, respectively, which are very similar to the *A. thaliana* HSFA5 sequence TKNIEQLTL localized in the extreme end and containing the nuclear export signal (NES, Nover et al., 2001; Kotak et al., 2004).

Finally, we found that one *P. tricornutum* HSF (Pt_1c) consists of a fusion of the HSF domain with an AP2 domain. AP2 is a DNA-binding domain found in plant transcription factors such as APETALA2 and EREBP (ethylene responsive element binding protein) (identified by Kamel Jabbari). Interestingly, HSF and AP2 transcription factors have been

shown to interact and to synergistically activate the expression of ethylene-responsive genes in *Helianthus annuus* (sunflower) (Diaz-Martin et al., 2005). The existence and expression of this fusion was verified by PCR using *P. tricornutum* cDNA for template and it was not found in *T. pseudonana* nor in the other stramenopiles examined here. However, we identified a gene model in *E. huxleyi* that displays both the HSF and AP2 domains.

2.2.4 Myb factors

The Myb family of transcription factors is characterized by the presence of a conserved DBD domain which comprises up to three imperfect repeat sequences (Myb motifs) of approximately 50 amino acids. Each repeat forms an independent mini-domain with a helix-turn-helix structure containing three regularly spaced tryptophan (W) residues. In invertebrates, there is a single Myb protein which represents a so called Myb3R factor, i.e., it contains three Myb motifs which are referred to as R1, R2 and R3 and are marked by indicative tryptophan residues, from W1 to W3, from W4 to W6, and from W7 to W9, respectively. Vertebrate genomes possess three genes encoding Myb3R proteins (Nomura et al., 1988), and *Arabidopsis* and tobacco genomes contain five such genes (Ito, 2005; Yanhui et al., 2006). According to the model of evolution of Myb proteins proposed by Lipsick (1996), Myb3R proteins may have been generated by successive intragenic domain duplication events in primitive eukaryotes.

In contrast to animals, plant Myb3R proteins constitute only a small fraction of their Myb transcription factor complement. Indeed, the *Arabidopsis* genome contains in total around 200 Myb genes, categorized according to the number of adjacent repeats in the Myb domain and the similarity of the latter to either R1, R2 or R3. Myb proteins containing two repeats (Myb2R) constitute the largest Myb family in plants (126, Yanhui et al., 2006). The Lipsick model of evolution of Myb proteins (1996) suggests that the Myb2R proteins arose

after loss of the sequences encoding R1 in an ancestral *Myb3R* gene. These proteins were reported to be involved in many physiological and biochemical processes including cell cycle control. Some are also involved in various defense and stress responses and in light and hormone signaling (see Yanhui et al., 2006). The conserved CDC5 (Cell Division Control protein 5) family, which does not function as a TF, also contains two repeats of the Myb domain which are only distantly related to those of the R2R3-type Myb domain (Stracke et al., 2001).

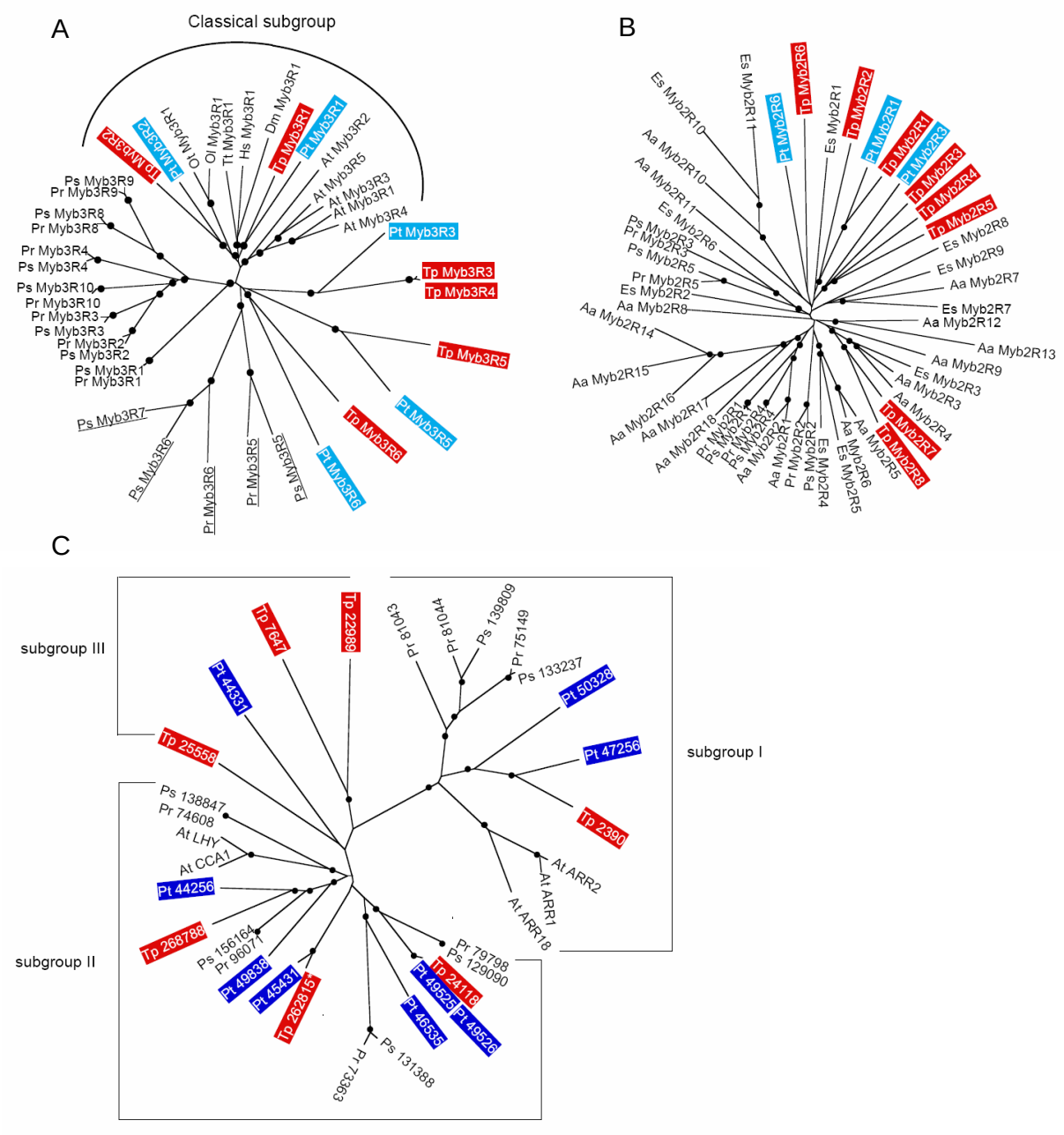
Other plant Myb proteins contain a single Myb domain (Myb1R), or sometimes just a partial one. They are fairly divergent and include factors that bind the consensus sequence of telomeric DNA (Yu et al., 2000). It has also been shown that Myb1R factors can act as transcriptional activators (Baranowskij et al., 1994). Among plant Myb1R proteins are two key clock genes, *CCA1* and *LHY*, containing a typical SHAQKYF motif within the Myb repeat (Wang et al., 1997). In *Arabidopsis*, the CCA1-like subfamily is composed of 25 proteins forming two main clades (Yanhui et al., 2006). Numerous other Myb-related proteins have also been identified in *Arabidopsis* (Yanhui et al., 2006). One of them is referred to as Myb4R because of the presence of four Myb motifs.

In diatoms, we identified a total of 34 *P. tricornutum* and 37 *T. pseudonana* Myb proteins, as well as 15 *P. tricornutum* and *T. pseudonana* proteins containing Myb domains only weakly similar to Myb proteins that were not considered further. They belong to five types: Myb3R, Myb2R, CDC5-like, Myb1R, and Myb-related proteins. *P. tricornutum* and *T. pseudonana* have 5 and 6 Myb3R proteins, respectively. They form 5 orthologous pairs, and one of the *P. tricornutum* Myb3R proteins (Pt_Myb3R3) has two almost identical *T. pseudonana* orthologs (Tp_Myb3R3 and Tp_Myb3R4) which clearly result from a recent duplication (Figure 2.2A). This observation is confirmed by the fact that the Tp_Myb3R3 and Tp_Myb3R4 genes are closely linked and are located on the same DNA strand. As shown in

Fig. 2a, two orthologous diatom pairs (Pt_Myb3R1/Tp_Myb3R1 and Pt_Myb3R2/Tp_Myb3R2) were found to be similar to Myb proteins from different species and were denoted a «classical» subgroup. It should be noted that *P. sojae* (Figure 2.2A), as well as other representatives of stramenopiles, *A. anophagefferens* and *E. siliculosus* (not shown), do not to have any orthologs of the «classical» Myb3R proteins.

Among stramenopile Myb2R proteins, we found eight *T. pseudonana* and only three *P. tricornutum* proteins. Phylogenetic analysis did not reveal significant similarity between them and Myb2R proteins of other stramenopile species. Pt and Tp Myb2R sequences seem to be diatom-specific, with only two *T. pseudonana* proteins (Tp_Myb2R7 and Tp_Myb2R8) being grouped with one *E. siliculosus* and two *A. anophagefferens* Myb2R sequences (Figure 2.2B). Comparison of stramenopile Myb2R and Myb3R amino acid sequences showed that the two Myb motifs of Myb2R (which start with W1 and W4, respectively) correspond to the second and third Myb motifs of Myb3R (which start with W4 and W7, respectively). This is in accordance with the Lipsick model of evolution of Myb proteins (1996) which suggests that the Myb2R proteins arose after loss of the sequences encoding R1 in an ancestral *Myb3R* gene. This conclusion is strengthened by the observation that the same intron positions are often found in the *Myb* regions corresponding to the R1 of Myb2R and to the R2 of Myb3R, and/or in those corresponding to the R2 of Myb2R and to the R3 of Myb3R (see Figure 2.2).

Figure 2.2: Phylogenetic relationships between stramenopile Myb3R (A), Myb2R (B), Myb1R (C). *P. tricornutum* and *T. pseudonana* sequences are displayed with red and blue background, respectively. The black dots indicate branches supported by bootstrap scores above 60% (over 1,000 iterations).



We also observed 9 *P. tricornutum* and 7 *T. pseudonana* Myb proteins which contain a single Myb domain (Myb1R proteins). Among these, some belong to the SHAQKYF-like family described in plants, including *CCA1* and *LHY* (Schaffer et al., 2001; Yanhui et al., 2006). Most appear to have a THAQKYF signature rather than SHAQKYF. These diatom Myb1R proteins can be divided into three subgroups (Figure 2.2C). Subgroup I includes two Pt and one Tp Myb1R proteins, as well as some Myb1R *Phytophthora* proteins, which seem to be related to *Arabidopsis* ARR (*Arabidopsis* response regulator) proteins. Subgroup II is weakly related to *Arabidopsis* proteins *CCA1* and *LHY*, although a clear orthology was not observed. Subgroup III appears to be diatom-specific (Fig. 2c). It should also be noted that *Phytophthora* SHAQKYF-like proteins are considerably more numerous than those of diatoms: 33 in *P. ramorum* and 35 in *P. sojae*, most of which also display a THAQKYF signature.

In this work, we also re-investigated the origins of the two Myb motifs found in the CDC5 family. The CDC5 protein was first described in *S. pombe* (Ohi et al., 1994) as being essential for progression of the G2 phase of the cell cycle. It was later demonstrated that a single copy of the CDC5-like protein is structurally and functionally conserved in all eukaryotes. The Arabidopsis *CDC5* gene was found to be constitutively expressed in all organs and during all conditions tested (Yanhui et al., 2006). The CDC5 protein was called "R2R3-like protein" (Yanhui et al., 2006). Here, detailed comparison of the amino acid sequences of CDC5 and Myb3R proteins in numerous species led us to conclude that whereas the second MYB motif of CDC5s does correspond to the R3 of Myb3R proteins, their first MYB motif corresponds not to the R2, but rather to the R1/R2 hybrid motif of Myb3R sequences. This correspondence is confirmed by the finding that in different species, including all higher and lower animals, as well as *Chlamydomonas*, *Volvox*, *Bigelowiella* and

S. pombe, the region of the *CDC5* gene encoding the R1/R2 motif has the intron *i*1 in exactly the same position as in numerous other *Myb3R* genes encoding the R1 motif.

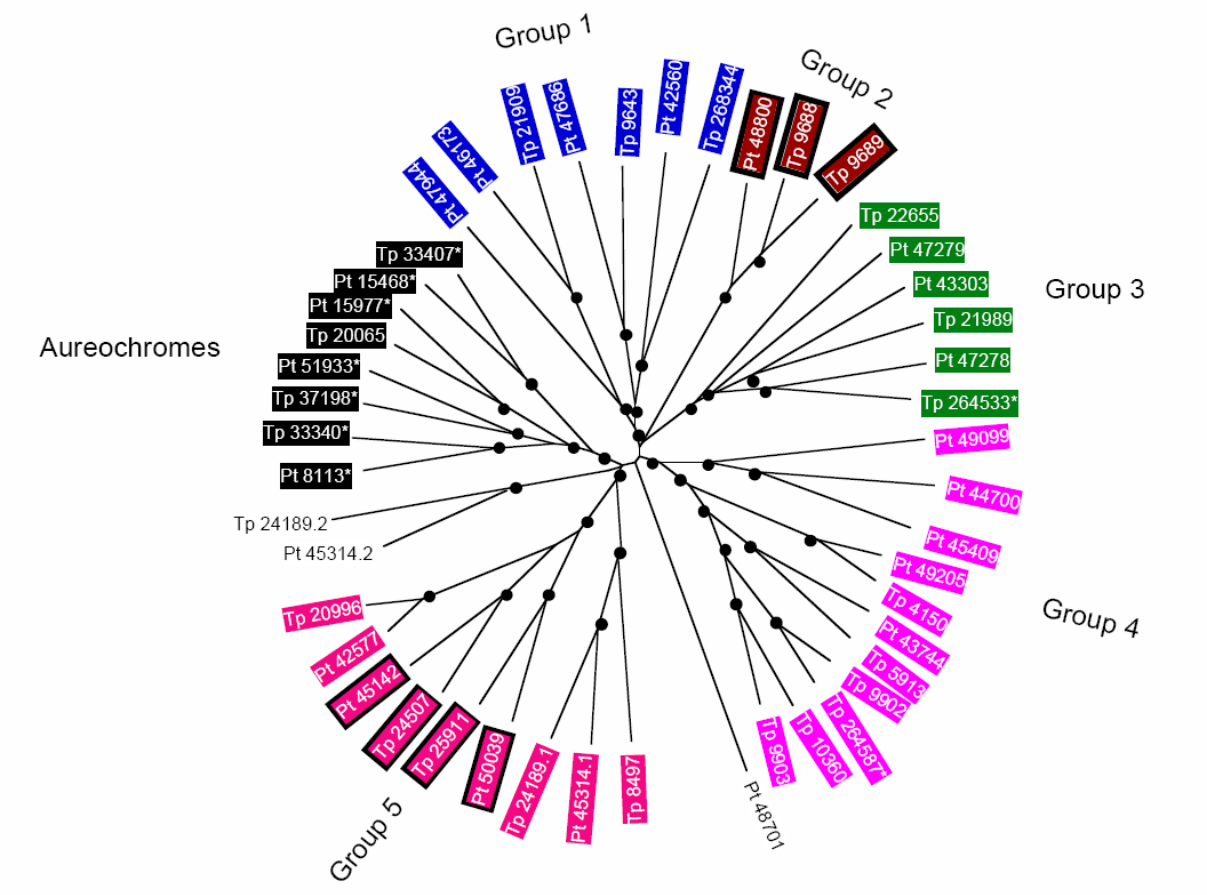
2.2.5 bZIP domain factors

The eukaryotic basic-leucine zipper (bZIP) transcription factors play critical roles in a myriad of cellular functions in organisms from different phylogenetic kingdoms. In particular, they regulate numerous response pathways in plants, such as light and stress signaling. The family is represented by multiple genes, encoding proteins that contain a basic region involved in DNA binding and nuclear import, and a leucine zipper dimerization domain (Hurst, 1995). The bZIP transcription factors in Arabidopsis (approximately 75 members) are classified into ten groups (Jacoby et al., 2002).

We identified 22 *P. tricornutum* and 23 *T. pseudonana* bZIP proteins, respectively, which were subdivided into six major phylogenetic groups, members of all groups being well resolved (Figure 2.3). Comparative phylogenetic analysis with representatives from plants, algae, and animals showed that there was no significant homology between diatom bZIP proteins and those of other species (not shown). On the other hand, some *P. tricornutum* and *T. pseudonana* bZIP proteins are similar to those of *A. anophagefferens*, *P. sojae* and *E. siliculosus*, indicating that they are stramenopile-specific.

One orthologous diatom pair (bZIP14) contains not one, but two bZIP domains. We found similar proteins in *Phytophthora* and *Ectocarpus*. In all of them, the first (5') and the second (3') bZIP domains are separated by approximately 100 amino acids. All 5' and 3' bZIP domains form two different subgroups, respectively (Supplementary Figure 1), indicating that the bZIP domain precursor duplicated before the divergence of stramenopiles, and that the two duplicated copies were transferred to diverse species.

Figure 2.3: Phylogenetic relashionships between the diatom bzip (IPR004827) domains. The black dots indicate branches supported by bootstrap scores above 60% (over 1,000 iterations).

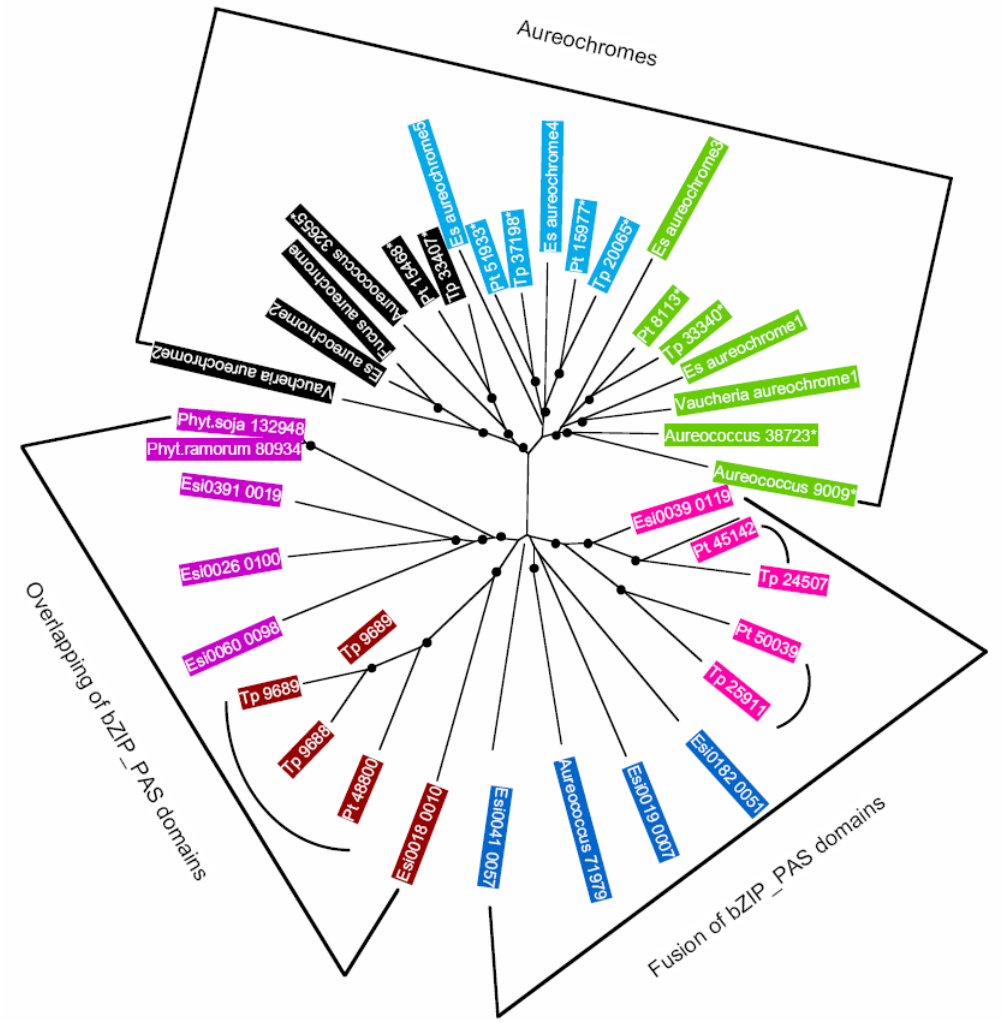


A new class of bZIP factors was recently described in the stramenopile algae *Vaucheria frigida* (Xanthophyceae) and *Fucus distichus* (Phaeophyceae) and was found to function as a blue light photoreceptor. These proteins were denoted aureochromes and contain a bZIP domain together with a PAS-related LOV (Light, Oxygen, or Voltage) domain (Takahashi et al., 2007), which is also found in the phototropin family of plant blue-light photoreceptors (see Crosson et al. 2003). We found four putative orthologous *P. tricornutum* and *T. pseudonana* protein pairs which appear to represent aureochromes (Figures 2.3 and 2.4).

Moreover, similar proteins are found in *A. anophagefferens* and *E. siliculosus*. Thus, aureochromes were found in all photosynthetic stramenopiles for which data is available. Three aureochrome subgroups were observed in stramenopiles.

In two other orthologous *P. tricornutum* and *T. pseudonana* bZIP protein pairs, we found a fusion of a putative bZIP domain with a PAS domain, which is related to the LOV domain. This PAS domain is rather similar to the photoreactive yellow protein (PYP) sensor domain, which is considered as a blue light photoreceptor for a negative phototactic response in bacteria (Borgstahl et al., 1995). This class of fusion protein may also have a cellular function involved in light perception and signal transduction, and similar proteins can be found in *A. anophagefferens* and *E. siliculosus*. Among these bZIP_PAS proteins, two subgroups can be distinguished. On the other hand, in some stramenopile proteins, putative bZIP and PAS domains overlap and form two subgroups. Together these results suggest that stramenopiles may have evolved two new classes of photoreceptors with the bZIP_PAS and bZIP_LOV fusions which both contain putative light-sensitive and DNA-binding domains. This is of particular importance considering that the stramenopiles found to retain these fusions are all photosynthetic organisms for which accurate light perception is likely to be critical. The importance of the aureochrome family as blue-light photoreceptors in marine stramenopiles is highlighted by the fact that the marine light spectrum is dominated by blue light wavelengths.

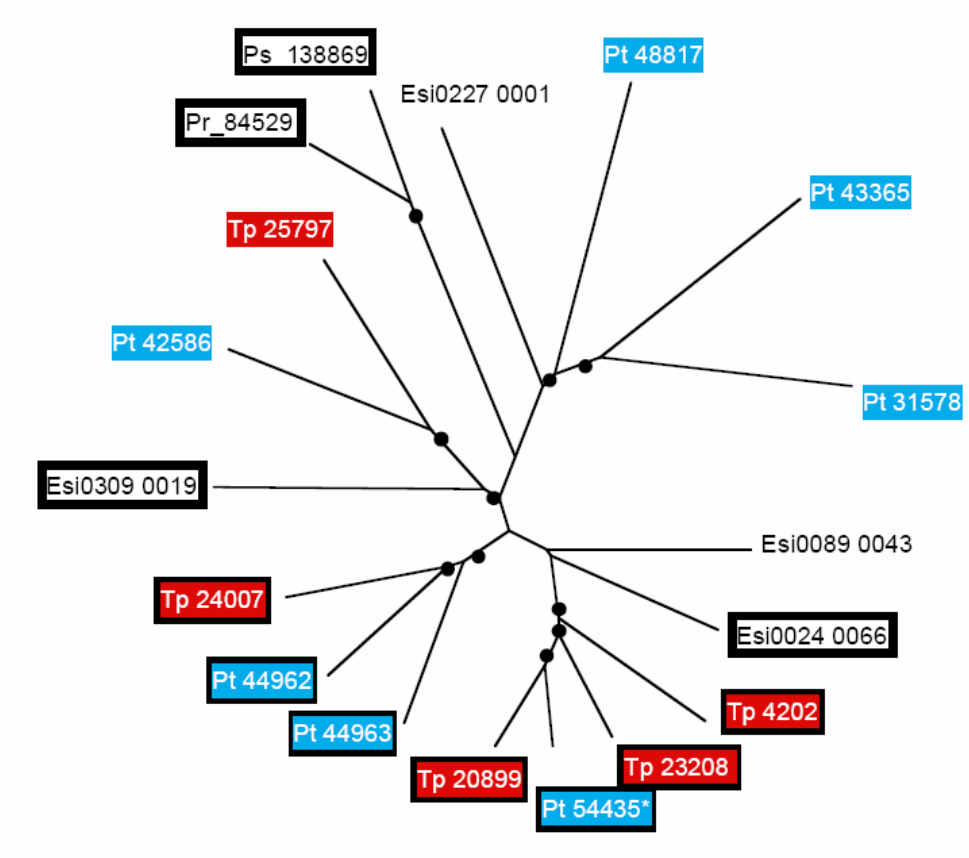
Figure 2.4: Phylogenetic relashionships between the bzip domains of proteins containing bZIP and PAS domains. The three subgroups of aureochromes are shown in a black, blue and green background, the two subgroups of bZIP_PAS fusions are shown in an azure and rose background, and the two subgroups of overlapping bZIP_PAS proteins are shown in a purple and chocolate background. The black dots indicate branches supported by bootstrap scores above 60% (over 1,000 iterations).



2.2.6 bHLH factors

The helix-loop-helix (HLH) proteins are a group of eukaryotic transcription factors playing pivotal roles during development of multicellular organisms. One HLH protein family, the basic helix-loop-helix (bHLH) proteins, possess an additional domain of basic amino acids spanning 12-14 residues, adjacent to the N-terminus of the HLH domain, which determines DNA binding specificity (Littlewood and Evan, 1995). bHLH proteins can be classified into discrete classes. We found 9 *P. tricornutum* and 5 *T. pseudonana* bHLH proteins. Half of them, as well as four *Phytophthora* and *Ectocarpus* bHLH proteins also contain a PAS domain (Figure 2.5). Such a fusion has been documented before in one of the bHLH family members of vertebrates (Kewley et al., 2004). However, whereas all the 11 stramenopile bHLH_PAS proteins contain only one PAS domain, the majority of animal bHLH_PAS proteins contain two PAS domains forming two different subgroups and their patchy distribution among eukaryotic supergroups suggests that these domains fused at least twice in the course of evolution.

Figure 2.5: Phylogenetic relashionships between the bHLH (IPR001092) domains from a selection of stramenopile bHLH factors. The black dots indicate branches supported by bootstrap scores above 60% (over 1,000 iterations).

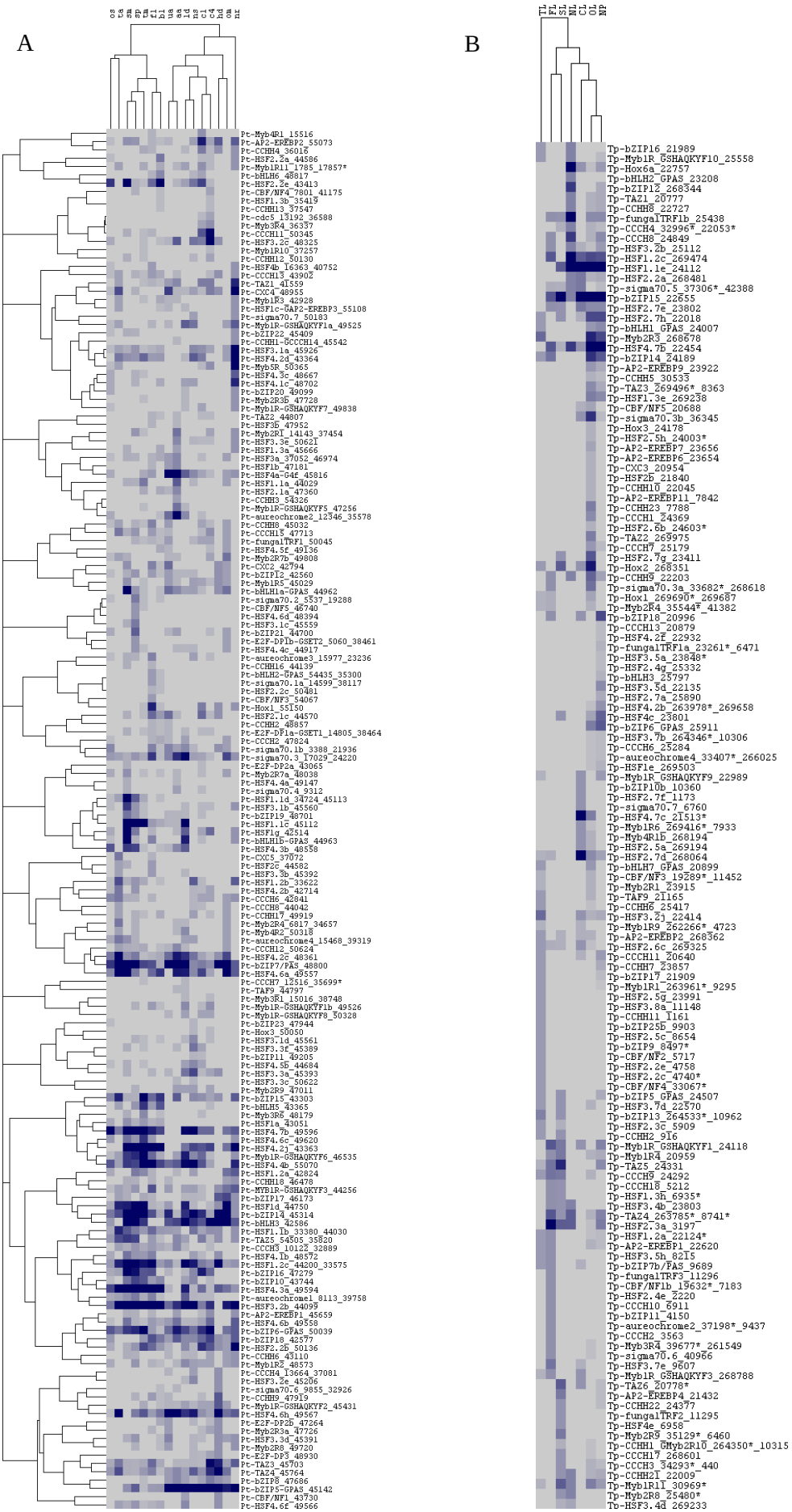


2.2.7 Expression analysis

We analyzed the expression pattern of the diatom TFs identified in this study across the 16 and 7 *P. tricornutum* and *T. pseudonana* EST libraries, respectively (Maheswari et al., 2008). After normalizing transcript abundance in each individual library to library size, we made a hierarchical clustering (Eisen et al., 1998) of the *P. tricornutum* and *T. pseudonana* TFs to identify groups of genes with similar expression pattern and the libraries with similar gene expression pattern. This method was used to identify libraries which share similar patterns of expression with reference to the presence or absence of a transcript and its relative abundance. Figure 2.6 shows the results visualized using Java Treeview (Saldanha, 2004).

We observe first of all that the expression pattern of TFs mirrors the overall similarities of expression patterns of all expressed genes in the different libraries (e.g, the similarities observed between the *P. tricornutum* "blue light" and "low iron"libraries) (Bowler et al., 2008), showing that this subset of genes is representative of the global gene expression pattern from the respective diatoms. We observe in both diatoms that different families of TFs show similar expression patterns, which may reflect biological interactions or cross-talk between different TFs in response to a particular condition. We also notice that the expression of some TFs is rather constitutive while others appear to be expressed more specifically, especially in *T. pseudonana*.

Figure 2.6 (next page): Hierarchical clustering showing the expression pattern of *P. tricornutum* TFs transcripts in 16 different growth conditions (A) and *T. pseudonana* TFs transcripts in 7 different growth conditions. Abbreviations of library names: original standard, (os), low decadienal (ld), silica minus (sm), silica plus (sp), triradiate morphotype (tm), oval morphotype (om), tropical accessions (ta), low iron (fl), blue light (bl), urea adapted (ua), ammonium adapted (aa), high decadienal (hd), nitrate starvation (ns), nitrate replete (nr), high CO2 for 1 day (c1), high CO2 for 4 days (c4).


A
B
Pt-Myb4R1_15516
Pt-AP2-EREBP2_55073
Pt-CCHH4_36016
Pt-HSF2.2a_44586
Pt-MybR1R11_17785_17857*
Pt-bHLH6_48817
Pt-HSF2.2e_43413
Pt-CBF/NF4_7801_41175
Pt-HSF1.3b_35415
Pt-CCHH13_37547
Pt-cdc5_13192_36588
Pt-Myb3R4_36337
Pt-CCCH11_50345
Pt-HSF3.2c_48325
Pt-MybR1R10_37257
Pt-CCHH12_50130
Pt-HSF4b_16363_40752
Pt-CCCH13_43902
Pt-TAZ1_41559
Pt-CXC4_48955
Pt-MybR3_42928
Pt-HSF1c-bAP2-EREBP3_55108
Pt-sigma70.7_50183
Pt-MybR-GSHAQKYF1a_49525
Pt-bZIP22_45409
Pt-CCHH1-bCCCH14_45542
Pt-HSF3.1a_45926
Pt-HSF4.2d_43364
Pt-Myb5R_50365
Pt-HSF4.3c_48667
Pt-HSF4.1c_48702
Pt-bZIP20_49099
Pt-Myb2R3b_47728
Pt-MybIR-GSHAQKYF7_49838
Pt-TAZ2_44807
Pt-HSF3b_47952
Pt-Myb2R1_14143_37454
Pt-HSF3.3e_50621
Pt-HSF1.3a_45666
Pt-HSF3a_37052_46974
Pt-HSF1b_47181
Pt-HSF4a-G4f_45816
Pt-HSF1.1a_44029
Pt-HSF2.1a_47360
Pt-CCHH3_54326
Pt-MybIR-GSHAQKYF5_47256
Pt-aureochrome2_12346_35578
Pt-CCHH8_45032
Pt-CCCH15_47713
Pt-fungalTRF1_50045
Pt-HSF4.5f_49136
Pt-Myb2R7b_49808
Pt-CXC2_42794
Pt-bZIP12_42560
Pt-MybIR5_45029
Pt-bHLH1a-GPA3_44962
Pt-sigma70.2_5537_19288
Pt-CBF/NF5_46740
Pt-HSF4.6d_48394
Pt-HSF3.1c_45559
Pt-bZIP21_44700
Pt-E2F-DP1b-GSET2_5060_38461
Pt-HSF4.4c_44917
Pt-aureochrome3_15977_23236
Pt-CCHH16_44139
Pt-bHLH2-GPA3_54435_35300
Pt-sigma70.1a_14599_38117
Pt-HSF2.2c_50481
Pt-CBF/NF3_54067
Pt-Hox1_55150
Pt-HSF2.1c_44570
Pt-CCHH2_48857
Pt-E2F-DP1a-GSET1_14805_38464
Pt-CCCH2_47824
Pt-sigma70.1b_3388_21936
Pt-sigma70.3_17029_24220
Pt-E2F-DP2a_43065
Pt-Myb2R7a_48038
Pt-HSF4.4a_49147
Pt-sigma70.4_9312
Pt-HSF1.1d_34724_45113
Pt-HSF3.1b_45560
Pt-bZIP19_48701
Pt-HSF1.1c_45112
Pt-HSF1g_42514
Pt-bHLH1b-GPA3_44963
Pt-HSF4.3b_48558
Pt-CXC5_37072
Pt-HSF2c_44582
Pt-HSF3.3b_45392
Pt-HSF1.2b_33622
Pt-HSF4.2b_42714
Pt-CCCH6_42841
Pt-CCCH8_44042
Pt-CCHH17_49919
Pt-Myb2R4_6817_34657
Pt-Myb4R2_50318
Pt-aureochrome4_15468_39319
Pt-CCCH12_50624
Pt-HSF4.2c_48361
Pt-bZIP7/PA3_48800
Pt-HSF4.6a_49557
Pt-CCCH7_12516_35699*
Pt-TAF9_44797
Pt-Myb3R1_15016_38748
Pt-MybIR-GSHAQKYF1b_49526
Pt-MybIR-GSHAQKYF8_50328
Pt-bZIP23_47944
Pt-Hox3_50050
Pt-HSF3.1d_45561
Pt-HSF3.3f_45389
Pt-bZIP11_49205
Pt-HSF4.5b_44684
Pt-HSF3.3a_45393
Pt-HSF3.3c_50622
Pt-Myb2R9_47011
Pt-bZIP15_43303
Pt-bHLH5_43365
Pt-Myb3R5_48179
Pt-HSF1a_43051
Pt-HSF4.7b_49596
Pt-HSF4.6c_49620
Pt-HSF4.2j_43363
Pt-MybIR-GSHAQKYF6_46535
Pt-HSF4.4b_55070
Pt-HSF1.2a_42824
Pt-CCHH18_46478
Pt-MYB1R-GSHAQKYF3_44256
Pt-bZIP17_46173
Pt-HSF1d_44750
Pt-bZIP14_45314
Pt-bHLH3_42586
Pt-HSF1.1b_33380_44030
Pt-TAZ5_54505_35820
Pt-CCCH3_10122_32889
Pt-HSF4.1b_48572
Pt-HSF1.2c_44200_33575
Pt-bZIP16_47279
Pt-bZIP10_43744
Pt-HSF4.3a_49594
Pt-aureochrome1_8113_39758
Pt-HSF3.2b_44099
Pt-AP2-EREBP1_45659
Pt-HSF4.6b_49558
Pt-bZIP6-GPA3_50039
Pt-bZIP18_42577
Pt-HSF2.2b_50136
Pt-CCHH6_43110
Pt-MybIR2_48573
Pt-CCHH4_13664_37081
Pt-HSF3.2e_45206
Pt-sigma70.6_9855_32926
Pt-CCHH9_47919
Pt-MybIR-GSHAQKYF2_45431
Pt-HSF4.6b_49567
Pt-E2F-DP2b_47264
Pt-Myb2R3a_47726
Pt-HSF3.3d_45391
Pt-Myb2R8_49720
Pt-E2F-DP3_48930
Pt-TAZ3_45703
Pt-TAZ4_45764
Pt-bZIP8_47686
Pt-bZIP5-GPA3_45142
Pt-CBF/NF1_43730
Pt-HSF4.6f_49566
TL
FL
SL
NL
CL
OL
NP
Tp-bZIP16_21989
Tp-MybIR_GSHAQKYF10_25558
Tp-Hox6a_22757
Tp-bHLH2_GPA3_23208
Tp-bZIP12_268344
Tp-TAZ1_20777
Tp-CCHH8_22727
Tp-fungalTRF1b_25438
Tp-CCCH4_32996*_22053*
Tp-CCCH8_24849
Tp-HSF3.2b_25112
Tp-HSF1.2c_269474
Tp-HSF1.1e_24112
Tp-HSF2.2a_268481
Tp-sigma70.5_37306*_42388
Tp-bZIP15_22655
Tp-HSF2.7e_23802
Tp-HSF2.7h_22018
Tp-bHLH1_GPA3_24007
Tp-Myb2R3_268678
Tp-HSF4.7b_22454
Tp-bZIP14_24189
Tp-AP2-EREBP9_23922
Tp-CCHH5_30533
Tp-TAZ3_269496*_8363
Tp-HSF1.3e_269238
Tp-CBF/NF5_20688
Tp-sigma70.3b_36345
Tp-Hox3_24178
Tp-HSF2.5h_24003*
Tp-AP2-EREBP7_23656
Tp-AP2-EREBP6_23654
Tp-CXC3_20954
Tp-HSF2b_21840
Tp-CCHH10_22045
Tp-AP2-EREBP11_7842
Tp-CCHH23_7788
Tp-CCCH1_24369
Tp-HSF2.6b_24603*
Tp-TAZ2_269975
Tp-CCCH7_25179
Tp-HSF2.7g_23411
Tp-Hox2_268351
Tp-CCHH9_22203
Tp-sigma70.3a_33682*_268618
Tp-Hox1_269690*_269687
Tp-Myb2R4_35544*_41382
Tp-bZIP18_20996
Tp-CCCH13_20879
Tp-HSF4.2f_22932
Tp-fungalTRF1a_23261*_6471
Tp-HSF3.5a_23848*
Tp-HSF2.4g_25332
Tp-bHLH3_25797
Tp-HSF3.5d_22135
Tp-HSF2.7a_25890
Tp-HSF4.2b_263978*_269658
Tp-HSF4c_23801
Tp-bZIP6_GPA3_25911
Tp-HSF3.7b_264346*_10306
Tp-CCCH6_25284
Tp-aureochrome4_33407*_266025
Tp-HSF1e_269503
Tp-MybIR_GSHAQKYF9_22989
Tp-bZIP10b_10360
Tp-HSF2.7f_1173
Tp-sigma70.7_6760
Tp-HSF4.7c_21513*
Tp-MybIR6_269416*_7933
Tp-Myb4R1b_268194
Tp-HSF2.5a_269194
Tp-HSF2.7d_268064
Tp-bHLH7_GPA3_20899
Tp-CBF/NF3_19289*_11452
Tp-Myb2R1_23915
Tp-TAF9_21165
Tp-CCHH6_25417
Tp-HSF3.2j_22414
Tp-MybIR9_262266*_4723
Tp-AP2-EREBP2_268362
Tp-HSF2.6c_269325
Tp-CCCH11_20640
Tp-CCHH7_23857
Tp-bZIP17_21909
Tp-MybIR1_263961*_9295
Tp-HSF2.5g_23991
Tp-HSF3.8a_11148
Tp-CCHH11_1161
Tp-bZIP25b_9903
Tp-HSF2.5c_8654
Tp-bZIP9_84977*
Tp-CBF/NF2_5717
Tp-HSF2.2e_4758
Tp-HSF2.2c_4740*
Tp-CBF/NF4_33067*
Tp-bZIP5_GPA3_24507
Tp-HSF3.7d_22570
Tp-bZIP13_264533*_10962
Tp-HSF2.5c_5909
Tp-CCHH2_916
Tp-MybIR_GSHAQKYF1_24118
Tp-MybIR4_20959
Tp-TAZ5_24331
Tp-CCCH9_24292
Tp-CCCH18_5212
Tp-HSF1.3h_6935*
Tp-HSF3.4b_23803
Tp-TAZ4_263785*_8741*
Tp-HSF2.3a_3197
Tp-HSF1.2a_22124*
Tp-AP2-EREBP1_22620
Tp-HSF3.5h_8215
Tp-bZIP7b/PA3_9689
Tp-fungalTRF3_11296
Tp-CBF/NF1b_19632*_7183
Tp-HSF2.4e_2220
Tp-CCHH10_5911
Tp-bZIP11_4150
Tp-aureochrome2_37198*_9437
Tp-CCCH2_3563
Tp-Myb3R4_39677*_261549
Tp-sigma70.6_40966
Tp-HSF3.7e_9507
Tp-MybIR_GSHAQKYF3_268788
Tp-TAZ6_20778*
Tp-AP2-EREBP4_21432
Tp-CCHH22_24377
Tp-fungalTRF2_11295
Tp-HSF4e_6958
Tp-Myb2R9_35129*_6460
Tp-CCHH1_GMyb2R10_264350*_10315
Tp-CCCH17_268601
Tp-CCCH3_34293*_440
Tp-CCHH21_22009
Tp-MybIR11_30969*
Tp-Myb2R8_25480*
Tp-HSF3.4d_269233

2.3 Materials and methods:

Diatom genomes were screened for the presence of gene models with similarity to DNA binding domains typical of the transcription factor families listed in Table 2.1. This computational analysis was performed using BLAST Analysis of domain composition of the candidate proteins was done by submitting their amino acid sequences to InterProScan (http://www.ebi.ac.uk/InterProScan/index.html). Multiple protein alignments were done using CLUSTALW followed by manual curation. The phylogenetic trees were constructed with the Neighbor Joining method using the CLUSTALX program (Thomson et al., 1997) and displayed using TreeView.

2.4 References:

Adams MD, Celniker SE, Holt RA, Evans CA, Gocayne JD, Amanatides PG, Scherer SE, Li PW, Hoskins RA, Galle RF, et al: The genome sequence of Drosophila melanogaster. *Science* 2000, 287:2185-2195.

Ahn SG, Liu PC, Klyachko K, Morimoto RI, Thiele DJ. The loop domain of heat shock transcription factor 1 dictates DNA-binding specificity and responses to heat stress. *Genes Dev*. 2001, 15(16):2134-45.

Armbrust EV, Berges JA, Bowler C, Green BR, Martinez D, Putnam NH, Zhou S, Allen AE, Apt KE, Bechner M, et al: The genome of the diatom Thalassiosira pseudonana: ecology, evolution, and metabolism. *Science* 2004, 306:79-86.

Baranowskij N, Frohberg C, Prat S, Willmitzer L. A novel DNA binding protein with homology to Myb oncoproteins containing only one repeat can function as a transcriptional activator. *EMBO J*. 1994, 13(22):5383-92.

Borgstahl GE, Williams DR, Getzoff ED: 1.4 A structure of photoactive yellow protein, a cytosolic photoreceptor: unusual fold, active site, and chromophore. *Biochemistry* 1995, 34:6278-6287.

Bowler C, Allen AE, Badger JH, Grimwood J, Jabbari K, Kuo A, Maheswari U, Martens C, Maumus F, Otillar RP, et al: The Phaeodactylum genome reveals the evolutionary history of diatom genomes. *Nature* 2008, 456:239-244.

Cicero MP, Hubl ST, Harrison CJ, Littlefield O, Hardy JA, Nelson HC. The wing in yeast heat shock transcription factor (HSF) DNA-binding domain is required for full activity. *Nucleic Acids Res*. 2001, 29(8):1715-23.

Crosson S, Rajagopal S, Moffat K. The LOV domain family: photoresponsive signaling modules coupled to diverse output domains. *Biochemistry* 2003, 42(1):2-10.

Diaz-Martin J, Almoguera C, Prieto-Dapena P, Espinosa JM, Jordano J: Functional interaction between two transcription factors involved in the developmental regulation of a small heat stress protein gene promoter. *Plant Physiol* 2005, 139:1483-1494.

Eisen MB, Spellman PT, Brown PO, Botstein D. Cluster analysis and display of genome-wide expression patterns. *Proc Natl Acad Sci U S A*. 1998 Dec 8;95(25):14863-8.

Erkine AM, Gross DS: Dynamic chromatin alterations triggered by natural and synthetic activation domains. *J Biol Chem* 2003, 278:7755-7764.

Field CB, Behrenfeld MJ, Randerson JT, Falkowski P: Primary production of the biosphere: integrating terrestrial and oceanic components. *Science* 1998, 281:237-240.

Hurst H.C. Transcription factors 1: bZIP proteins. *Protein Profile* 1995, 2 101-168.

Ito M. Conservation and diversification of three-repeat Myb transcription factors in plants. Review. *J Plant Res*. 2005, 118(1):61-69.

Jakoby M, Weisshaar B, Dröge-Laser W, Vicente-Carbajosa J, Tiedemann J, Kroj T, Parcy F. bZIP transcription factors in Arabidopsis. *Trends Plant Sci*. 2002, 7(3):106-11.

Kewley RJ, Whitelaw ML, Chapman-Smith A: The mammalian basic helix-loop-helix/PAS family of transcriptional regulators. *Int J Biochem Cell Biol* 2004, 36:189-204.

Kotak S, Port M, Ganguli A, Bicker F, von Koskull-Doring P. Characterization of C-terminal domains of Arabidopsis heat stress transcription factors (Hsfs) and identification of a new signature combination of plant class A Hsfs with AHA and NES motifs essential for activator function and intracellular localization. *Plant J*. 2004, 39(1):98-112.

Leng X, Blanco J, Tsai SY, Ozato K, O'Malley BW, Tsai MJ. Mouse retinoid X receptor contains a separable ligand-binding and transactivation domain in its E region. *Mol Cell Biol*. 1995, 15(1):255-63.

Levine M, Tjian R: Transcription regulation and animal diversity. *Nature* 2003, 424:147-151. Lipsick JS. One billion years of Myb. Review. *Oncogene* 1996, 13(2):223-235.

Littlewood TD, Evan GI: Transcription factors 2: helix-loop-helix. *Protein Profile* 1995, 2:621-702.

Liu TD and Marzluf GA. Characterization of pco-1, a newly identified gene which regulates purine catabolism in Neurospora. *Curr. Genet*. 2004, 46(4):213-27.

Maheswari U, Mock T, Armbrust EV, Bowler C: Update of the Diatom EST Database: a new tool for digital transcriptomics. *Nucleic Acids Res* 2008.

Morimoto RI, Sarge KD, Abravaya K. Transcriptional regulation of heat shock genes. A paradigm for inducible genomic responses. Review. *J Biol Chem*. 1992, 267(31):21987-90.

Morimoto RI. Regulation of the heat shock transcriptional response: cross talk between a family of heat shock factors, molecular chaperones, and negative regulators. *Genes Dev*. 1998, 12(24):3788-96.

Nomura N, Takahashi M, Matsui M, Ishii S, Date T, Sasamoto S, Ishizaki R. Isolation of human cDNA clones of myb-related genes, A-myb and B-myb. *Nucleic Acids Res*. 1988, 16(23):11075-89.

Nover L, Bharti K, Doring P, Mishra SK, Ganguli A, Scharf KD. Arabidopsis and the heat stress transcription factor world: how many heat stress transcription factors do we need? *Cell Stress Chaperones* 2001, 6(3):177-89.

Ohi R, McCollum D, Hirani B, Den Haese GJ, Zhang X, Burke JD, Turner K, Gould KL. The Schizosaccharomyces pombe cdc5+ gene encodes an essential protein with homology to c-Myb. *EMBO J*. 1994, 13(2):471-83.

Peteranderl R, Rabenstein M, Shin YK, Liu CW, Wemmer DE, King DS, Nelson HC. Biochemical and biophysical characterization of the trimerization domain from the heat shock transcription factor. *Biochemistry* 1999, 38(12):3559-69.

Riano-Pachon DM, Correa LG, Trejos-Espinosa R, Mueller-Roeber B: Green transcription factors: a chlamydomonas overview. *Genetics* 2008, 179:31-39.

Riechmann JL, Heard J, Martin G, Reuber L, Jiang C, Keddie J, Adam L, Pineda O, Ratcliffe OJ, Samaha RR, et al: Arabidopsis transcription factors: genome-wide comparative analysis among eukaryotes. *Science* 2000, 290:2105-2110.

Saldanha AJ, Java Treeview--extensible visualization of microarray data. *Bioinformatics* 2004 Nov 22;20(17):3246-8.

Schaffer R, Landgraf J, Accerbi M, Simon V, Larson M, Wisman E: Microarray analysis of diurnal and circadian-regulated genes in Arabidopsis. *Plant Cell* 2001, 13:113-123.

Schoffl F, Prandl R, Reindl A. Regulation of the heat-shock response *Plant Physiol*. 1998, 117(4):1135-41.

Sorger PK, Pelham HR. Yeast heat shock factor is an essential DNA-binding protein that exhibits temperature-dependent phosphorylation. Cell 1988, 54(6):855-64.

Shamovsky I, Nudler E: New insights into the mechanism of heat shock response activation. *Cell Mol Life Sci* 2008, 65:855-861.

Stracke R, Werber M, Weisshaar B. The R2R3-MYB gene family in *Arabidopsis thaliana*. Review. *Curr Opin Plant Biol*. 2001, 4(5):447-56.

Takahashi F, Yamagata D, Ishikawa M, Fukamatsu Y, Ogura Y, Kasahara M, Kiyosue T, Kikuyama M, Wada M, Kataoka H. AUREOCHROME, a photoreceptor required for photomorphogenesis in stramenopiles. *Proc Natl Acad Sci U S A* 2007, 104(49):19625-30.

Thompson JD, Gibson TJ, Plewniak F, Jeanmougin F and Higgins DG. The ClustalX windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. *Nucleic Acids Res*. 1997, 24:4876-4882.

Triezenberg SJ. Structure and function of transcriptional activation domains. Review. *Curr Opin Genet Dev*. 1995, (2):190-6.

Wang ZY, Kenigsbuch D, Sun L, Harel E, Ong MS, Tobin EM. A Myb-related transcription factor is involved in the phytochrome regulation of an Arabidopsis Lhcb gene. *Plant Cell* 1997, 9(4):491-507.

Wisniewski J, Orosz A, Allada R, Wu C. The C-terminal region of Drosophila heat shock factor (HSF) contains a constitutively functional transactivation domain. *Nucleic Acids Res*. 1996, 24(2):367-74.

Wu C. Heat shock transcription factors: structure and regulation. *Annu Rev Cell Dev Biol*. 1995, 11:441-69.

Yanhui C, Xiaoyuan Y, Kun H, Meihua L, Jigang L, Zhaofeng G, Zhiqiang L, Yunfei Z, Xiaoxiao W, Xiaoming Q, Yunping S, Li Z, Xiaohui D, Jingchu L, Xing-Wang D, Zhangliang C, Hongya G, Li-Jia Q. The MYB transcription factor superfamily of Arabidopsis: expression analysis and phylogenetic comparison with the rice MYB family. *Plant Mol Biol*. 2006, 60(1):107-24.

Yu EY, Kim SE, Kim JH, Ko JH, Cho MH, Chung IK. Sequence-specific DNA recognition by the Myb-like domain of plant telomeric protein RTBP1. *J Biol Chem*. 2000, 275(31):24208-14.

Chapter III:

Potential Impact of Stress Activated Retrotransposons on Genome Evolution in a Marine Diatom

3.1 Introduction

Transposable elements (TEs) are mobile genetic sequences found within the genomes of most organisms. Sequences derived from TEs represent a genomic fraction of 3% in baker's yeast (Kim et al., 1998), ~20% in fruit fly (Kaminker et al., 2002; Kapitonov and Jurka, 2002; Quesneville et al., 2003), 45% in human (Smit, 1999; Lander et al., 2001) and over 80% in maize (SanMiguel et al., 1996 and 1998). They are thought to be important contributors to genome evolution by inserting into genes or genetic regulatory elements, thereby disrupting gene function, altering levels of gene expression, triggering chromosomal rearrangements, and adding to or subtracting from the physical size of a host genome (Kumar and Bennetzen, 1999). TEs are classified into two groups based on their mode of transposition: retrotransposons or Class 1 TEs which replicate through reverse transcription of an mRNA intermediate, and DNA transposons or Class 2 TEs that use a "cut and paste" mechanism.

A pervasive group of retrotransposons are those flanked by long terminal repeats (LTRs), also typical of retroviruses to which they are related. The LTR direct sequence repeats flank the internal region that encodes both structural and enzymatic proteins with homology to the GAG and POL proteins of retroviruses. The *gag* gene encodes structural proteins that form the virus-like particle (VLP), inside which reverse transcription takes place. The *pol* gene encodes several enzymatic functions, including a protease (PR) that cleaves the POL polyprotein, a reverse transcriptase (RT) that copies the retrotransposon RNA into cDNA, a ribonuclease H domain (RH), and an integrase (IN) that integrates the cDNA into the genome. Two main groups of LTR retrotransposons (LTR-RTs) are found throughout eukaryotes, and are distinguished by the organization of their *pol* genes and similarities among their encoded RT proteins (Xiong and Eickbush, 1990). These groups are referred to as *Ty1/copia* elements (Pseudoviridae) and *Ty3/gypsy* elements (Metaviridae), which respectively display a PR, IN, RT, RH and PR, RT, RH, IN gene organization.

The unicellular chlorophyll *c*-containing algal class Bacillariophyceae (diatoms) is among the most successful and diversified groups of photosynthetic eukaryotes, with possibly over 100,000 extant species (Round et al., 1990) widespread in all kinds of humid and open water environments. The contribution of diatom photosynthesis to marine primary productivity has been estimated to be around 40% (Nelson et al., 1995; Raven and Waite, 2004). Diatoms have a peculiar genetic makeup because they are likely to have emerged following a secondary endosymbiotic process between a photosynthetic eukaryote, most probably red algal-like, and a heterotrophic eukaryote (Falkowski et al., 2004). They are traditionally divided into two orders: the centric diatoms which are radially symmetrical and are thought to have arisen around 180 Million years ago (Mya), followed by the pennate diatoms around 90 Mya which are bilaterally symmetrical. Genome sequences of the centric diatom *Thalassiosira pseudonana* and the pennate diatom *Phaeodactylum tricornutum* have recently become available (Armbrust et al., 2004; Bowler et al., 2009). Because diatoms are single celled organisms that typically reproduce mitotically, the activity of LTR-RTs might have particularly profound effects on genome evolution since any non-lethal retroelement insertion will be transmitted to subsequent generations.

In an analysis of the *T. pseudonana* genome, Armbrust and collaborators identified several TEs (Armbrust et al., 2004). In the current work, we have identified additional TEs in both diatom genomes and we show that LTR-RTs are the most abundant elements, particularly in *P. tricornutum* where they have amplified enormously. Phylogenetic analysis of the RT domain shows that diatom *Ty1/copia*-like elements belong to different lineages, and that two of them are diatom specific. Examination of the CAMERA metagenomic database reveals that these elements are also widespread in different oceans. The potential ecological relevance of these elements for driving genome and population evolution and heterogeneity

has been assessed by examining their expression in response to stress as well as their distribution in *P. tricornutum* accessions collected from different locations worldwide.

3.2 Results

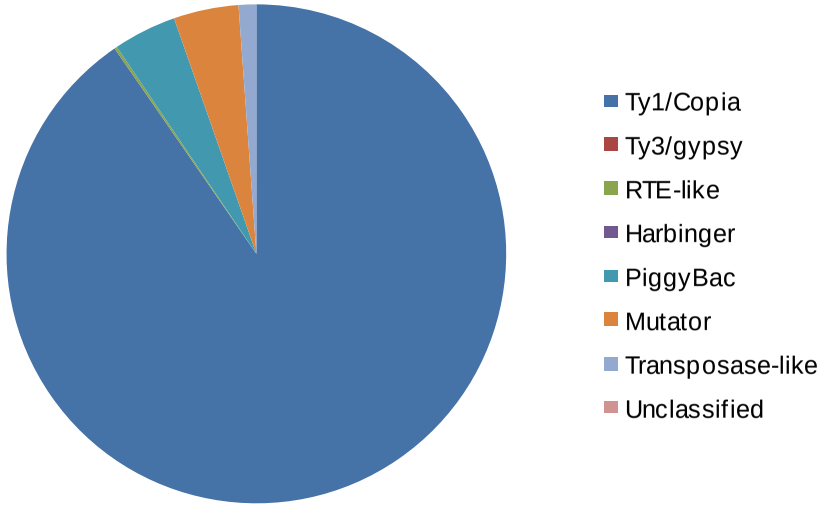
3.2.1 Expansion of LTR Retrotransposons in the *P. tricornutum* genome

We first examined the TE content of diatom genomes. In the *T. pseudonana* genome Armbrust and collaborators identified some *Ty1/copia* and *Ty3/gypsy*-like elements, a family of RTE-like non-LTR retrotransposons, Mutator-like (here denoted as TpMuDR1) and *Harbinger*-like DNA transposons, as well as some unknown unclassified repeated sequences (Armbrust et al., 2004; Kapitonov and Jurka, 2004). In the present work, we could identify additional LTR-RT elements in the *T. pseudonana* genome (Figure 3.1). We also identified numerous *Ty1/copia*-like elements in the *P. tricornutum* genome as well as an RTE-like element, two distinct families of *Mutator*-like elements (one being closely related to *TpMuDR1* elements), and two other different types of uncharacterized transposase-containing elements (one being weakly related to *piggyBac* transposons and for which we also found a homolog in the *T. pseudonana* genome (see Materials and Methods)). *Ty3/gypsy*-like elements were not found in the *P. tricornutum* genome (Figure 3.1A).

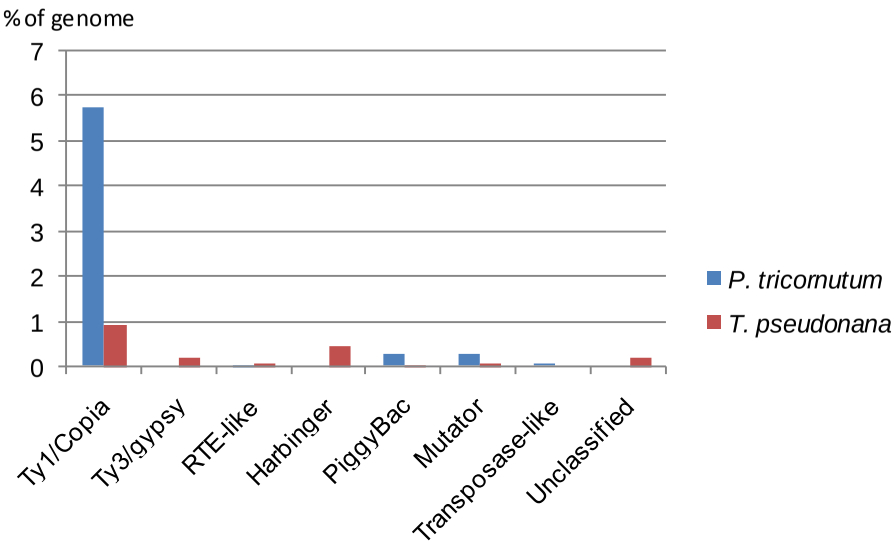
To analyse the contribution of TEs to diatom genomes we used the diatom TE DNA sequences to run the RepeatMasker program (Smit et al., 1996-2004) on both genomes. In total, we found that TEs contribute 1,665 kb (6.4%) of the *P. tricornutum* genome and 590 kb (1.9%) of the *T. pseudonana* genome. Of these, LTR-RTs are the most abundant in both genomes and constitute 90% and 58% of the *P. tricornutum* and *T. pseudonana* TE complement, respectively (Figure 3.1A and 3.1B). Harbinger elements also appear to represent a significant proportion in *T. pseudonana*. In total, the RepeatMasker output indicated that sequences deriving from LTR-RTs make up 5.8% of the *P. tricornutum* genome and 1.1% of the *T. pseudonana* genome (Figure 3.1C). It thus appears that Ty1/copia-like LTR-RTs have significantly expanded in the *P. tricornutum* genome.

Figure 3.1 (next page): Composition of the TE complements in the *P. tricornutum* and *T. pseudonana* genomes. (A and B) Pie chart representing the relative abundance of different TEs to the *P. tricornutum* (A) and *T. pseudonana* (B) TE complements. (C) Histogram representing percent genome coverage across the diatom TE complements. (D) Pie chart representing the relative contribution of the different CoDi groups to the *P. tricornutum* LTR-RT complement.

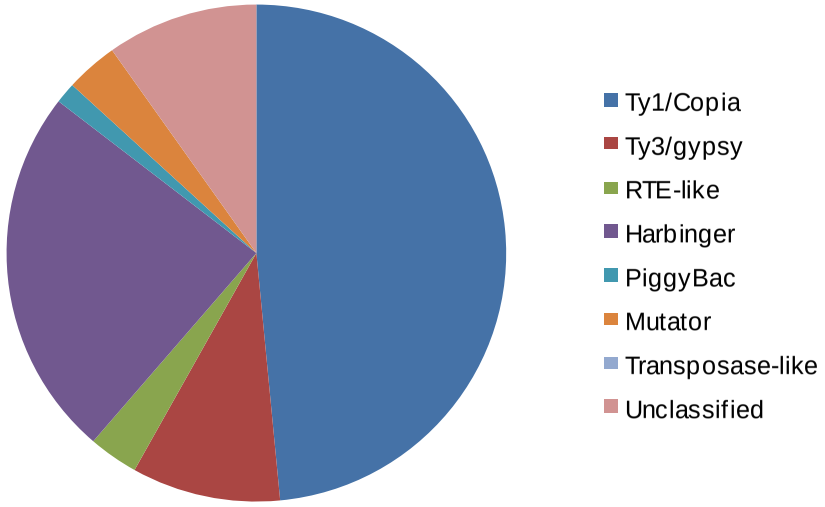
A


Ty1/Copia
Ty3/gypsy
RTE-like
Harbinger
PiggyBac
Mutator
Transposase-like
Unclassified

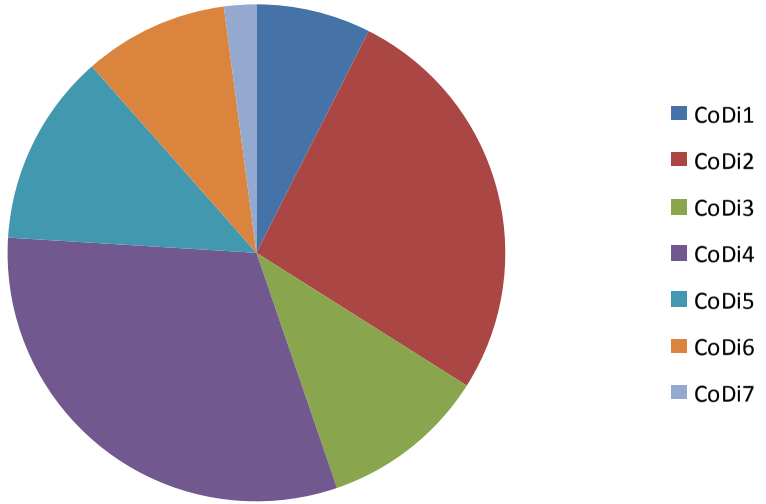
C


% of genome
7
6
5
4
3
2
1
0
Ty1/Copia
Ty3/gypsy
RTE-like
Harbinger
PiggyBac
Mutator
Transposase-like
Unclassified
P. tricornutum
T. pseudonana

B


Ty1/Copia
Ty3/gypsy
RTE-like
Harbinger
PiggyBac
Mutator
Transposase-like
Unclassified

D


CoDi1
CoDi2
CoDi3
CoDi4
CoDi5
CoDi6
CoDi7

3.2.2 Classification of LTR retrotransposon sequences

To further investigate the diatom LTR-RT elements, we manually screened the *P. tricornutum* and *T. pseudonana* nuclear genomes for the presence of putatively autonomous LTR-RTs (see Materials and Methods), and found a total of 42 and 13 putative active elements in the final unmasked assemblies of the *P. tricornutum* and *T. pseudonana* nuclear genomes, respectively. Most of these have greater than 95% identical LTR pairs and display only one or no stop codon/frameshifts between the *gag* and *pol* genes (Supplementary Table 3.1 and Materials and Methods). All the selected sequences from *P. tricornutum* and 11 from *T. pseudonana* belonged to the *Ty1/copia* class with *pol* domains ordered as expected (PR, IN, RT, RH), and the two remaining sequences from *T. pseudonana* belonged to the *Ty3/gypsy* class with *pol* domains also ordered in a typical fashion (PR, RT, RH, IN).

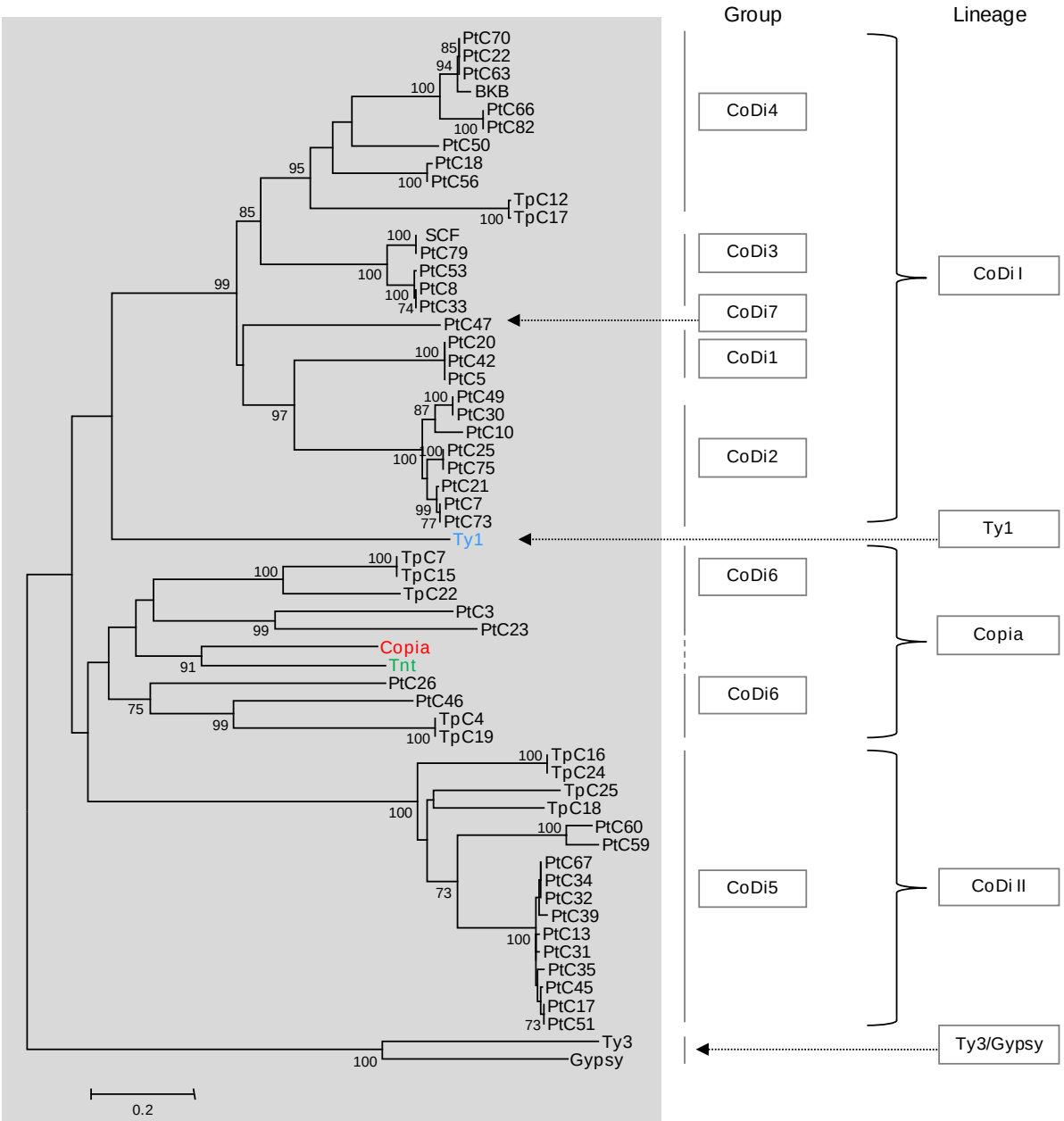
The 53 *Ty1/copia*-like elements identified in the *P. tricornutum* and *T. pseudonana* genomes were classified on the basis of RT domain sequence similarity (see Materials and Methods). Seven groups of *Ty1/copia*-like retroelements were identified and denoted CoDi1 to CoDi7 (*Ty1/Copia*-like elements from Diatoms) (Supplementary Table 3.1, Figure 3.1). While the CoDi1 to CoDi5 groups are quite homogeneous, the CoDi6 group consists of a set of diverse elements. The CoDi7 group is composed of a single element from *P. tricornutum* (PtC47). The CoDi1-2-3-7 groups are specific to *P. tricornutum* whereas the CoDi4-5-6 groups are composed of elements found in both diatom genomes. It appears that the CoDi2 and CoDi4 groups are major components of LTR-RT expansion in the *P. tricornutum* genome (Figure 3.1D).

3.2.3 Phylogenetic analysis

We constructed a phylogenetic tree from a CLUSTALW multiple alignment of the RT domains from the *Ty1/copia*-like shown in Supplementary Table 3.1 as well as reference

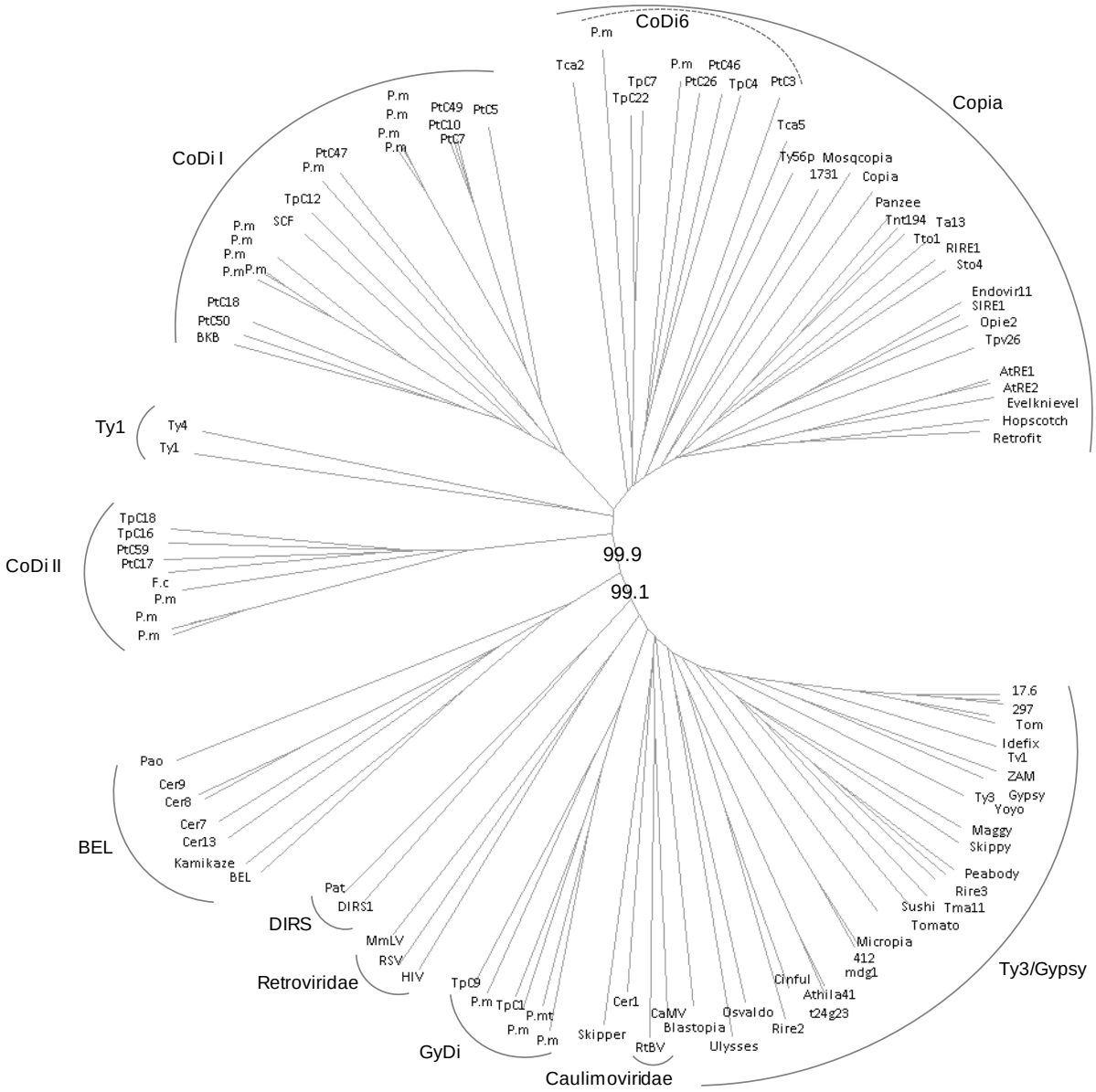
sequences for the Ty1 and Copia lineages (*Tnt* from tobacco, *copia* from fruit fly, and *Ty1* from budding yeast) (Figure 3.2). We observed a distribution of sequences into seven clusters corresponding to the groups defined previously (Supplementary Table 3.1). The most homogeneous clusters represent the groups CoDi1-2-3 composed of sequences present only in *P. tricornutum*. The PtC47 element representing the CoDi7 group appears distantly linked to the CoDi1-2 groups. The lineage linking the CoDi1-2-3-4-7 groups was denoted CoDiI (Figure 3.2). Like CoDi4, the CoDi5 group is composed of sequences from both the centric and the pennate diatom and constitutes a separate lineage we called CoDiII. Finally, the elements from the CoDi6 group which includes elements from both genomes cluster into a highly heterogeneous lineage together with the marker elements *Tnt* and *copia*. In this tree, we can therefore recognize a class of diatom *Ty1/copia*-like elements most closely related to known elements from the *Copia* lineage as well as two diatom-specific lineages, CoDiI and CoDiII (Figure 3.2).

Figure 3.2: Phylogenetic tree showing the relationships between the CoDis and other *Ty1/copia*-like elements. This tree uses the RT domains from *Ty3* and *gypsy* as outgoup and was constructed with the NJ method with the MEGA4 software (Kumar et al., 2004). The bootstrap values were calculated over 1,000 iterations and bootstrap scores over 70% are shown.



To better clarify the evolutionary relationships between the LTR-RTs from diatoms and other retrotransposable elements, we studied RT sequences from a representative subset of elements from each CoDi group defined on the basis of our previous analysis (Supplementary Table 3.1 and Figure 3.2) and RT sequences that we identified from the pennate diatoms *Fragilariopsis cylindrus*, *Pseudo-nitzschia multistriata*, and *Pseudo-nitzschia multiseries*. A phylogenetic representation of diatom RT domains with those belonging to the major lineages of LTR retrotransposons and retroviruses (see Materials and Methods) showed that the heterogeneous CoDi6 group appears closest to the major Copia lineage (Figure 3.3), which includes sequences from animals, plants, yeast, and heterokonts (diatoms), which confirms the origin of the Copia lineage as deeply rooted in eukaryotes. This tree also confirms the distant evolutionary relationships that link the elements from the CoDiI lineage to the Ty1 and Copia lineages and the even more distant relationships that link the CoDiII lineage to these other elements. We also note that the RT sequences from the other diatoms cluster in the CoDiI, CoDiII and Copia lineages, and that the *Ty3/gypsy*-like elements from *T. pseudonana*, *P. multiseries* and *P. multistriata* also segregate together in a diatom-specific cluster (Figure 3.3).

Figure 3.3 (next page): Phylogenetic tree showing the relationships between CoDis and other LTR-RT and retroviral lineages. The bootstrap values were calculated over 1,000 iterations and are indicated for two basal nodes. The tree was constructed with the NJ method using the SplitsTree4 software (Huson and Bryant, 2006). Species abbreviations: P. mt (*Pseudonitzschia multistriata*); P. m (*Pseudonitzschia multiseries*); F. c (*Fragilariopsis cylindrus*).


CoDi6
Copia
CoDi I
Ty1
CoDi II
BEL
DIRS
Retroviridae
GyDi
Caulimoviridae
Ty3/Gypsy
99.9
99.1
P.m
Tca2
TpC7
TpC22
PtC46
PtC26
TpC4
PtC3
Tca5
Ty56p
Mosqcopia
1731
Copia
Panzee
Tnt194
Ta13
Tto1
RIRE1
Sto4
Endovir11
SIRE1
Opie2
Tpv26
AtRE1
AtRE2
Evelknievel
Hopscotch
Retrofit
PtC49
PtC5
PtC10
PtC7
PtC47
TpC12
SCF
PtC18
PtC50
BKB
Ty4
Ty1
TpC18
TpC16
PtC59
PtC17
F.c
Pao
Cer9
Cer8
Cer7
Cer13
Kamikaze
BEL
Pat
DIRS1
MmLV
RSV
HIV
TpC9
TpC1
P.mt
Skipper
Cer1
CaMV
Blastopia
Osvaldo
Ulysses
RtBV
Rire2
Cinful
Athila41
t24g23
mdg1
412
Micropia
Tomato
Sushi
Tma11
Rire3
Peabody
Skippy
Maggy
Ty3
Gypsy
Yoyo
ZAM
Tv1
Idefix
Tom
297
17.6

3.2.4 Expression of LTR retrotransposons in *P. tricornutum* and *T. pseudonana*

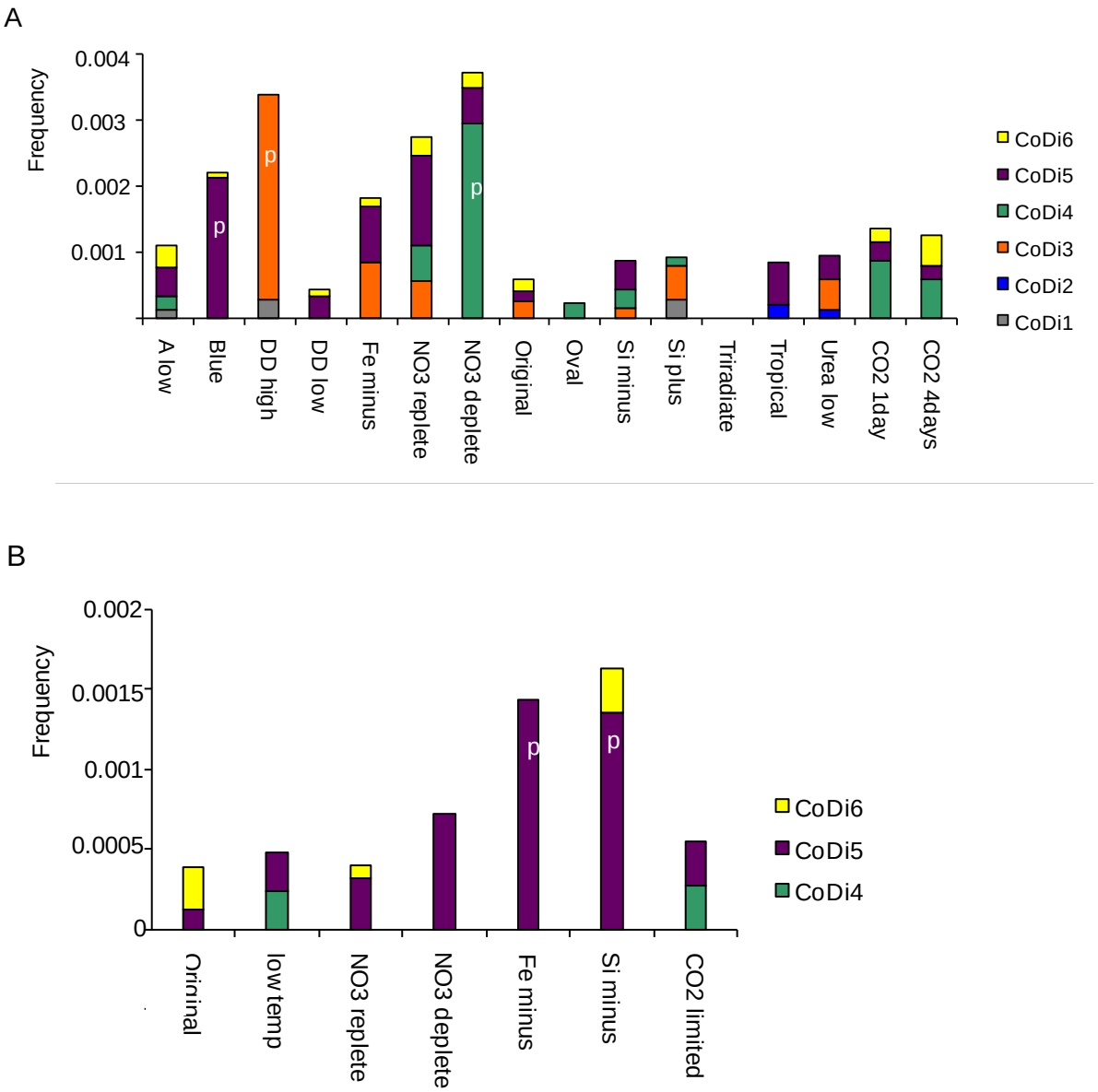
To examine TE expression, the complete nucleotide sequences of the full length elements from *P. tricornutum* listed in Supplementary Table 3.1 were searched in the diatom digital gene expression database (available at http://www.biologie.ens.fr/diatomics/EST3/) (Maheswari et al., 2008) using BLAST. This database comprises more than 200,000 ESTs from *P. tricornutum* and *T. pseudonana* cells grown in a range of different conditions, many of which correspond to different abiotic stresses. The global EST profile of each CoDi group reveals a pattern of higher expression levels under some stress conditions (Figure 3.4). In particular, we focused on two *P. tricornutum* CoDi1 lineage elements that were strongly induced under conditions of nitrate starvation and following exposure to the toxic reactive aldehyde decadienal (DD) (Figure 3.4). These were denoted *Blackbeard* (*Bkb*) and *Surcouf* (*Scf*), respectively, and the contribution of CoDi3 and CoDi4 to the nitrate deplete and DD high libraries are due exclusively to these elements. qRT-PCR was subsequently used to confirm their upregulation in response to nitrogen starvation and following exposure to DD (Table 3.1).

Table 3.1 : Gene expression of specific *Ty1/copia*-like alleles in cultures of *P. tricornutum* grown under particular stress conditions.

Treatment	Target Element	Reference Gene	$2^{-\Delta\Delta CT}$ (fold change)*
DD2 2 hrs	*Surcouf*	TBP	62.83 (38.35-102.94)
DD2 6 hrs	*Surcouf*	TBP	106.64 (78.14-145.54)
DD2 30 hrs	*Surcouf*	TBP	26.48 (15.27-45.92)
DD2 4 days	*Surcouf*	TBP	2.27 (1.56-3.28)
24 hrs (N limitation)	*Blackbeard*	18S rDNA	3.51 (2.67-4.61)
2 weeks (N limitation)	*Blackbeard*	18S rDNA	92.21 (61.59-138)

*Fold change with respect to expression levels in untreated cultures

Figure 3.4: Abundance of CoDi-encoding ESTs in different conditions. (A) EST frequencies of the *P. tricornutum* CoDi elements listed in Supplementary Table 3.1 within the 16 *P. tricornutum* cDNA libraries described and available at www.biologie.ens.fr/diatomics/EST3/. CoDi7 group does not have any EST support. (B) EST frequencies of the *T. pseudonana* CoDi elements listed in Supplementary Table 3.1 within the 7 *T. pseudonana* cDNA libraries described and available at www.biologie.ens.fr/diatomics/EST3/. (A and B) Letter p indicates statistically-supported a (Pearson's Chi squares p= 0.0000) higher EST frequency of a CoDi group in this condition respect to the original library (non-stressed).



3.2.5 Regulation of *Blackbeard*

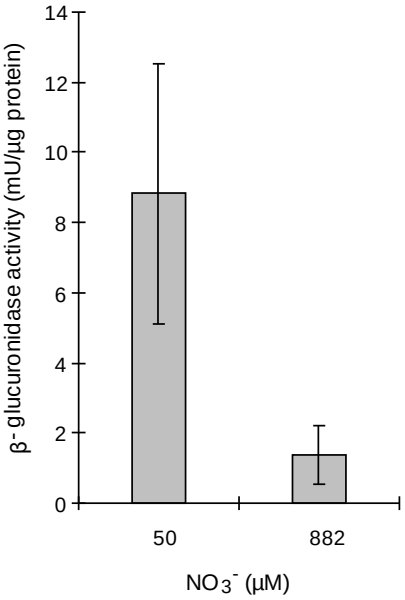
In an effort to better understand *Blackbeard* expression in response to nitrogen limitation, we examined its transcriptional and chromatin-level regulation. Because *cis*-acting elements regulating LTR-RT expression are typically found within LTRs (Pouteau et al., 1991; Servant et al., 2008), we generated a construct containing the *Blackbeard* LTR fused to the *β-glucoronidase* (GUS) reporter gene. Although the *Blackbeard* LTR is only 163 bp, spectrophotometric GUS measurements on *P. tricornutum* lines transformed with this construct showed that it was sufficient to activate transcription in response to nitrate starvation (Figure 3.5A). This shows that the *Blackbeard* LTR alone contains sufficient *cis*-regulatory element information to drive *Blackbeard* expression in response to nitrate limitation.

Cytosine methylation is commonly found in the DNA sequences of transposable elements (at least in genomes in which methylation occurs) and is thought to be involved in the heterochromatin formation and maintenance that controls TE mobility. TE mobilization has been shown to be associated with DNA hypomethylation (Chandler and Walbot, 1986; Scortecci et al., 1997; Miura et al., 2001), and hypomethylation has also been found to accompany active transposition in response to stress (Liu et al., 2004; Hashida et al., 2006). Although nothing is known about such phenomena in diatoms, it was clearly of interest to determine whether they also occur in these organisms. We therefore assessed whether the *Blackbeard* element was methylated using McrPCR. In this method, DNA is digested with McrBC which cleaves DNA containing methylcytosine. Consequently, PCR using McrBC-digested DNA as template leads to a decrease of amplification at methylated (cut) loci with respect to untreated DNA. We first observed that all LTR-RTs tested were methylated in the *P. tricornutum* genome under normal growth conditions (data not shown), demonstrating that

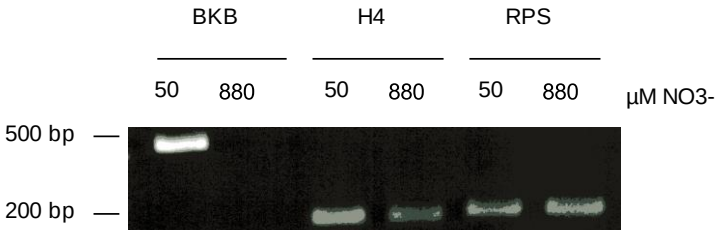
DNA methylation does occur in this diatom. We then compared McrPCR amplification levels using DNA extracted from *P. tricornutum* cells grown in normal and nitrate limiting conditions. Figures 3.5B and 3.5C show that the induction of *Blackbeard* in response to nitrate limitation was accompanied by a decrease in cytosine methylation, suggesting that chromatin remodeling occurs at the *Blackbeard* locus in response to nitrate limitation. Preliminary results from bisulfite sequencing indicate that methylation at the *Blackbeard* locus occurs in a CpG context (data not shown).

Figure 3.5: Regulation of *Blackbeard* expression. (A) Effect of nitrate limitation on the expression of the pLTRbkb-GUS-FcpA construct in transgenic *P. tricornutum* cells. Data represent the average with standard error from seven independent cultures after two weeks nitrate limitation (50 μM NO3-) compared to standard growth medium (882 μM NO3-). (B) Verification of *Blackbeard* transcriptional activation by semi-quantitative RT PCR in the cultures used for McrPCR. (C) McrPCR on *Blackbeard* and H4 and RPS controls using DNA extracted from *P. tricornutum* cells grown under normal and nitrate-limited conditions.

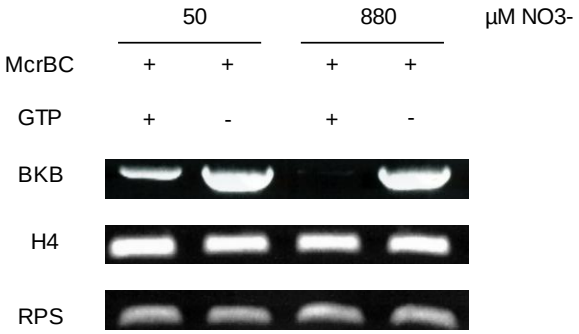
A


β- glucuronidase activity (mU/µg protein)
14
12
10
8
6
4
2
0
50
882
NO3- (µM)

B


BKB
H4
RPS
50
880
50
880
50
880
µM NO3-
500 bp
200 bp

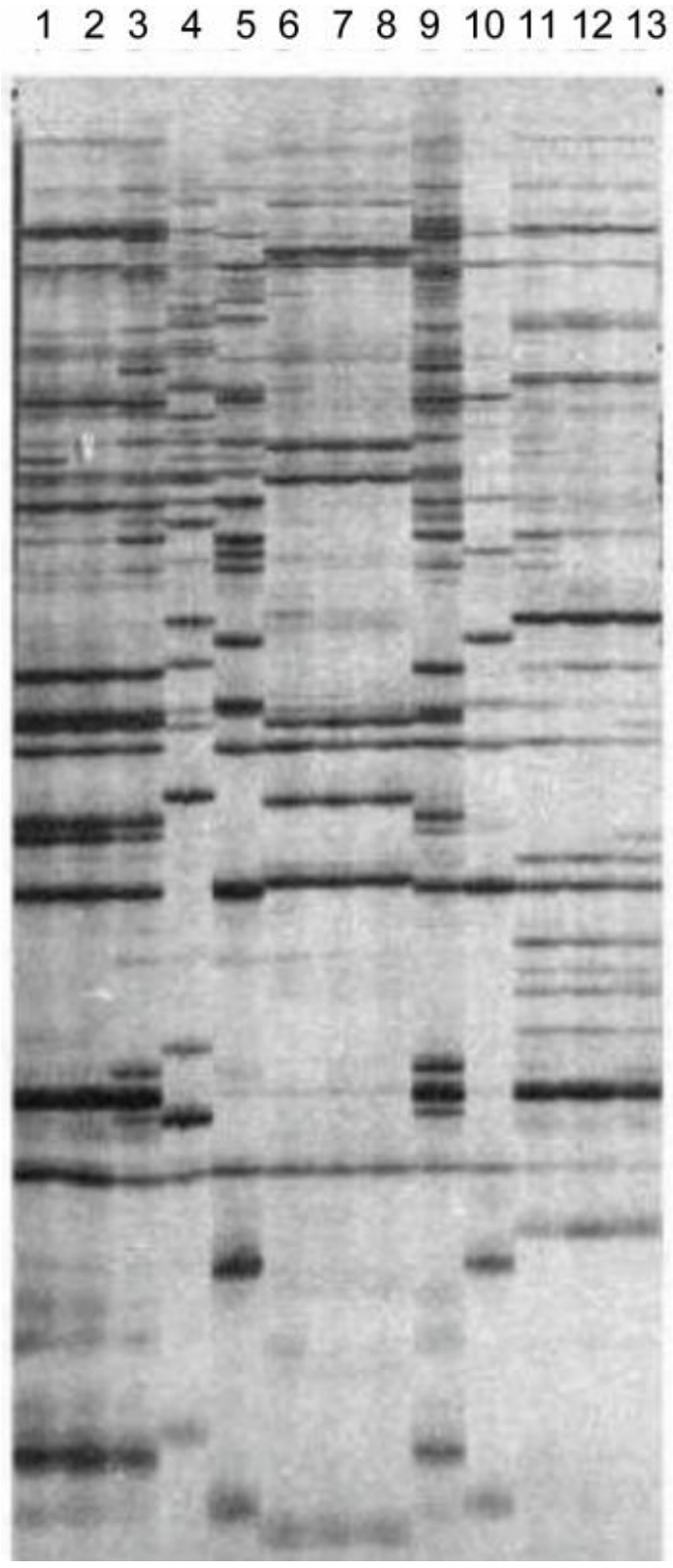
C


50
880
µM NO3-
McrBC
+
+
+
+
GTP
+
-
+
-
BKB
H4
RPS

3.2.6 Insertion polymorphism between *P. tricornutum* accessions

Although suggestive, the induction of *Bkb* expression by nitrate limitation is not proof that it can actually drive genome rearrangements by *de novo* insertion in the genome. In order to better evaluate this possibility, we assessed the distribution of *Bkb* elements in thirteen *P. tricornutum* accessions collected from different locations worldwide by Sequence Specific Amplified Polymorphism (SSAP) (De Martino et al., 2007) (see Materials and Methods). SSAP amplifies the region between a PCR primer site near the end of an element and an adjacent restriction site in the flanking genomic DNA. This global analysis revealed clear differences in *Bkb* insertion profiles in different accessions, demonstrating that it has been transposing in natural environments (Figure 3.6). We were able to confirm the same phenomenon with two other elements, *Scf* and *PtC34* (data not shown). We subsequently cloned several bands from the SSAP gel in order to determine some insertion sites in accessions other than the sequenced genotype (Supplementary Table 3.2). None of the sequences we obtained were inserted inside genes, and most were inserted into intergenic regions, sometimes very close to coding sequences. For example, a PtC34 insertion found in Pt6-7-8 is located 82 bp upstream of the 5' UTR of the gene encoding uroporphyrinogen-III synthase, which catalyses the sixth step of heme biosynthesis. We also found several sequences corresponding to *Bkb* and *Scf* inserted into other TEs (Supplementary Table 3.2).

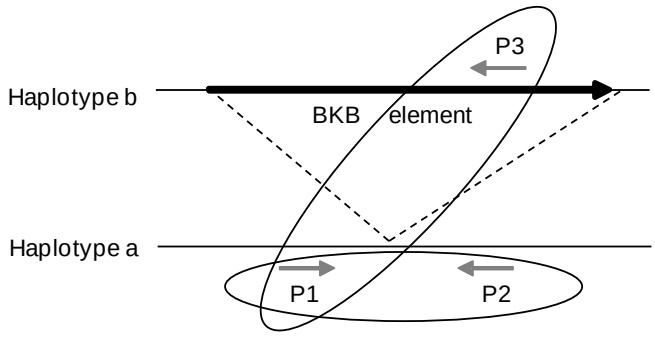
Figure 3.6: Sequence Specific Amplified Polymorphism analysis of *Bkb* in 13 *P. tricornutum* accessions. Each amplified insertion is revealed as a band on a sequencing gel and genomic DNA from the different accessions produces a characteristic fingerprint of bands.



3.2.7 Two distinct haplotypes at loci containing TEs

Analysis of sequencing reads around several TE insertion sites revealed that many were inserted in just one of the haplotypes and that the other haplotype was apparently intact. As an example, the *Blackbeard* insertion is shown in Figure 3.7A and Supplementary Figure 3.1. For this (and all other) insertion events, we could verify by PCR that the allelic specificity is conserved in all accessions in which they are found (accessions Pt1, Pt2, Pt3, and Pt9 for *Bkb*), whereas in the other accessions we could only detect the empty locus (Figure 3.7). Because the oldest of these accessions was collected more than one hundred years ago (De Martino et al., 2007), we can conclude that the *Bkb* insertion must have occurred before this time.

A



B

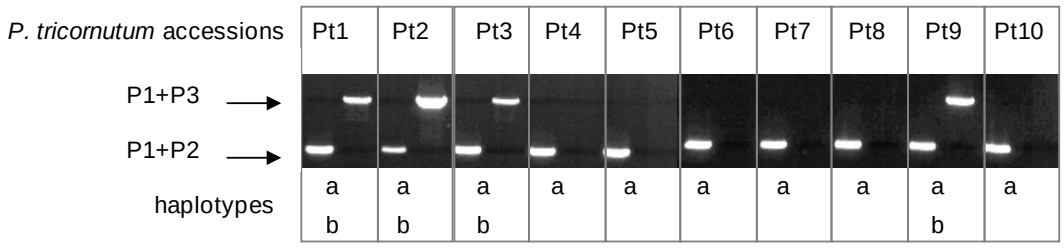


Figure 3.7 (previous page): Analysis of the *Blackbeard* locus. (A) Schematic representation of the primer pairs used to perform PCR at the *Blackbeard* locus. Primer pairs are embedded within ovals and dashed lines indicate the projection of the *Bkb* locus found in haplotype b to its native target site on haplotype a. (B) Haplotype analysis by PCR to assess the presence/absence of the *Blackbeard* insertion in ten *P. tricornutum* accessions. Haplotypes a and b respectively refer to the absence and presence of the *Blackbeard* insertion.

3.2.8 TE-mediated recombination in the *P. tricornutum* genome

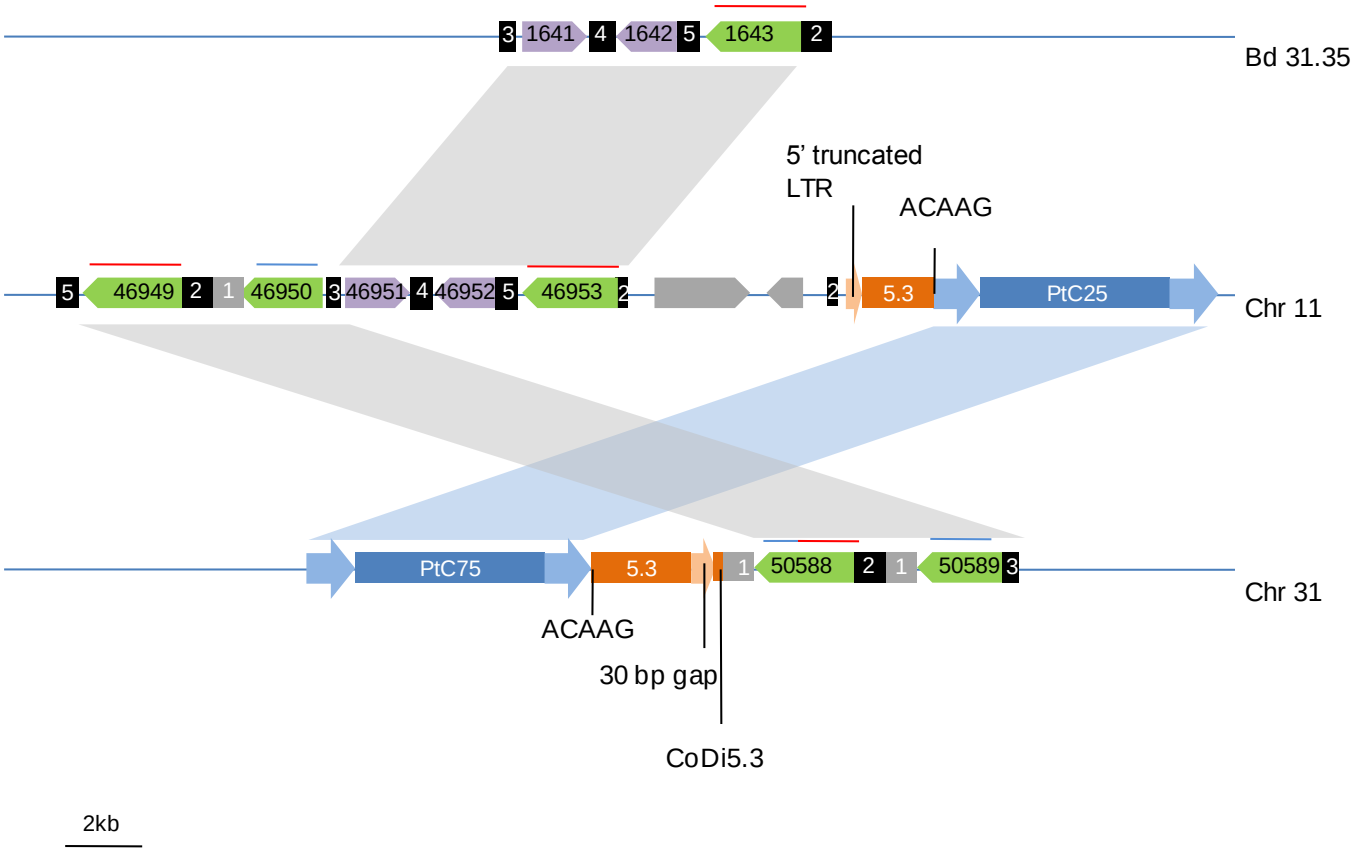
To shed light on the potential impact of LTR-RTs on genome dynamics, we analyzed some signatures of intra- or inter-element recombination in the *P. tricornutum* and *T. pseudonana* genomes (Devos and Bennetzen, 2002). Unequal intrastrand homologous recombination between LTRs of different elements belonging to the same family is a typical example and can result in a net loss of the DNA in between the elements involved. Five examples of this were found in our study of the *P. tricornutum* genome, all of which resulted in clearly recognizable recombinant products in which apparently intact elements with more than 99% identical LTRs lacked the target-site duplication (TSD) (see Supplementary Table 3.1) and were therefore expected to be the product of homologous recombination between two family members. On the other hand, we found no example of this kind in the *T. pseudonana* genome.

We also noticed that the two elements constituting the CoDi2.3 family, PtC25 (on chromosome 11) and PtC75 (on chromosome 31), both lacked a TSD (Supplementary Table 3.1). Closer examination of these loci revealed evidence that these two elements have been co-involved in a recombination event (Figure 3.8). Specifically, we found that the 5' flanking region of PtC25 consists of a truncated CoDi5.3 element and that the 3' flanking sequence of

PtC75 also consists of a truncated CoDi5.3 which is the exact continuation of the PtC25-flanking entity but in addition contains a duplication an ACAAG motif. The most parsimonious explanation for this organization is that either PtC25 or PtC75 inserted inside a CoDi5.3 element and that this insertion generated duplication of the target site (ACAAG). Subsequently, PtC25 and PtC75 engaged in a recombination event that split the CoDi5.3 element into the two halves found on chromosome 11 and 31.

Furthermore, these two genomic regions contain a group of 5 orthologs of an unknown gene family (see Figure 3.8 and supplementary Figure 3.2). The segment containing the two copies located on chromosome 31 and their intergenic region is located less than 1 kb downstream of the CoDi5.3-like element and is highly similar (>97% identity) to the segment containing two of the copies located on chromosome 11 and their intergenic region. The Pt2_50888 gene in fact appears to be the product of recombination between two distinct orthologs as its beginning and downstream region is similar to the Pt2_46949 locus and its end and upstream region appears most similar to the Pt2_46950 and Pt2_50889 loci (Supplementary Figure 3.2). A >7 kb region between the Pt2_46950 locus and the CoDi5.3 segment is also duplicated elsewhere in the genome. These loci therefore provide compelling evidence for TE-mediated recombination events in the *P. tricornutum* genome.

Figure 3.8: Schematic representation of the PtC25 and PtC75 recombinant loci. LTR-RT of the CoDi5.3 (orange) and CoDi2.3 (blue) groups are drawn with their LTRs (flanking arrows). Gene family 1 (green) and gene family 2 (purple) and other genes (grey) are drawn as arrows. Gene family 1 is further distinguished by red and/or blue bar on top and similar colors indicate similar sequences (see supplementary Figure 3.5). Black or grey boxes with identical numbers indicate similar intergenic regions. Grey parallelograms project large duplicated regions from chromosome to chromosome. The blue parallelogram indicates the high similarity between the PtC25 and PtC75 elements. We indicate a 30 bp gap found in the CoDi5.3 segment flanking PtC75. We also indicate that the PtC25-associated CoDi5.3 entity contains a 5' truncated LTR which starts precisely where the gap described on chromosome 31 ends, further consolidating the historical link between these two loci. Bd 31.35 indicates a scaffold that could not be successfully mapped during *P. tricornutum* genome assembly.

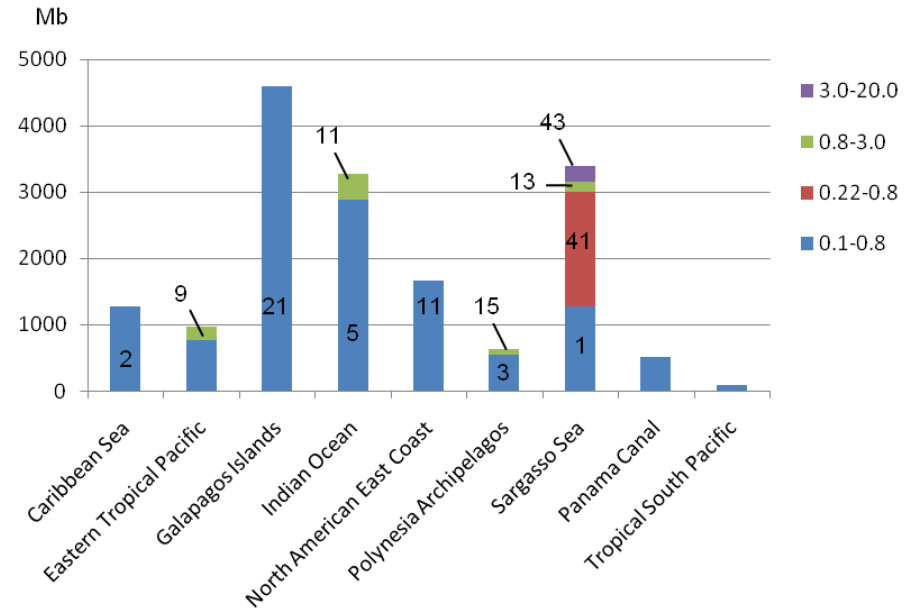

3 1641 4 1642 5 1643 2
Bd 31.35
5’ truncated
LTR
ACAAG
5 46949 2 1 46950 3 46951 4 46952 5 46953 2
2 5.3 PtC25
Chr 11
PtC75 5.3 1 50588 2 1 50589 3
Chr 31
ACAAG
30 bp gap
CoDi5.3
2kb

3.2.9 A high diversity of RT domains from micro-planktonic organisms

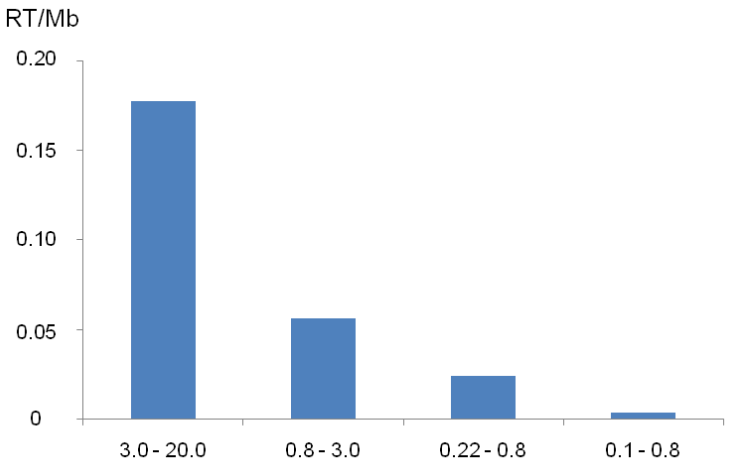
Very little or no data about RT sequences are available from other eukaryotic clades that include planktonic organisms of ecological importance such as dinoflagellates and coccolithophores. In order to investigate deeper the diversity of LTR-RTs that can be found in planktonic organisms, we used our diatom TE dataset to screen the CAMERA metagenomic database (http://camera.calit2.net/), which contains sequences from environmental samples collected during the Global Ocean Sampling (GOS) and Sargasso Sea surveys (Venter et al., 2004; Rusch et al., 2007). These sequences are derived from micro-organisms that were trapped on filters of different sizes (0.1-0.8µm, 0.22-0.8µm, 0.8-3.0µm, 3.0-20.0µm) from the surface water of various parts of the globe including Caribbean Sea, Eastern tropical Pacific, Galapagos Islands, North American East coast, Polynesia Archipelagos, and Sargasso Sea. The size of the database for each filter and at each geographical position is indicated in Figure 3.9A.

Figure 3.9: Distribution of the GOS RT sequences. (A) Size of dataset in megabases (Mb) for each filter across the different geographic positions examined. Numbers indicate the number of RT hits for each filter. (B) Frequency of RT hits across the different filters.

A

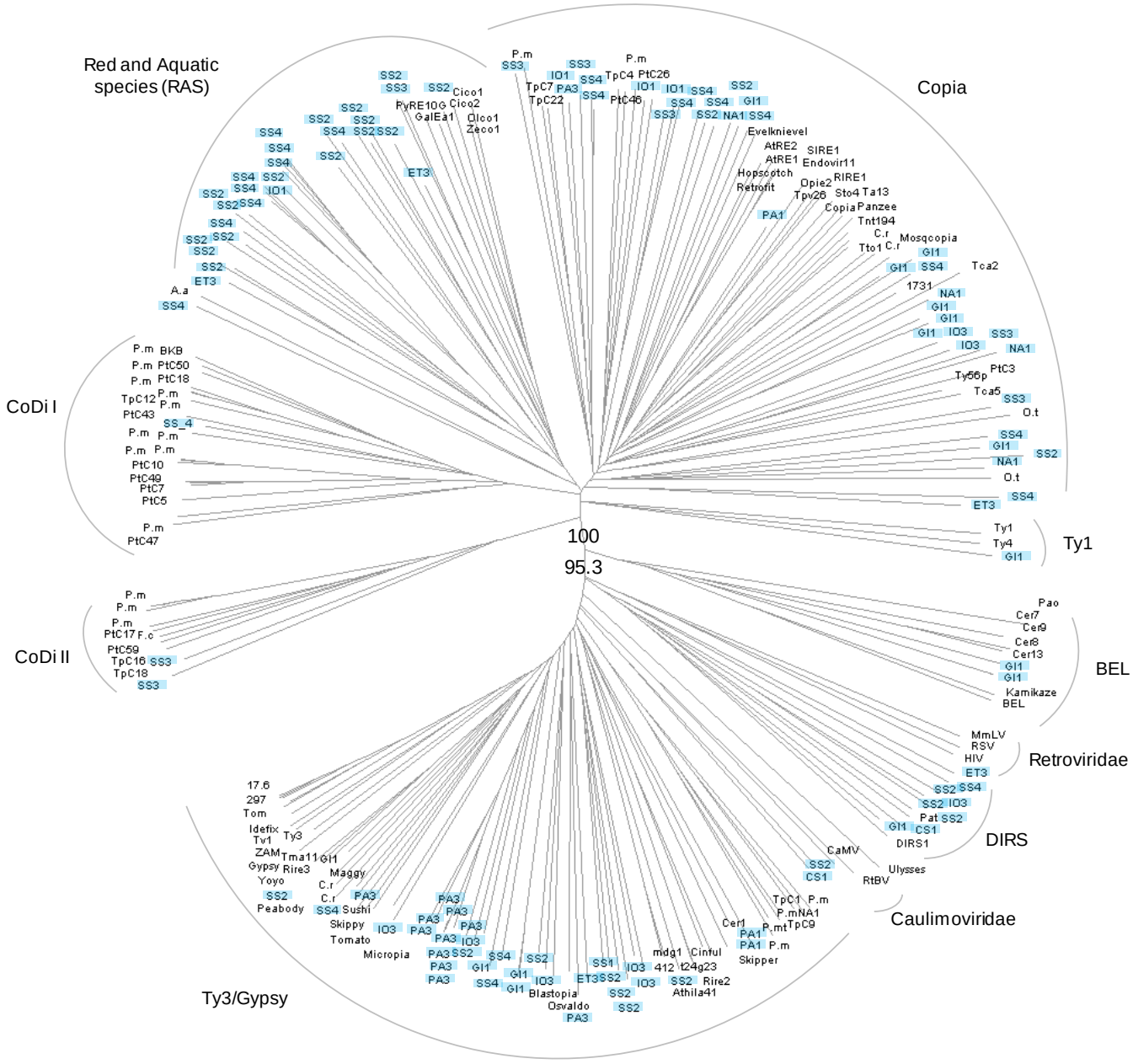


B



We queried by BLAST our entire set of RT domains against the CAMERA protein dataset and retrieved a total of 175 subject sequences (Figure 3.9A), all of which have an LTR-RT for best hit by BLAST comparison with Genbank (data not shown). After normalizing the number of hits from each filter size by its cognate sample size, we observed that the larger the pore size of the filter, the more abundant is the RT domain, with about 0.18 RT domains per Mb of sequence from the 3.0-20.0 µm filters (Figure 3.9B). A total of 115 of these sequences could be included unambiguously in our RT domain alignment and were used to build a phylogenetic tree in which we also incorporated RT sequences from the green algae *Chlamydomonas reinhardtii* and *Ostreococcus tauri*, the brown alga *Aureococcus anophagefferens*, and the RT domain from the *PyRE10G* element found in the red alga *Porphyra yezoensis* (Peddigari et al., 2008) (Figure 3.10). As expected, we observed an enormous diversity of GOS sequences. It was found that GOS RT domains clustered with all the LTR-RT lineages described here, including the CoDiI and CoDiII lineages. However, RT domains belonging to the Ty3/gypsy, Copia, and the recently identified red/aquatic species (RAS) lineage (Terrat et al., 2008) are the most abundant in the dataset analyzed. We also noticed that the RAS-like lineage appears to be quite a diverse assemblage (Figure 3.10). These RAS-like elements appear to be the most abundant in the Sargasso Sea samples, especially from the 0.22-0.8 µm filters (data not shown).

Figure 3.10 (next page): Phylogenetic tree showing the relationships between the reverse transcriptase domains from the CAMERA database, retroviruses, and LTR retrotransposons. The tree was constructed with the NJ method using the SplitsTree4 software (Huson and Bryant, 2006). The bootstrap values were calculated over 1,000 iterations and are indicated for two basal nodes. GOS sequences are labeled by a two- letter code indicating their geographic provenance: Caribbean Sea (CA), Eastern Tropical Pacific (ET), Galapagos Islands (GI), Indian Ocean (IO), North American East Coast (NA), Polynesia Archipelagos (PA), Sargasso Sea (SS); followed by a number indicating filter size: 0.1-0.8 (1), 0.22-0.8 (2), 0.8-3.0 (3), 3.0-20.0 (4). These labels appear with blue background. Species abbreviations: P. mt (*Pseudonitzschia multistriata*)*;* P. m (*Pseudonitzschia multiseries*)*;* F. c (*Fragilariopsis cylindrus*)*;* C. r (*Chlamydomonas Reinhardtii*)*;* O. t (*Ostreococcus tauri*)*;* A.a (*Aureococcus anophagefferens*).


Red and Aquatic species (RAS)
Copia
CoDi I
CoDi II
Ty1
BEL
Retroviridae
DIRS
Caulimoviridae
Ty3/Gypsy
100
95.3

3.3 Discussion

In this work, we have identified seven groups of *Ty1/copia*-like LTR-retrotransposons in diatom genomes. Four groups (CoDi1-2-3 and CoDi7) were found only in the *P. tricornutum* genome whereas elements belonging to the CoDi4-5-6 groups were detected in both diatom genomes. The presence of both classes suggests either that they were present in the diatom common ancestor and that the CoDi1-3 groups became extinct in the lineage leading to the centric species *T. pseudonana*, or that representatives of each group have been separately introduced horizontally in pennate and centric diatoms. The topology of the tree presented in Figure 3.2 shows that CoDi3 and CoDi4 are bootstrap-supported sister groups that share a common ancestor after the separation from CoDi1 and CoDi2. This, together with the fact that we could not detect traces of diverged remnant copies from the CoDi1-3 groups in the *T. pseudonana* genome by BLAST searches (data not shown) favors the horizontal transfer hypothesis to explain the presence of CoDi4 elements in the *T. pseudonana* genome.

Ty3/gypsy-like elements were found in the *T. pseudonana* genome but not in the *P. tricornutum* genome. We also identified RT sequences corresponding to *Ty3/gypsy*-like elements from the pennate diatoms *P. multiseries* and *P. multistriata* which clearly cluster with the GyDi elements (Figure 3.3). Although the number of diatom species for which data is available is low, this suggests that *Ty3/gypsy*-like elements were likely present in the diatom common ancestor, and that these elements have been lost in *P. tricornutum*.

Figure 3.10 shows the retrotransposon sequences found in the CAMERA dataset. Although the vast majority of the sequences derived from these environmental genomic surveys are of bacterial and archaeal origin (Rusch et al., 2007), the authors counted 69 18S rRNA sequences in the analysis of the Sargasso Sea data (Venter et al., 2004) and 98 in the GOS sequence collection (Doug Rusch, personal communication). Thus, some small eukaryotes were also present in these datasets. The observed higher abundance of RT domains in the

fractions containing the larger cells is consistent with higher relative eukaryote/prokaryote abundance in these samples. The RT sequences studied display a huge diversity including some clustering in the CoDiI and CoDiII lineages, which likely testifies for the presence of diatoms in the samples. The other RT sequences may reflect the diversity of LTR retrotransposons populating the genomes of diverse tiny marine eukaryotes such as green, red, or brown algae, dinoflagellates, haptophytes, or euglenoids. For example, the abundance of RAS-clustering sequences in the Sargasso Sea fractions may be indicative of the presence of red algae, although analysis of these eukaryotic fractions did not reveal a particular abundance of red algae (Pigagneau et al., 2008). It will therefore be important to determine which eukaryotic branch or branches the RAS-like sequences collected come from. In addition to the CoDiI, CodiII, and RAS sequences, other discrete clusters shown in Figure 3.10 are exclusively composed of RT sequences from the CAMERA database and are likely to represent RT domains from organisms for which we have little or no genomic knowledge.

The mutagenic potential of LTR retrotransposons (Grandbastien et al., 1989) and the effects of their accumulation (Bennetzen and Kellogg, 1997) and recombination (Vicient et al., 1999) together suggest that active retrotransposons may be major contributors to genome diversification. Accumulated data indicates that retrotransposons in plants (Wessler, 1996), animals and fungi respond to various forms of stress. It has also been shown in natural wild barley populations living on each side of a canyon that LTR retrotransposon dynamics contribute to genome diversity in response to sharp microclimatic divergence (Kalendar et al., 2000). LTR-RTs are hence thought to play a key role in long term adaptation of natural populations exposed to stress by generating genetic diversity within populations. Evidence presented here suggests that this may also be the case in diatoms. For example, *Blackbeard* is one of the most highly expressed genes in the EST library derived from *P. tricornutum* cells grown under nitrate starvation and *Surcouf* is highly expressed in response to DD treatment

(Table 3.1, Figure 3.4). If these expression levels correlate with completion of the retrotransposition cycle, which ends with *de novo* insertions, then nitrate starvation, DD, and perhaps other environmental stressors could lead to an increase in genetic diversity in *P. tricornutum*. LTR-RTs may therefore be major drivers of genetic diversity in *P. tricornutum* populations. Although we have not been able to observe *de novo* insertion of *Bkb* or *Scf* elements following stress, this claim is supported by the different insertions that have been observed in *P. tricornutum* accessions isolated from different locations around the world (Figure 3.6).

The significance of these findings is strengthened by the ecological relevance and common occurrence of stress in marine environments. Nitrogen is the most widespread limiting nutrient for marine phytoplankton (Falkowski, 1997), and transitions between nitrate starved stratified waters and nitrate replete upwelling conditions are a major influence governing marine diatom population oscillations (Smetacek, 1999). Conversely, diatom-derived unsaturated aldehydes can regulate intercellular signalling, stress surveillance, and defence against grazers (Vardi et al., 2006 and 2008; Ianora et al., 2004). Diatoms can sense these aldehydes accurately, whereby subthreshold levels serve as an early-warning protective mechanism, and lethal doses initiate a cascade leading to autocatalytic cell death. Activation of *Surcouf* only after exposure to high levels of aldehydes supports a threshold-dependent response in which activation only occurs under acute stress conditions. Furthermore, the fact that significant aldehyde concentrations are only produced by nutrient-stressed and wounded diatoms suggests a possible role in long term adaptation to abiotic and biotic stress (Fontana et al., 2007; Ribalet et al., 2007a and 2007b; Casotti et al., 2005).

Finally, sexual reproduction in *P. tricornutum* has never been documented. Here, we have seen that *Blackbeard* and *Surcouf* insertions occurred at least a century ago and that both (as well as all other insertions tested; data not shown) have remained in a heterozygous state

until now, in accordance with rare or absent meiotic cycles and only limited crossing overs between chromosome pairs in *P. tricornutum*. The maintenance of LTR-RT insertions in a heterozygous state in the *P. tricornutum* genome could increase the genetic variability between haplotypes and hence enhance adaptation capacity to changing environments. Furthermore, the observation that the *Blackbeard* element is hypomethylated in response to nitrate starvation provides a direct link between environmental stress and chromatin remodeling in diatoms. Such phenomena can confer phenotypic plasticity to an individual species, especially if they are heritable, and may be more useful for environmental adaptation than DNA-based modifications, which are irreversible and more likely to lead to speciation and therefore reproductive isolation. It is therefore possible that epigenetic modifications, combined with TE-mediated genomic rearrangements, maintain population diversity in *P. tricornutum*, as opposed to sex-driven chromosomal recombination. The potential capacity of such processes to monitor and to respond rapidly to changing environmental conditions may have contributed to the evolutionary and ecological success of diatoms in contemporary oceans.

3.4 Materials and Methods

Identification of transposable elements

TE complements from the *P. tricornutum* (http://genome.jgi-psf.org/Phatr2/Phatr2.home.html) and *T. pseudonana* (http://genome.jgi-psf.org/Thaps3/Thaps3.home.html) nuclear genomes were established by BLAST search (Altschul et al., 1990) using the Repbase library (Jurka et al., 2005) or single TE sequences, redundancy search and search for structural features such as ORFs larger than 1000 amino acids (which are characteristic of LTR-RT) and subsequent BLAST comparison with GenBank. When necessary, full length sequences were determined by examining multi-copy alignment. We then searched for the presence of LTRs upstream and downstream of the DNA sequence corresponding to the ORFs containing a polyprotein. LTR size sometimes varied by a few nucleotides between pairs and the length of the longest LTR is reported in Supplementary Table 3.1. The target site duplication was examined in the genomic sequence directly flanking the LTRs. The DNA sequences between LTR pairs were translated in order to eventually identify another ORF (denoted ORF1 in Supplementary Table 3.1) upstream of the ORF containing the polyprotein (ORF2). ORF2 and ORF1 were then submitted to InterProScan (http://www.ebi.ac.uk/InterProScan/). The domain composition and order found in ORF2 was established by performing multiple alignments of the putatively active *Ty1/copia*-like elements from *P. tricornutum* and *T. pseudonana* with *Ty1* from yeast and *Copia* from fruit fly and of the putatively active *Ty3/gypsy*-like elements from *T. pseudonana* with *Gypsy* from fruit fly, and *Ty3* from yeast.

RT domains from *P. multiseries*, *P. multistriata*, *F. cylindrus* (Krell and Gloeckner, 2004), *O. tauri*, *C. reinhardtii, A. anophagefferens* as well as from the GOS and Sargasso Sea metagenomic surveys were found using the RT amino acid sequences from the diatom LTR-

RTs identified in this work and a set of RT domains assembled by Gao and collaborators (including elements from the *Ty3/gypsy*, *Ty1/copia*, DIRS, and BEL groups) as digital probes in BLAST searches (Altschul et al., 1990) directly on the respective cDNA, genomic, and metagenomic databases.

To estimate the number of solo LTRS in both diatom genomes, we used two separate libraries to run the RepeatMasker program: one with LTR sequences only and the other containing the full length sequences. Outputs were mixed and LTR hits were considered solo LTRs if there was no other hit within the upstream and downstream flanking 100 bp.

Classification based on sequence similarity and structural features

We included the *Blackbeard* element in our analysis although it appears to be haplotype-specific and is absent from the final assembly of the *P. tricornutum* genome (see Results). The seven CoDi groups were divided into 26 distinct families on the basis of nucleotide pairwise distances. Further analysis of these elements revealed common structural features that were highly similar within multi-copy families (Supplementary Table 3.1). Overall, the full length diatom retroelement sequences measure between 5182 bp (TpC22) and 8062 bp (PtC26). LTR length varies from 153 bp to 844 bp in the CoDi4.4 and CoDi3.2 families, respectively, and percent identity between LTR pairs varies from 94% to 100%, meaning that all the elements examined are likely to have inserted relatively recently in their respective genomes. The LTR TG/CA terminal inverted repeat is found in 23 out of 26 families and is missing only in CoDi3.2, CoDi4.2 and CoDi4.3. In some cases, such as the CoDi2.2 family, the terminal repeat is longer and contains up to 8 conserved nucleotides. The duplicated target site or direct repeat (DR) is quite heterogeneous within the groups although the *P. tricornutum* elements from the CoDi5 group consistently differ in a few A/T insertions between duplicates (for which the target site was found). Within the GAG-encoding region of these elements,

InterProScan detected tandem CCHC zinc fingers in the elements belonging to the CoDi6.2-6.3-6.4-6.5 families (this domain is commonly found within this region of *Ty1/copia*-like elements).

Finally, the selected *Ty3/gypsy*-like elements from *T. pseudonana* represent two rather closely related groups called GyDi1 and GyDi2 (*Ty3/Gypsy* from Diatoms). Structural features of these elements are also shown in Supplementary Table 3.1. We submitted one element from each family to GenBank (accession numbers are shown in Supplementary Table 3.1).

Phylogenetic analysis

Multiple alignments were performed with the CLUSTALW program (Higgins and Sharp, 1988). Genetic distances were calculated with the Poisson correction method (Nei and Chakraborty, 1976) for amino acid sequences and phylogenetic trees were constructed with the Neighbor-Joining method (Saitou and Nei, 1987). These evolutionary analyses were performed with the MEGA4 and SplitsTree4 platforms (Kumar et al., 2004; Huson and Bryant, 2006).

In addition to the RT sequences identified in this work, phylogenetic trees presented in Figure 3.3 includes RT domains from Ty1/Copia, Ty3/Gypsy, DIRS and BEL LTR-RT lineages (Gao et al., 2003) (dataset obtained from Voytas D.F lab website at http://www.public.iastate.edu/~voytas/Data%20from%20Pubs.html), as well as RT sequences from the Retroviridae human immunodeficiency virus type 1 (HIV), Rous sarcoma virus (RSV), and moloney murine leukaemia virus (MmLV) and from the Caulimoviridae Cauliflower mosaic virus (CaMV) and Rice tungro bacilliform virus (RtBV).

In addition to the sequences used in Figure 3.3 and 115 RT sequences from the CAMERA metagenomic database (http://camera.calit2.net/), Figure 3.5 is built from a CLUSTALW alignment including also four RT sequences from *C. reinhardtii*, two from the *O. tauri*, one

RT sequence from *A. anophagefferens* (http://www.jgi.doe.gov/), and the RT sequence of the *PyRE10G* element from *P. yezoensis* (AB286055). For all phylogenetic analysis, the residues used were a modification of those originally identified by Toh et al. (1983 and 1985) in retroviral, human hepatitis B virus (HBV), cauliflower mosaic virus (CaMV), and several retrotransposon sequences from Drosophila (Xiong and Eickbush, 1990).

Cell culture and accessions

Axenic cultures of *P. tricornutum* Bohlin clone Pt1 8.6 (CCMP2561) were obtained from the culture collection of the Provasoli-Guillard National Center for Culture of Marine Phytoplankton, Bigelow Laboratory for Ocean Sciences, USA. Cultures were grown in f/2 medium (Guillard, 1975) made with 0.2-µm-filtered and autoclaved local seawater supplemented with f/2 vitamins and inorganic nutrients (filter sterilized and added after autoclaving). Cultures were incubated at 18°C under cool white fluorescent lights at approximately 75 $\mu mol.m^{-2}.s^{-1}$ constant light and maintained in exponential phase in semi-continuous batch cultures. Sterility was monitored by occasional inoculation into peptone-enriched media to check for bacterial growth (Andersen et al., 1997).

In order to evaluate the effect of nitrate stress on *Blackbeard* mRNA levels, cells were transferred to media modified with 50 µM NO_3^- and maintained in exponential phase in semi-continuous batch cultures. Samples were collected after 24 hrs and after 2 weeks exposure to nitrogen limitation. In order to evaluate the effect of diatom-derived reactive aldehydes on *Surcouf* transcript abundance, 2 liters of exponential *P. tricornutum* culture was treated with 2 µg/mL (2E,4EZ)-decadienal (DD) and control culture was treated with equivalent volume of methanol (DD solvent). Samples of 250 mL were collected in the indicated time points (0, 2, 6, 30, 96 hr) after exposure to DD treatment. (2E,4*E/Z*)-decadienal (DD) was obtained from Acros Organics USA. DD was dissolved in methanol, and concentrations were determined by

measuring absorption at the lambda max for DD of 274 nm, using a Hewlett-Packard 8453 spectrophotometer. Diatom cells were harvested by centrifugation for 15 min at 3,000 *g*, washed with 12 mL of PBS, aliquoted into 2 mL Eppendorf tubes, and pelleted for 3 min at 10,000 *g*. Cell pellets were frozen instantly in liquid nitrogen and stored at -80°C before proceeding with RNA extraction.

The original sampling location of the *P. tricornutum* accessions Pt1-10 have been recently described in (De Martino et al., 2007). We recently obtained three additional *P. tricornutum* accessions that we included for our SSAP analysis. Pt11 and Pt13 were sampled in 2008, respectively, in the Gulf of Naples and the Gulf of Salerno, Italy. Pt12 was obtained from the Roscoff culture collection.

SSAP

SSAP experiments were conducted as previously described (Petit et al., 2007). Genomic DNA (500ng) was digested with MspI and ligated to an MspI adaptor obtained by the annealing of two primers (Adap-MspI-C :5'-CGT TCT AGA CTC ATC-3' and Adap-MspI-L : 5'-GAC GAT GAG TCT AGA A-3'). SSAP amplification was done by using a non labeled adaptor primer Msp1 (5'-GAT GAG TCT AGA ACG GC-3') and one of the following 33P-labelled LTR primers (*Bkb*, *Scf* and PtC34). Amplified products were separated on 6% denaturing polyacrylamide gels and exposed after drying to Kodak BioMax XAR films (Carestream Health Inc, Rochester). List of LTR primers: *Bkb*-Rev: 5'-ACG ATA ACC GAC CAG AAT CG-3'; *Scf*-Rev: 5'-CCC GAA AAA CAT TGC CTC TA-3'; PtC34-Rev: 5'-ATC GGA TCC AGG ACT TTG TG-3'

RNA purification and reverse transcription

mRNA levels of *Blackbeard* and *Surcouf* were analyzed using q-RT-PCR from triplicate samples collected from biological replicates of nitrate starved or DD-treated exponential grown cultures. Total RNA was extracted from approximately 10^8 cells using TRIzol Reagent (Invitrogen) and contaminating DNA was removed with TURBO DNase via treatment (Ambion), both according to manufacturer's protocols. RNA was then reverse transcribed into first strand cDNA with the SuperScript™ III First-Strand Synthesis System for RT-PCR (Invitrogen) using oligo-dT primers. Gene transcription was measured using the Brilliant® SYBR® Green QPCR Core Reagent Kit and the Stratagene MX3000P QPCR machine (Stratagene). Primers used for real-time PCR were *Surcouf* Fwd, 5'-CGA CCA CCG GCA TAC TTA TT-3', *Surcouf* Rev 5'-GGT TGT ACC GCA AGG CTA TG–3', *Blackbeard* Fwd 5'-GTG TTC TTG CTG CAA ATG GA-3', *Blackbeard* Rev 5'-ATT CAT CGG GGT CAC CAA TA-3', 18S rDNA Fwd 5'-CAT CCT TGG GTG GAA TCA GT-3' and 18S rDNA 5'-TGC GCA AAC CAA CAA AAT AG-3'. Additional primer sets were designed for Histone H4 and for TBP (TATA box binding protein) which served as a housekeeping gene for normalizing expression of the target gene (Siaut et al., 2007). For each treatment, we evaluated each of the housekeeping genes and selected the one that showed the least amount of variation across conditions.

GUS assay

The pLTRbkb-GUS-FcpA plasmid was constructed from the FcpBp-GUS-FcpA vector (Falciatore et al., 1999) in which the FcpB promoter has been removed by KpnI/SalI digestion and replaced by ligation with a PCR fragment corresponding to *Blackbeard* LTR amplified using the Fwd 5'-CTT AGT GGT ACC TAG AAA AAC CCC ACG TCA AGC-3' and Rev 5'-CTT AGT GTC GAC GAT AAA CTA GAA AAC TGC AAC GAT AAC-3' and digested

with KpnI/SalI. The pLTRbkb-GUS-FcpA vector was introduced into *P. tricornutum* by microparticle bombardment using a Biolistic PDS-1000/He Particle Delivery System (Bio-Rad, Hercules, CA, USA) as described by Falciatore et al. (1999).

For β-glucuronidase (GUS) assays, 7 colonies carrying the pLTRbkb-GUS-FcpA construct were grown to mid-log phase in media containing 50 or 882 μM NO_3^-. Two weeks after cells were transferred to 50 μM NO3-, 20 ml cultures were collected by centrifugation at 3,800 rpm for 15 min at 4°C and resuspended in 120 μl freshly prepared GUS extraction buffer (50 mM $NaPO_4$ pH 7.0, 10 mM β–mercaptoethanol, 0.1% Triton X-100), twice frozen in liquid nitrogen and thawed at 37°C, and finally centrifuged at 12,000 rpm for 5 min at 4°C. Soluble proteins were quantified with the Bio-Rad Protein Assay. The fresh extracts were used for spectrophotometric GUS assays performed by incubating at least 10 μg of total protein extract with the GUS enzyme substrate p-nitrophenyl glucuronide (PNPG) at a final concentration of 1 mM, in a total reaction volume of 1 ml. After a one hour incubation at 37°C, the colorimetric reaction was stopped by adding 0.4 ml 2.5 M 2-amino-2-methyl-1,3-propandiol and the absorbance measured at 415 nm. The enzymatic GUS activity was calculated on the base of the O.D. recorded and the molar extinction coefficient of the GUS substrate p-nitrophenol. One unit is defined as the quantity of enzyme that produces one nanomole of product in one minute at 37 °C (Jefferson et al., 1987).

McrPCR

P. tricornutum cells grown were grown as described above under normal and nitrate-limited conditions for two weeks. DNA and RNA were extracted from 20 mL of culture for each condition. After cDNA synthesis from RNA samples (as described above) *Blackbeard* expression was verified by semi-quantitative RT PCR using the primers used for Q-PCR (see above) and primers amplifying the H4 and RPS housekeeping genes as controls (Siaut et al.,

2007). For McrPCR, 1µg of DNA from each sample was incubated for 1 hour at 37°C with 20 units McrBC endonuclease supplemented with 100 µg/ml bovine serum albumin and 1 mM guanosine triphosphate. Negative controls were obtained with the same experimental procedure but replacing guanosine triphosphate with water. The enzyme was subsequently inactivated by incubation at 65°C for 10 minutes. Digestion efficiency of the *Blackbeard* locus was measured by semi-quantitative PCR using forward genomic primer -AAT ATT GGT CTT CGG CAA CG-3' and the *Blackbeard*-specific reverse primer 5'-GCT TCC GTC AAA CAC TCA CA-3' and we used the primers amplifying the H4 and RPS genes as controls (see above).

PCR haplotype/accession analysis

Polymerase chain reactions were performed using template DNA extracted from cultures of the ten different *P. tricornutum* accessions (see previous). The primers used to assess the presence of the two different haplotypes at the *Blackbeard* locus in DNA extracts from the ten accessions were the genomic Fwd 5'-AAT ATT GGT CTT CGG CAA CG-3' paired with the genomic Rev 5'-TTT GAC CCT ATT GGC TAC CG-3' or paired with the *Blackbeard*-specific Rev 5'-GCT TCC GTC AAA CAC TCA CA-3'. The primers used to assess the presence of the two different haplotypes at the *Surcouf* locus were the genomic Fwd 5'-TGT CTA TTG ACA TTT TGG AAG GTG-3'paired with the genomic Rev 5'-AGA TTC ATC AAT GGA TCA TCT CTC-3' or paired with the *Surcouf*-specific Rev 5'-GGG TAC CTG CTC CAT ATG TAG GTT-3'. Additional primer sets were designed for the other insertions analyzed.

3.5 References:

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

Andersen RA, Morton, S.L., and Sexton, J.P: CCMP - Provasoli-Guillard National Center for Culture of Marine Phytoplankton 1997 list of strains. *Journal of Phycology* 1997, 33 (suppl):1-75.

Armbrust V, Berges J, Bowler C, Green B, Martinez D, Putnam N, Zhou S, Allen A, Apt K, Bechner M, et al: The Genome of the Diatom Thalassiosira Pseudonana: Ecology, Evolution, and Metabolism. *Science* 2004, 306:79-86.

Bennetzen JL, Kellogg EA: Do Plants Have a One-Way Ticket to Genomic Obesity? *The Plant cell* 1997, 9:1509-1514.

Bowler C, Allen AE, Badger JH, Grimwood J, Jabbari K, Kuo A, Maheswari U, Martens C, Maumus F, Otillar RP, et al: The Phaeodactylum genome reveals the evolutionary history of diatom genomes. *Nature* 2008, 456:239-244.

Casotti R, Mazza S, Brunet C, Vantrepotte V, Ianora A, Miralto A: Growth inhibition and toxicity of the diatom aldehyde 2-trans, 4-trans-decadienal on Thalassiosira weissflogii (Bacillariophyceae). *Journal of Phycology* 2005, 41:7-20.

Chandler VL, Walbot V: DNA modification of a maize transposable element correlates with loss of activity. *Proc Natl Acad Sci U S A* 1986, 83:1767-1771.

De Martino A, Meichenin A, Shi J, Pan KH, Bowler C: Genetic and phenotypic characterization of Phaeodactylum tricornutum (Bacillariophyceae) accessions. *Journal of Phycology* 2007, 43:992-1009.

Falciatore A, Casotti R, Leblanc C, Abrescia C, Bowler C: Transformation of Nonselectable Reporter Genes in Marine Diatoms. *Mar Biotechnol (NY)* 1999, 1:239-251.

Falkowski PG: Evolution of the nitrogen cycle and its influence on the biological sequestration of CO2 in the ocean. *Nature* 1997, 387:272-275.

Falkowski PG, Katz ME, Knoll AH, Quigg A, Raven JA, Schofield O, Taylor FJ: The evolution of modern eukaryotic phytoplankton. *Science* 2004, 305:354-360.

Fontana A, d'Ippolito G, Cutignano A, Romano G, Lamari N, Gallucci AM, Cimino G, Miralto A, Ianora A: LOX-induced lipid peroxidation mechanism responsible for the detrimental effect of marine diatoms on Zooplankton grazers. *Chembiochem* 2007, 8:1810-1818.

Gabriel A, Boeke JD: Reverse transcriptase encoded by a retrotransposon from the trypanosomatid Crithidia fasciculata. *Proc Natl Acad Sci U S A* 1991, 88:9794-9798.

Gao X, Havecker ER, Baranov PV, Atkins JF, Voytas DF: Translational recoding signals between gag and pol in diverse LTR retrotransposons. *RNA* 2003, 9:1422-1430.

Grandbastien MA, Spielmann A, Caboche M: Tnt1, a mobile retroviral-like transposable element of tobacco isolated by plant cell genetics. *Nature* 1989, 337:376-380.

Guillard RRL: Culture of phytoplankton for feeding marine invertebrates. In *Culture of Marine Invertebrate Animals.* Edited by Smith WLaCMH. New York, USA: Plenum Press; 1975

Hashida SN, Uchiyama T, Martin C, Kishima Y, Sano Y, Mikami T: The temperature-dependent change in methylation of the Antirrhinum transposon Tam3 is controlled by the activity of its transposase. *Plant Cell* 2006, 18:104-118.

Higgins DG, Sharp PM: CLUSTAL: a package for performing multiple sequence alignment on a microcomputer. *Gene* 1988, 73:237-244.

Huson DH, Bryant D: Application of phylogenetic networks in evolutionary studies. *Mol Biol Evol* 2006, 23:254-267.

Ianora A, Miralto A, Poulet SA, Carotenuto Y, Buttino I, Romano G, Casotti R, Pohnert G, Wichard T, Colucci-D'Amato L, et al: Aldehyde suppression of copepod recruitment in blooms of a ubiquitous planktonic diatom. *Nature* 2004, 429:403-407.

Jefferson RA, Kavanagh TA, Bevan MW: GUS fusions: beta-glucuronidase as a sensitive and versatile gene fusion marker in higher plants. *EMBO J* 1987, 6:3901-3907.

Jurka J, Kapitonov VV, Pavlicek A, Klonowski P, Kohany O, Walichiewicz J: Repbase Update, a database of eukaryotic repetitive elements. *Cytogenet Genome Res* 2005, 110:462-467.

Kalendar R, Tanskanen J, Immonen S, Nevo E, Schulman AH: Genome evolution of wild barley (Hordeum spontaneum) by BARE-1 retrotransposon dynamics in response to sharp microclimatic divergence. *Proc Natl Acad Sci U S A* 2000, 97:6603-6607.

Kaminker JS, Bergman CM, Kronmiller B, Carlson J, Svirskas R, Patel S, Frise E, Wheeler DA, Lewis SE, Rubin GM, et al: The transposable elements of the Drosophila melanogaster euchromatin: a genomics perspective. *Genome Biol* 2002, 3:RESEARCH0084.

Kapitonov VV, Jurka J: Molecular paleontology of transposable elements in the Drosophila melanogaster genome. *Proc Natl Acad Sci U S A* 2003, 100:6569-6574.

Kapitonov VV, Jurka J: Harbinger transposons and an ancient HARBI1 gene derived from a transposase. *DNA Cell Biol* 2004, 23:311-324.

Kim JM, Vanguri S, Boeke JD, Gabriel A, Voytas DF: Transposable elements and genome organization: a comprehensive survey of retrotransposons revealed by the complete Saccharomyces cerevisiae genome sequence. *Genome Res* 1998, 8:464-478.

Krell A, Gloeckner, G: Analysis of an osmotic stress induced cDNA library of the psychrophilic diatom Fragilariopsis cylindrus. 2004.

Kumar A, Bennetzen JL: Plant retrotransposons. *Annu Rev Genet* 1999, 33:479-532.

Kumar S, Tamura K, Nei M: MEGA3: Integrated software for Molecular Evolutionary Genetics Analysis and sequence alignment. *Brief Bioinform* 2004, 5:150-163.

Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W, et al: Initial sequencing and analysis of the human genome. *Nature* 2001, 409:860-921.

Liu ZL, Han FP, Tan M, Shan XH, Dong YZ, Wang XZ, Fedak G, Hao S, Liu B: Activation of a rice endogenous retrotransposon Tos17 in tissue culture is accompanied by cytosine demethylation and causes heritable alteration in methylation pattern of flanking genomic regions. *Theor Appl Genet* 2004, 109:200-209.

Maheswari U, Mock T, Armbrust EV, Bowler C: Update of the Diatom EST Database: a new tool for digital transcriptomics. *Nucleic Acids Res* 2008.

Miura A, Yonebayashi S, Watanabe K, Toyama T, Shimada H, Kakutani T: Mobilization of transposons by a mutation abolishing full DNA methylation in Arabidopsis. *Nature* 2001, 411:212-214.

Nei M, Chakraborty R: Empirical relationship between the number of nucleotide substitutions and interspecific identity of amino acid sequences in some proteins. *J Mol Evol* 1976, 7:313-323.

Nishihara H, Okada N: Retroposons: genetic footprints on the evolutionary paths of life. *Methods Mol Biol* 2008, 422:201-225.

Peddigari S, Zhang W, Takechi K, Takano H, Takio S: Two different clades of copia-like retrotransposons in the red alga, Porphyra yezoensis. *Gene* 2008, 424:153-158.

Petit M, Lim KY, Julio E, Poncet C, Dorlhac de Borne F, Kovarik A, Leitch AR, Grandbastien MA, Mhiri C: Differential impact of retrotransposon populations on the genome of allotetraploid tobacco (Nicotiana tabacum). *Mol Genet Genomics* 2007, 278:1-15.

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

Pouteau S, Huttner E, Grandbastien MA, Caboche M: Specific expression of the tobacco Tnt1 retrotransposon in protoplasts. *EMBO J* 1991, 10:1911-1918.

Quesneville H, Nouaud D, Anxolabehere D: Detection of new transposable element families in Drosophila melanogaster and Anopheles gambiae genomes. *J Mol Evol* 2003, 57 Suppl 1:S50-59.

Raven JA, Waite AM: The evolution of silicification in diatoms: inescapable sinking and sinking as escape? *New Phytologist* 2004, 162:45-61.

Ribalet F, Berges JA, Ianora A, Casotti R: Growth inhibition of cultured marine phytoplankton by toxic algal-derived polyunsaturated aldehydes. *Aquatic Toxicology* 2007, 85:219-227.

Ribalet F, Wichard T, Pohnert G, Ianora A, Miralto A, Casotti R: Age and nutrient limitation enhance polyunsaturated aldehyde production in marine diatoms. *Phytochemistry* 2007, 68:2059-2067.

Round FE, Crawford, R.M., Mann, D.G: *The Diatoms: Biology and Morphology of the Genera.* London, UK: Cambridge University Press; 1990.

Rusch DB, Halpern AL, Sutton G, Heidelberg KB, Williamson S, Yooseph S, Wu D, Eisen JA, Hoffman JM, Remington K, et al: The Sorcerer II Global Ocean Sampling expedition: northwest Atlantic through eastern tropical Pacific. *PLoS Biol* 2007, 5:e77.

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

SanMiguel P, Gaut BS, Tikhonov A, Nakajima Y, Bennetzen JL: The paleontology of intergene retrotransposons of maize. *Nat Genet* 1998, 20:43-45.

SanMiguel P, Tikhonov A, Jin YK, Motchoulskaia N, Zakharov D, Melake-Berhan A, Springer PS, Edwards KJ, Lee M, Avramova Z, Bennetzen JL: Nested retrotransposons in the intergenic regions of the maize genome. *Science* 1996, 274:765-768.

Scala S, Carels N, Falciatore A, Chiusano ML, Bowler C: Genome properties of the diatom Phaeodactylum tricornutum. *Plant Physiol* 2002, 129:993-1002.

Scortecci KC, Dessaux Y, Petit A, Van Sluys MA: Somatic excision of the Ac transposable element in transgenic Arabidopsis thaliana after 5-azacytidine treatment. *Plant Cell Physiol* 1997, 38:336-343.

Servant G, Pennetier C, Lesage P: Remodeling yeast gene transcription by activating the Ty1 long terminal repeat retrotransposon under severe adenine deficiency. *Mol Cell Biol* 2008, 28:5543-5554.

Siaut M, Heijde M, Mangogna M, Montsant A, Coesel S, Allen A, Manfredonia A, Falciatore A, Bowler C: Molecular toolbox for studying diatom biology in Phaeodactylum tricornutum. *Gene* 2007, 406:23-35.

Smetacek V: Diatoms and the ocean carbon cycle. *Protist* 1999, 150:25-32.

Smit AF: Interspersed repeats and other mementos of transposable elements in mammalian genomes. *Curr Opin Genet Dev* 1999, 9:657-663.

Smit AFA, Hubley, R., Green, P: RepeatMasker Open-3.0. 1996-2004.

Stekel DJ, Git Y, Falciani F: The comparison of gene expression from multiple cDNA libraries. *Genome Res* 2000, 10:2055-2061.

Terrat Y, Bonnivard E, Higuet D: GalEa retrotransposons from galatheid squat lobsters (Decapoda, Anomura) define a new clade of Ty1/copia-like elements restricted to aquatic species. *Mol Genet Genomics* 2008, 279:63-73.

Toh H, Hayashida H, Miyata T: Sequence homology between retroviral reverse transcriptase and putative polymerases of hepatitis B virus and cauliflower mosaic virus. *Nature* 1983, 305:827-829.

Toh H, Kikuno R, Hayashida H, Miyata T, Kugimiya W, Inouye S, Yuki S, Saigo K: Close structural resemblance between putative polymerase of a Drosophila transposable genetic element 17.6 and pol gene product of Moloney murine leukaemia virus. *EMBO J* 1985, 4:1267-1272.

Vardi A, Bidle KD, Kwityn C, Hirsh DJ, Thompson SM, Callow JA, Falkowski P, Bowler C: A diatom gene regulating nitric-oxide signaling and susceptibility to diatom-derived aldehydes. *Curr Biol* 2008, 18:895-899.

Vardi A, Formiggini F, Casotti R, De Martino A, Ribalet F, Miralto A, Bowler C: A stress surveillance system based on calcium and nitric oxide in marine diatoms. *PLoS Biol* 2006, 4:e60.

Venter JC, Remington K, Heidelberg JF, Halpern AL, Rusch D, Eisen JA, Wu D, Paulsen I, Nelson KE, Nelson W, et al: Environmental genome shotgun sequencing of the Sargasso Sea. *Science* 2004, 304:66-74.

Vicient CM, Suoniemi A, Anamthawat-Jonsson K, Tanskanen J, Beharav A, Nevo E, Schulman AH: Retrotransposon BARE-1 and Its Role in Genome Evolution in the Genus Hordeum. *Plant Cell* 1999, 11:1769-1784.

Wessler SR: Turned on by stress. Plant retrotransposons. *Curr Biol* 1996, 6:959-961.

Xiong Y, Eickbush TH: Origin and evolution of retroelements based upon their reverse transcriptase sequences. *EMBO J* 1990, 9:3353-3362.

Chapter IV:

Epigenetics in *P. tricornutum*

4.1 Introduction

The modern usage of the word epigenetic refers to heritable traits that do not involve changes to the underlying DNA sequence (the Greek prefix *epi-* in *epigenetics* implies features that are "in addition to" genetics). Among the several layers of control of gene expression, epigenetic regulation is highly dynamic but can be preserved when cells divide. Most modifications only occur within the course of an individual organism's lifetime and over rounds of cell division, but some are inherited transgenerationally. Epigenetic regulation commonly involves the remodeling of chromatin. Chromatin is a combination of DNA, RNA, and protein that makes up chromosomes. In addition to packaging DNA into a small volume to fit in the cell, a primary function of chromatin is to control gene expression and it is organized along the chromosomes in relaxed (euchromatin) or more condensed (heterochromatin) domains that respectively influence gene expression positively or negatively. Highly expressed genes require the looser structure constituted by euchromatin that allows regular access by RNA polymerase and transcription factors. Transposable elements instead are mostly maintained in a transcriptionaly silent mode in compact chromatin structures to hamper their potential harmful effects to an organism.

Chromatin is primarily organized in nucleosomes which consist in a complex of 147 base pairs of DNA wrapped around an octamer of the four core histones (H3, H4, H2A, H2B). The linker histone H1 binds the nucleosome and the entry and exit sites of the DNA, thus locking the DNA into place and allowing the formation of higher order structure. Covalent chemical modifications of both histones and DNA lead to different levels of chromatin compaction through the action of enzymatic activities that reorganize nucleosomes. Core histones are predominantly globular except for their N-termini. These histone ‘‘tails’’, especially lysine residues therein, are the targets of different types of modifications including

methylation, acetylation, and phosphorylation. DNA modification consists of cytosine methylation. Operationally, these modifications function either by disrupting chromatin contacts or by affecting the recruitment of non-histone proteins to the chromatin, and their presence dictates the higher-order chromatin structure in which DNA is packaged.

There are over 60 different residues on histones where modifications have been detected either by specific antibodies or by mass spectrometry (Kouzarides, 2007). Additional complexity comes partly from the fact that lysine residues can be mono-, di-, or trimethylated. This vast array of modifications gives enormous potential for functional responses to signaling conditions within the cell. The identification of the enzymes that catalyze histone modifications has been the focus of intense research over the last 10 years. Enzymes have been identified for histone acetylation (Sterner and Berger, 2000), methylation (Zhang and Reinberg, 2001), phosphorylation (Nowak and Corces, 2004), and so on. Most modifications have been found to be dynamic and the characterization of enzymes that remove these modifications have been the subject of intense research. For example, it was not clear until recently whether histone methylation on lysine (K) was an irreversible epigenetic mark or if an antagonizing reaction exists. Three classes of enzymes capable of such an activity have now been characterized. PAD4 (Peptidylarginine deaminase 4), which was the first to be discovered, is not strictly a histone demethylase but a histone deaminase that converts methyl-arginine to citrulline (Wang et al., 2004). LSD1 (Lysine Specific Demethylase 1) was the first member identified from the second class and directly reverses H3 (K4) and H3 (K9) methylation by an oxidative reaction using flavin cofactor (Shi et al., 2004). The last and largest class of demethylases contain the Jumonji C domain (JmjC) and catalyses histone lysine demethylation through an oxidative reaction that requires Fe (II) and α-ketoglutarate as cofactors (Jung et al., 2005). In this chapter I have assessed the presence of genes encoding

proteins involved in chromatin-based mechanisms in the P. tricornutum genome as well as the existence of the related processes.

4.2 Results and Discussion

4.2.1 Histone modifications

4.2.1.1 Histone Modifiers

The *P. tricornutum* genome was screened manually for the presence of genes encoding putative histone modifiers and a rich repertoire of such proteins was found (Table 4.1). Enzymes responsible for most types of modifications assessed were found, except for proline isomerization. We identified 8 different proteins putatively involved in the acetylation of different histone lysine residues and as many as 16 genes encoding candidate proteins with the reverse function. Several homologs of histone methyltransferases were also identified. In addition, we found a gene model encoding a protein containing a SET domain and Jumonji C (JmjC) domain (identified by Kamel Jabbari). This fusion is particularly interesting because it associates two domains with antagonistic activities: the SET domain known to mediate methylation on lysine and the JmjC domain which mediates lysine demethylation. Therefore this chimeric protein may have a dual activity, maybe on different residues such as demethylating H3 (K9) and methylating H3 (K4) from the same nucleosome. The existence of transcripts corresponding to this fusion was confirmed by PCR on *P. tricornutum* cDNA. Intriguingly, such a SET-JmjC fusion is also found within the fungus *Neurospora crassa* but in the opposite orientation (JmjC-SET). We also found several other candidates for histone lysine demethylation. Although no PADI4 homolog was detected, we identified an LSD1 homolog, as well as various JmjC-containing proteins with significant similarity to known histone demethylases. Besides the SET-JmjC fusion, two other fusions involving the JmjC domain were identified: JmjC-CHROMO and JmjC-MBT. No protein with similar composition could be found in other eukaryotes except in *T. pseudonana*. Both fusions are of particular interest because they associate the JmjC lysine demethylase domain with domains

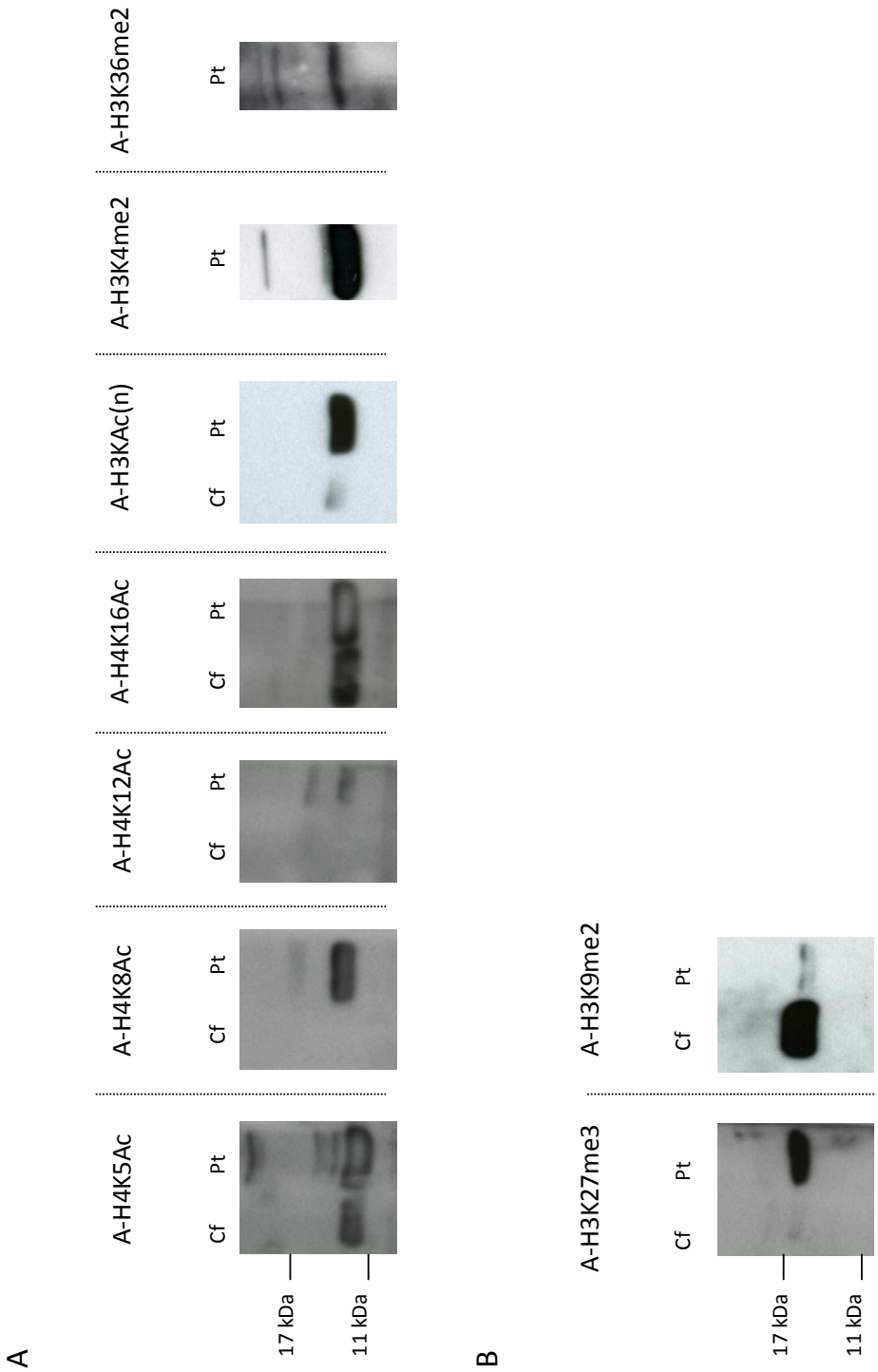
that have been shown to interact with chromatin: CHROMO and MBT. Indeed, the CHROMO domain has been shown to interact with mono-, di-, and trimethylated H3 (K9) (Flanagan et al., 2005). Further, it was recently shown in fission yeast that the CHROMO domain-containing protein Swi6/HP1 interacts with the JmjC-domain-only protein Epe1 (Zofall et al., 2006). MBT domains have been shown to bind monomethylated H3 (K4) and different degrees of methylation on H4 (K20) depending on the MBT domain tested (Kim et al., 2006). Furthermore, it was recently shown that the Drosophila MBT domain protein L3MBTL1 binds at least two nucleosomes and can compact nucleosomal arrays dependent on mono- and dimethylation of histone H4 lysine 20 (Trojer et al., 2007). Putative histone arginine (R) methyltranserases, serine/threonine kinases, and ubiquitylases were also identified in the *P. tricornutum* genome. All together these data suggest the existence of complex epigenetic regulations through the specific modification of histones in *P. tricornutum*.

In order to confirm the presence of these histone modifications in *P. tricornutum*, I used different antibodies that were available in the laboratory to probe various histone modifications commonly associated with eu- or heterochromatin by western blot. Several modified histones typically associated with open chromatin such as dimethylated H3 (K4), dimethylated H3 (K36), and poly-acetylated H3 (R) were detected (Figure 4.1). I also detected modified histones typically associated with condensed chromatin such as dimethylated H3 (K9) and trimethylated H3 (K27) (Figure 4.1). Thus, the *P. tricornutum* chromosomes are likely made of different domains that are more or less prone to transcription. TEs, for example, are likely to be typified by heterochromatic marks, as is the case in most eukaryotes.

Table 4.1:

Histone Modifiers	Residues Modified	Homolog in *P. tricornutum*
Lysine Acetyltransferases (KATs)		
HAT1 (KAT1)	H4 (K5, K12)	54343
GCN5 (KAT2)	H3 (K9, K14, K18, K23, K36)	46915
Nejire (KAT3)	H3 (K14, K18) H4 (K5, K8)	45703
CBP/P300 (KAT3A/B)	H2A (K5) H2B (K12, K15)	45764, 54505
MYST1 (KAT8)	H4 (K16)	24733, 24393
ELP3 (KAT9)	H3	50848
Lysine Deacetylases (HDACs)		
RDP3 (Class I HDACS)	-	46592, 24019
HDA1 (Class II HDACS)	-	35869, 45906, 51026, 49800
NAD+ dependant (Class III HDACS)	H4 (K16)	16859, 45909, 52135, 45850, 24866, 45909, 52718, 21543, 39523, 45850
Lysine Methyltransferases		
MLL	H3 (K4)	40183, 54436, 42693, 47328, 49473, 49476
ASH1	H3 (K4)	50375?
E(Z)	H3 (K9, K27)	32817
SET+JmjC	-	bd1647
Lysine Demethylases (KDM)		
LSD1 (KDM1)	H3 (K4, K9)	51708, 44106
FBXL (KDM2)	H3 (K36)	42595
JMJD2 (KDM4)	H3 (K9, K36)	48747
JMJ-MBT	-	48109
JMJ-CHROMO	-	Pt1-40322
Arginine Methyltransferases		
CARM1	H3 (R2, R17, R36)	17184, 54710
PRMT5	H3 (R8), H4 (R3)	49565
Serine/Threonine Kinases		
Aurora	H3 (S10)	35463
Ubiquitilases		
RING20	H2B (K120)	50465

Figure 4.1: Pictures of western blot experiments revealing the presence of several modified histones in *P. tricornutum* (Pt) proteome. I some cases, we used a cauliflower (Cf) protein extract as a positive control. (A) Histone modifications commonly associated with euchromatin. (B) Histone modifications commonly associated with heterochromatin.



4.2.1.2 Effectors

Many chromatin modifications can serve as recognition sites for the recruitment of effector proteins that read basal modifications induced by biological messages (Kouzarides, 2007). Several distinct binding modules have been identified in various nuclear proteins, coupling a particular histone modification with cognate effector proteins. Effector proteins may alter chromatin structure by binding two or more nucleosomes as found with Heterochromatin Protein 1 (HP1) and MBT group proteins (Ruthenburg et al., 2007). This phenomenon likely occurs through multimerization and whether it connects adjacent or remote nucleosomes remains an open question. The CHROMO domain protein HP1 is a fundamental unit of heterochromatin packaging that is typically enriched at the centromeres and telomeres of nearly all eukaryotic chromosomes, with the notable exception of budding yeast. Members of the HP1 family are characterized by two conserved domains: an N-terminal chromodomain and a C-terminal chromoshadow domain. Intriguingly, the only *P. tricornutum* protein containing a CHROMO domain with similarity to HP1 is the JmjC-CHROMO fusion mentioned above. As well, the sole *P. tricornutum* protein with an MBT domain is the JmjC-MBT fusion. The TUDOR domain is also often found in proteins with histone modifier domains and has been shown to recognize dimethylated H4 (K20) as well as other histone modifications (Huang et al., 2006). Other proteins with TUDOR domains have been found to be involved in RNA binding. Interestingly, we found in both diatom genomes a protein in which a TUDOR domain is fused to a linker histone H1 domain, in addition to copies similar to regular H1. These represent the only candidates for TUDOR-containing chromatin components in diatom genomes. All together, these observations suggest that diatoms have evolved lineage-specific fusions involving chromatin modifiers and effector domains that likely reflect diatom-specific regulation of chromatin dynamics.

4.2.1.3 Chromatin Extraction and Immunoprecipitation

A key step towards understanding the function of epigenetic marks consists in the identification of their target loci along the chromosomes. Conversely, deciphering the different aspects of transcriptional regulation of a specific gene involves the identification of the different players that constitute and bind nucleosomes and DNA such as TFs and modified histones. Chromatin immunoprecipitation (ChIP) is a technique used to identify specific DNA sequences that are bound *in vivo* to proteins of interest. It involves formaldehyde fixation of chromatin to crosslink the DNA-binding proteins to the DNA. After shearing the DNA into small fragments, specific DNA-protein complexes are isolated by immunoprecipitation with protein-specific antibodies. The DNA isolated from the complex can then be used to quantify loci of interest by quantitative PCR. Alternatively, genome-wide studies can be addressed by hybridizing the immunoprecipitated material onto a genomic array (ChIP on chip) or by deep sequencing (ChIP-Seq). ChIP is routinely used for plants, fungi, and animals. Adapting the ChIP protocol to diatoms is therefore a key step towards the study of epigenetic phenomena. For example, identification of the targets of some of the fusions mentioned above may help to elucidate their functions.

To apply ChIP to *P. tricornutum*, cell lysis and chromatin shearing was achieved successfully using a buffer commonly used for ChIP on mice neurons (Figure 4.2) (see Materials and Methods). Setting up a ChIP protocol may be facilitated by using antibodies against proteins that directly bind DNA and have abundant DNA targets. We therefore used an antibody raised against RNA polymerase II for initial immunoprecipitation trials because it directly binds all transcribed genes in a genome. To control the efficiency of the procedure, we used the immunoprecipitated material for quantitative PCR with primers corresponding to transcribed genes and TE-specific loci as positive and negative controls, respectively (Figure 4.3). After validation of the method used, we tried to apply it to dimethylated H3 (K9) which

is a less abundant chromatin-associated protein usually associated with silenced regions such as TEs and absent from highly transcribed regions. As expected, quantitative PCR using the same gene-specific and TE-specific primer pairs gave results opposite to the ones obtained using the anti-polymerase II antibody (Figure 4.3). These results suggest that the method used is ready to be applied to diatoms to address specific questions.

Figure 4.2: Pictures of agarose gels used to assess the efficiency of chromatin extraction and sonication. (A) Profiles of *P. tricornutum* chromatin extracted using two yeast and mouse ChIP buffers and after mild sonication. (B) Chromatin profile after successive sonication cycles. (C) Chromatin profile after mild electrophoresis.

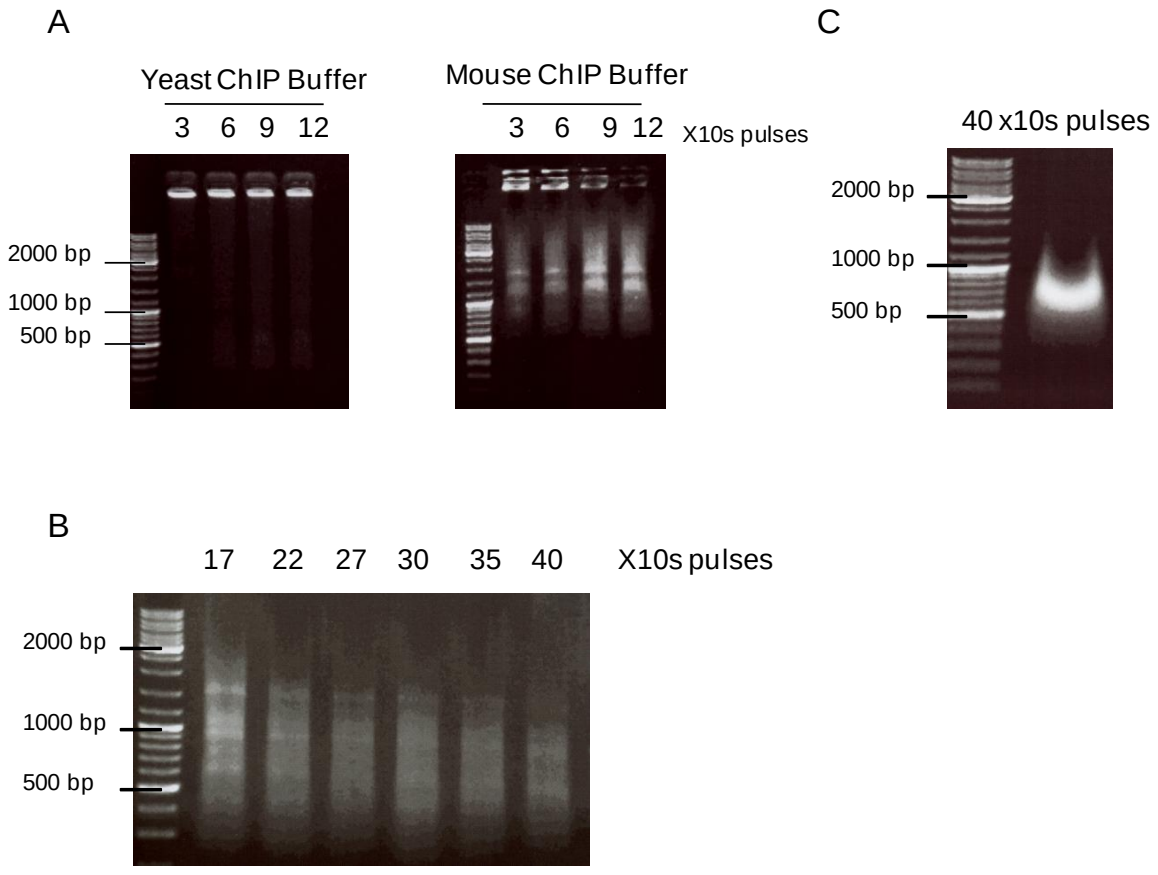
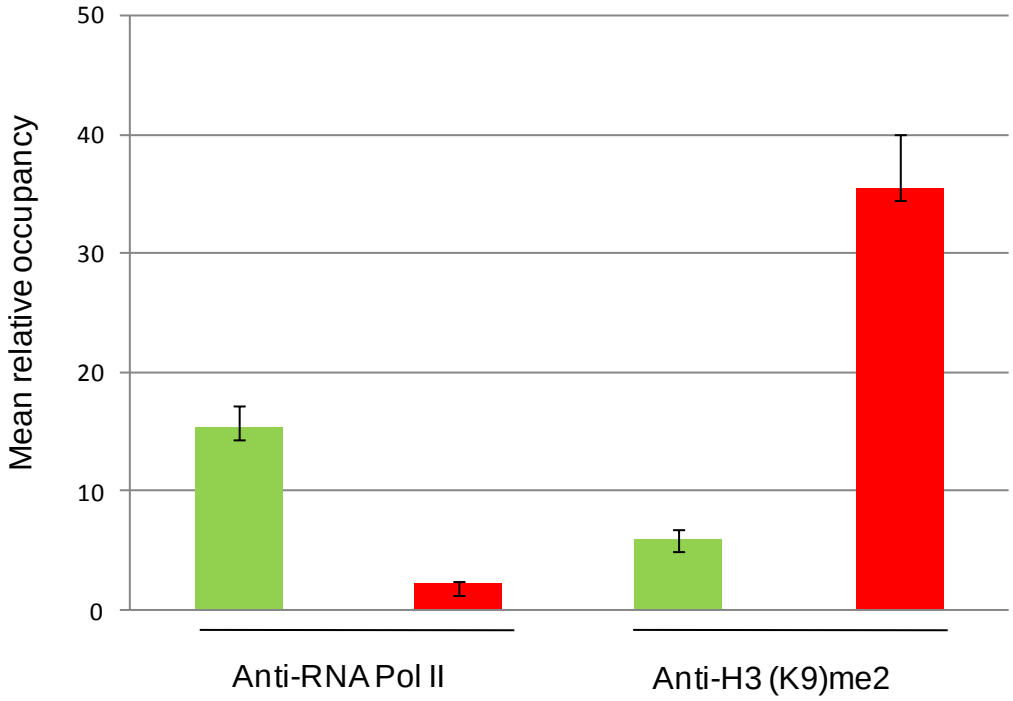


Figure 4.3: Histogram presenting the quantitative PCR results obtained using immunoprecipitated material. The bars represent the mean relative occupancy for 6 genes (green) and 6 TE loci (red) using anti-RNA polymerase II and anti-dimethylated H3 (K9) antibodies. Each bar represents the average of two experiments and error bars indicate the standard deviation between both replicates.



4.2.2 DNA methylation

Adenine or cytosine methylation is an ancestral phenomenon that is part of the restriction/modification system of many bacteria, in which DNA is methylated periodically throughout the genome. When introduced into the cell, foreign DNA that is not methylated in this manner is degraded by sequence-specific restriction enzymes. Therefore, the methylation of native DNA acts as a primitive immune system, allowing the bacteria to protect themselves from infection by bacteriophage. In eukaryotes, DNA methylation occurs on cytosine, specifically on the number 5 carbon (C5) of the cytosine pyrimidine ring. It can be copied after DNA replication and inherited, and subsequently removed without changing the original DNA sequence and it is therefore considered an epigenetic phenomena. Functionally, DNA methylation may impact gene transcription by impeding the binding of transcriptional regulators or by recruiting proteins with methyl-CpG-binding domains (MBD). MBD proteins may then recruit additional histone modifiers, effectors, and other chromatin remodelling complexes to form compact and silent chromatin. Functionally, two major types of activities can be distinguished among C5-MTases, namely “maintenance” and “de novo” methylation (Riggs, 1975; Holliday and Pugh, 1975). Maintenance methylation occurs after DNA replication on hemimethylated symmetric motifs (CpG and CpNpG), whereas methylation that occurs at previously unmethylated cytosines (in symmetrical or asymmetrical sites) is known as de novo methylation.

In mammals, DNA methylation is essential for normal development as indicated by the death of mouse embryos with defective DNA methylation (Li et al., 1992), and is associated with a number of key processes. CpG islands (CGIs) are discrete CpG-rich regions present in the promoters of 50%–70% of human genes. Methylated CpG islands are strongly and heritably repressed, and used to ensure mono-allelic expression of imprinted genes (Ferguson-Smith and Surani, 2001) and X-chromosome inactivation (Jones and Takai, 2001),

silencing of parasitic DNA, and tissue-specific gene silencing (Shen et al., 2007). Many human diseases, including cancer, are also caused and impacted by abnormal methylation (Jones and Laird, 1999). In plants as well, cytosine DNA methylation is important for the silencing of repeated sequences and genomic imprinting (Nowack et al., 2007). In addition, recent genome-wide studies in *Arabidopsis thaliana* have revealed that many endogenous genes are also methylated either within their promoters or within their transcribed regions, and that gene methylation is highly correlated with transcription levels (Zilberman et al., 2007). DNA methylation is also well documented in the filamentous fungi *Neurospora crassa* and *Ascobolus immersus* and is associated with two mechanisms named Repeat-Induced Point mutation (RIP) and Methylation Induced Premeiotically (MIP), respectively. In these species, methylation resides in the asymmetrical sites (CpN) located in the repeated sequences and was shown to protect their genomes against the deleterious effect of repeated sequences (Selker, 1997). The repeated sequences are identified and methylated de novo during the sexual phase and the methylation is maintained through vegetative development. Methylation of the repeated elements reduces recombination rates in *A. immersus* (Maloisel and Rossignol 1998), and decreases the similarity between repeated sequences in *N. crassa* (RIP) (Selker et al. 2003).

DNA methylation has not yet been described for other eukaryotic lineages including stramenopiles. The only data about DNA methylation in diatoms was published in 1992 by Eric Jarvis who used HPLC to analyze DNA nucleoside composition in several species of microalgae (Jarvis et al., 1992). The authors described generally low levels of methylation in the diatoms studied including *P. tricornutum*. More than fifteen years later, I reevaluated this phenomenon in diatoms during my PhD.

4.2.2.1 DNA methyltransferases in diatoms

DNA (cytosine-5) methyltransferases (C5-MTases) catalyze the transfer of a methyl group from S-adenosyl-l-methionine to the C5 position of cytosine residues in DNA. In animals and fungi, DNA methylation typically occurs in the symmetrical CpG context (Jaenisch and Bird, 2003). In contrast, plant cytosines are methylated in symmetrical (CpG and CpNpG) and asymmetrical (CpNpNp) contexts, where N can be any nucleotide but guanine. Most insights about the mechanism of this enzymatic reaction have come from studies of bacterial C5-MTases (Cheng and Roberts, 2001). All MTases share a catalytic domain containing 10 conserved small motifs, 6 of which being strongly conserved and suggesting a common origin (Kumar et al. 1994). Eukaryotic C5-MTases are larger proteins, but the sequence homology they share with bacterial enzymes and experimental data suggest that they act using the same catalytic mechanism.

A number of eukaryotic C5-MTases have been identified by sequence similarity since the first was reported in 1988 (Bestor et al., 1988). Based on sequence similarity within their C-terminal catalytic domains, the known eukaryotic C5-MTases had been grouped into several classes, represented by the mammalian DNMT1, DNMT2, DNMT3a/3b and the chromomethylases (CMTs) (Chen and Li, 2004; Goll, 2005). However the DNMT2 family, which is found in most eukaryotes, was recently shown to methylate tRNA instead of DNA (Goll et al., 2006). Members of DNMT1 and DNMT3 families are found in the genome of most eukaryotes (although fungal relatives show greater divergence). In contrast, CMTs are specific to flowering plants. In addition, rare cases of organisms encoding MTases that do not fall in these families have been described (see below). A prominent feature of these enzymes is the presence of N-terminal extensions in addition to the catalytic domains that contain the highly conserved C5-MTase motifs and the variable nature of these extensions is believed to be the structural basis for the differences in biological functions. So far, all DNMT1 ortholog

tested have been shown to be responsible for maintenance activity, while the DNMT3 and CMT families were reported to have de novo methylation activity.

I performed an in silico screening of the *P. tricornutum* genome for the presence of putative C5-MTases. Four gene models were found to contain a domain with significant similarity with the DNA methyltransferase domain. Surprisingly, the first one, Pt2_ 45072, presents sequence homology with proteins from fungi and green algae that are not annotated or annotated as "DNA repair protein RAD8" or "SNF2 family helicase". We did a scan for protein domains in these sequences and found that they contain a Snf2-type DEXDc/HELICc helicase domain in their C-terminus as well as a RING-type zinc finger, but in addition they display a DNA methyltransferase domain at their C-terminus as well. We subsequently observed that the gene model flanking Pt2_ 45072 contains a helicase domain, suggesting that both models in fact represent a single gene. In support of this, we found an EST covering part of the DNA sequence separating these two models using BLAST search against the *P. tricornutum* cDNA database. I looked further in the literature to find description of similar structures and found that such domain association had been reported in the fungi *A. nidulans*, *A. fumigatus* (Ascomycetes) and *C. neoformans* (Basidiomycetes) in a study on the evolutionary diversification of DNA methyltransferases in eukaryotic genomes (Ponger and Li., 2005).

Here, we found by BLAST that such proteins are also present in *T. pseudonana*, the Phaeophyceae *Aureococcus anophagefferens*, Prasinophyceae *Ostreococcus tauri*, *Ostreococcus lucimarinus*, and two sequenced *Micromonas pusilla* ecotypes (CCMP1545 and RCC299) as well as in the genome of various ascomycetes and basidiomycetes. These proteins were not found in the genomes of green algae such as *Chlamydomonas reinhardtii* and *Chlorella sp*. nor in the oomycetes *Phythophtora sojae* and *P. ramorum* and the brown alga *Ectocarpus siliculosus*. All proteins collected were larger than 2,000 amino acids. A

phylogenetic analysis of their DNA methyltransferase domain revealed that these sequences share a common ancestry and cluster in a lineage distinct to the DNMT1, DNMT2, DNMT3, and CMT lineages (Figure 4.4). We propose to name this lineage the HDMT family after helicase domain containing DNA methyltransferases. The patchy distribution of the HDMT family across eukaryotic lineages is intriguing. The phylogeny of HDMTs does not respect the known evolutionary relationships between the different lineages involved and suggests that HDMTs have been the subject of repeated horizontal transfers (Figure 4.5). For example, the DNA MTase domain of *T. pseudonana* HDMT presents 42% and 97% identity with its homolog in *P. tricornutum* and *A. anophagefferrens*, respectively suggesting a gene transfer between a centric diatom and some close Aureococcus ancestor. The clustering of *Ostreococcus* and *Micromonas* species suggests the presence of HDMTs in their Prasinophyceae ancestor.

Figure 4.4: Phylogenetic tree constructed with the Neighbor-Joining method showing the relationships between the DNA methyltransferase domains of different families of C5-MTases: bacterial C5-MTases, DNMT1, DNMT2, DNMT3, and HDMTs. Sequences presenting unexpected phylogeny are displayed with taxon-specific color code: diatoms (orange), green algae (green), brown algae (yellow), and fungi (blue). Abbreviations : *P. tricornutum* (Pt), *T. pseudonana* (Tp), *M. pusilla* (Mp), *O. tauri* (Ot), and *O. lucimarinus* (Ol). Bootstraps of the main branches are shown (score in % over 1000 iterations).

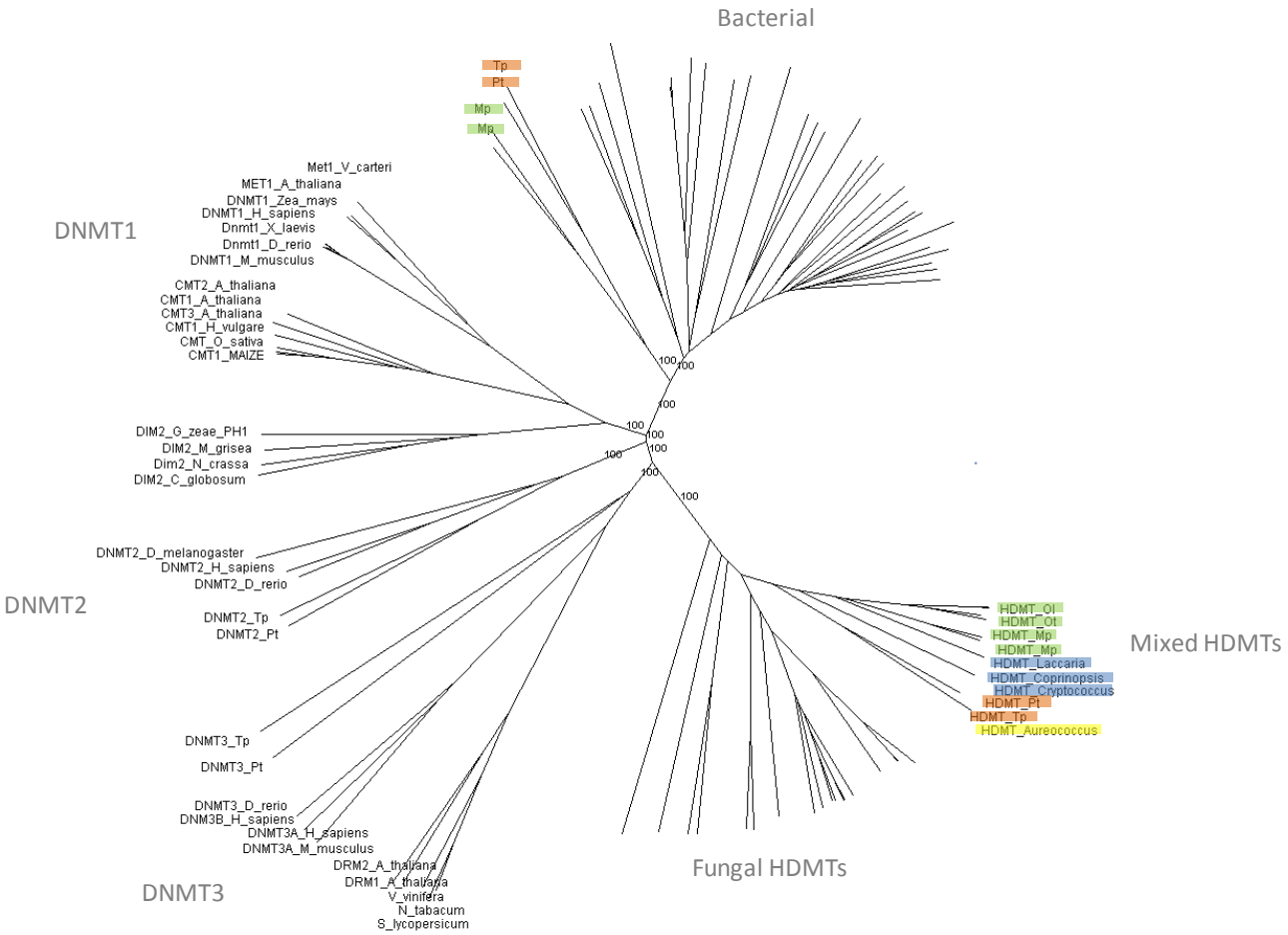
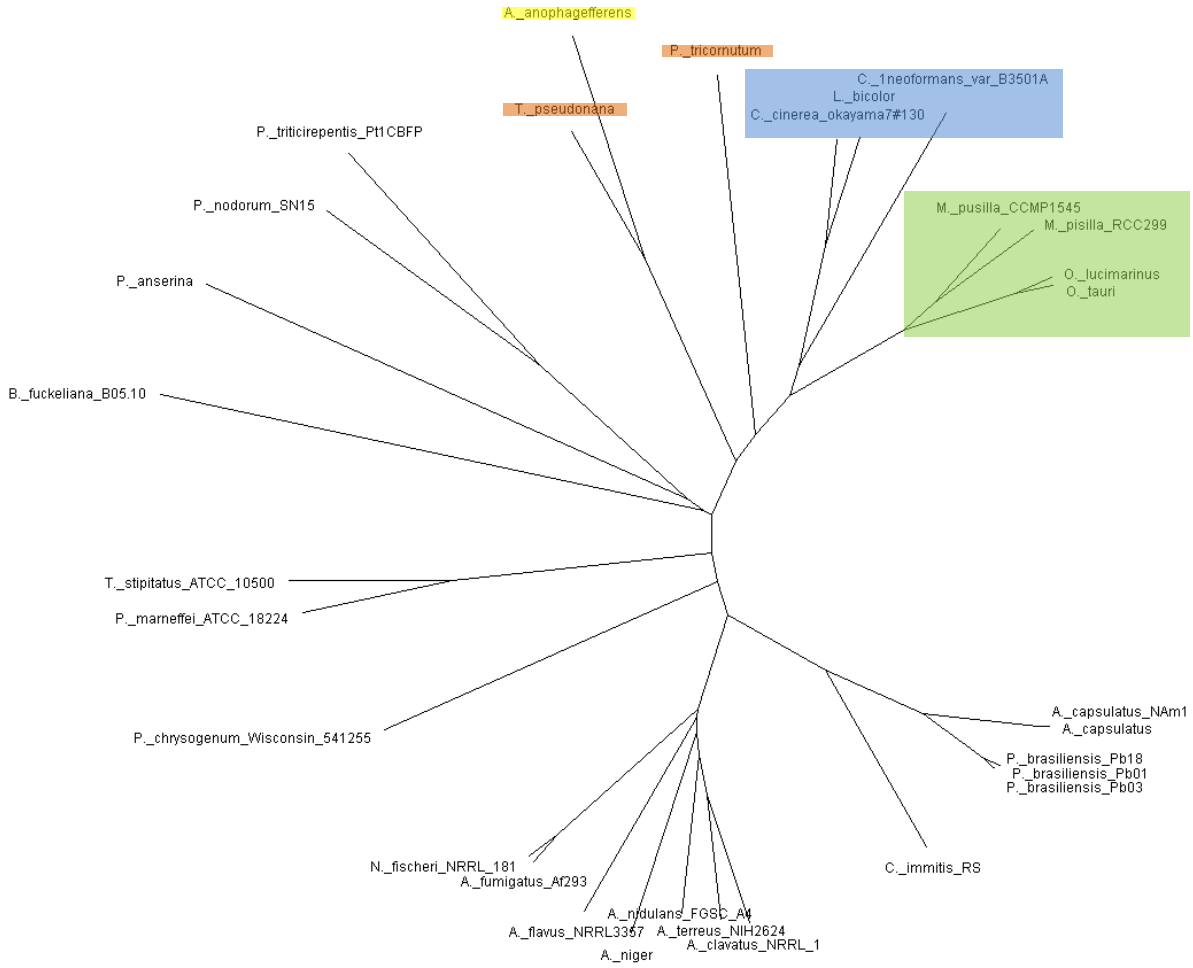


Figure 4.5: Bootstrap supported phylogenetic tree constructed with the Neighbor-Joining method presenting the relationships between the DNA methyltransferase domains of the HDMTs identified in this work. Sequences presenting unexpected phylogeny are displayed with taxon-specific color code as in Figure 4.4.



Interestingly, some members of the SWI2/SNF2 family of ATP-dependent chromatin-remodeling factors have been shown to be important for cytosine methylation. The Helicase DDM1 (Decreased DNA Methylation 1) was first identified in a forward genetic screen for genes that affect global cytosine methylation patterns in *A. thaliana*. Plant DDM1 mutants for show a 70% decrease in methylation (primarily at repetitive DNA), but the mutant plants are viable and fertile (Miura et al., 2001). Similarly, abolishment in lymphoid-specific helicase (Lsh), which is the mammalian homolog of DDM1, in mice perturbs DNA methylation patterns with a 50% reduction in cytosine methylation at repetitive sequences (Jeddeloh et al., 1999). Lsh was further shown to directly interact and cooperate with the C5-MTases DNMT1 and DNMT3B and with the histone deacetylases HDAC1 and HDAC2 to silence transcription (Myant and Stancheva, 2008). In addition, mutations in the SWI2/SNF2-type helicase ATRX (X-encoded alpha thalassemia mental retardation) leads to changes in the pattern of methylation of several highly repeated sequences including the rDNA arrays and subtelomeric repeats (Gibbons et al., 2000). The direct involvement of certain SWI2/SNF2 helicases in DNA methylation processes suggest that HDMT-type C5-MTases play a specific role in the organisms where they are found.

Phylogenetic analysis of the three other *P. tricornutum* proteins with DNA methyltransferase domains suggest that Pt2_ 16674 and Pt2_ 46156 belong to the DNMT2 and DNMT3 families, respectively, and that each has a homolog in the *T. pseudonana* genome (Figure 4.4). Unexpectedly, the last putative C5-MTase in *P. tricornutum*, of which a homolog is also found in *T. pseudonana*, appears to be more closely related to the bacterial C5-MTases (Figure 4.4) although it clusters with sequences from *Micromonas* species collected after BLAST comparison against the NCBI database. This phylogeny is also atypical and cannot be explained parsimoniously by common ancestry and subsequent loss in almost all eukaryotic species sequenced. It may therefore reflect a horizontal gene transfer

between eukaryotic genomes subsequent to a transfer from bacteria. The inter-eukaryotic transfer has likely been directed from an ancestral diatom towards the green algae since this protein is found in both centric and pennate diatoms, that diverged around 90 Mya. As a consequence, the initial transfer towards eukaryotes likely involved a common diatom ancestor. The presence of C5-MTases of bacterial origin in diatoms supposes their functional adaptation to eukaryotic genomes into purposeful biological roles. Finally, we could not identify any candidate gene model for a putative DNMT1 homolog in diatom genomes, which is also the case in numerous other species (Ponger and Li., 2005). We were also unable to identify any genes encoding proteins with methyl-CpG-binding domains (MBDs) in the diatom genomes.

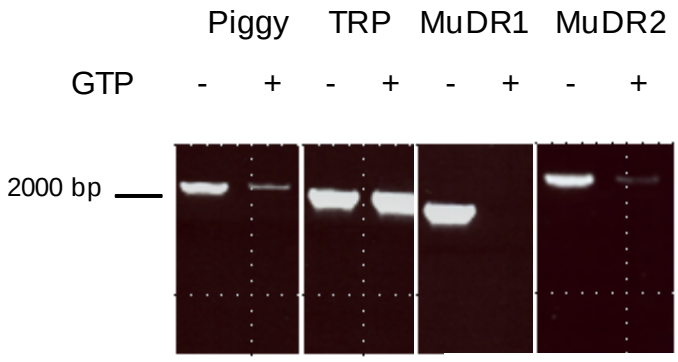
4.2.2.2 DNA methylation in *P. tricornutum*

We have seen that the *P. tricornutum* genome encodes diverse C5-MTases. DNA methylation can have different patterns and be associated with various functions in different organisms, although other organisms evolved without it such as *Caenorhabditis elegans*. What is the function of DNA methylation in *P. tricornutum*? A key step towards answering this question consists in establishing the pattern of 5-methyl cytosine (5mC) across the genome. As mentioned above, a common feature of DNA methylation pattern and function across eukaryotes is the mark of DNA sequences corresponding to parasitic DNA such as transposable elements and to serve the genome by limiting their potentially harmful effects due to mobilization and recombination.

To study DNA methylation in *P. tricornutum* we thus first addressed whether transposable elements were the target of such marks. As mentioned above, all the LTR retrotransposons tested using the Mcr-PCR method were observed to be methylated (Chapter 2). We have shown that the *Blackbeard* locus was hypomethylated in response to nitrate

starvation. Further, we tested loci corresponding to different class II TEs in the *P. tricornutum* genome and found that DNA from PiggyBac-like elements as well as the two families of Mutator-like elements were also methylated (Figure 4.6). In this assay, however, DNA sequence of the *P. tricornutum* transposase-like element was found to be free of methylation (note that these lack terminal inverted repeats). The fact that most TEs appear to be methylated in the *P. tricornutum* genome was not unexpected. But is DNA methylation restricted to TEs in this species? For example, in *A. thaliana* repeated sequences and transposable elements (TEs) are densely methylated but one third of the genes also have low but detectable methylation in their transcribed regions. What is the DNA methylation pattern in *P. tricornutum*? To challenge this question we decided to address a genome-wide analysis of DNA methylation in diatoms.

Figure 4.6: Picture of an agarose gel showing the products obtained by PCR on different families of *P. tricornutum* TEs after McrBC digestion in the presence or absence of GTP.

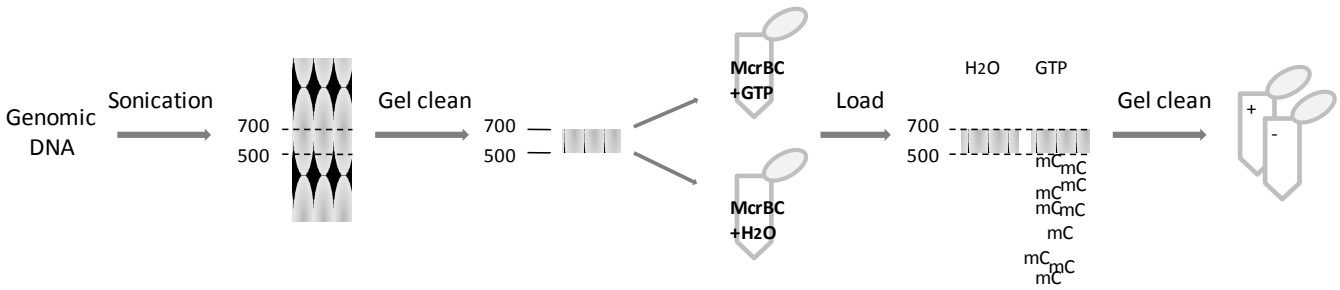


Such a study can be undertaken through different approaches. The first one, called Me-DIP (Methylated-DNA ImmunoPrecipitation), is an immunocapturing approach to enrich DNA that is methylated using an antibody that specifically recognizes 5mC. Differentially (antibody/control) precipitated DNA can then be identified by deep sequencing or hybridization on a whole-genome tiling array, as was done to uncover genome-wide DNA methylation pattern in *A. thaliana* (Zilberman et al., 2007). Genomic tiling microarrays, which represent contiguous stretches of chromosomes without bias towards coding sequences, allow DNA methylation of all sequence types to be assayed simultaneously at high resolution. The second method, called bisulfite sequencing, involves the treatment of DNA with sodium bisulfite that converts cytosine residues to uracil, but leaves 5mC residues unaffected. Upon amplification by PCR, the DNA polymerase inserts a thymine as complementary base of uracil. Sequencing of PCR products thus enables to identify 5mCs at the single-base resolution, as applied recently to the *A. thaliana* genome (Cokus et al., 2008). Third, genome-wide methylation patterns can be obtained using the restriction enzyme McrBC to selectively exclude the methylated fraction of DNA from a genome of interest. The sample is then compared to untreated DNA using a tiling array hybridization (Lippman et al., 2005). This method was recently used to identify differences in DNA methylation patterns on chromosome 4 from different *A. thaliana* ecopytes (Vaughn et al., 2007). We decided to use this last approach to address genome-wide DNA methylation patterns in the *P. tricornutum* genome. This aim constitutes a pioneering work on whole genome DNA methylation pattern since genome-wide data is only available for *A. thaliana* until now.

In order to achieve this, two requirements need to be addressed: to exclude the methylated DNA fraction efficiently from genomic DNA samples, and to obtain the whole *P. tricornutum* genome tiled on a microarray. To attain the first requirement, we adapted a protocol that was published by Lippman et al. (2005) with modifications referred to as

Window McrBC Restriction (WMR) (Figure 4.7). With WMR, DNA is sonicated to an average size of around 500-700 bp and subsequently size-fractionated by electrophoresis in an agarose gel. A window corresponding to 500-700 bp fragments is cut from the gel and cleaned using a column. The eluted DNA is split into two tubes with all reagents for McrBC digestion, except that McrBC-cofactor (GTP) is not added in the "control" tube but is replaced by water. After incubation, both samples are loaded on an agarose gel. At this step, DNA fragments that were cut by McrBC will migrate out of the 500-700 bp window. A mild smear of digested DNA down the window of the treated sample and not with the control suggests that part of the material was excluded from the window (Figure 4.8A). Finally, windows from both samples are recovered and cleaned (Figure 4.7).

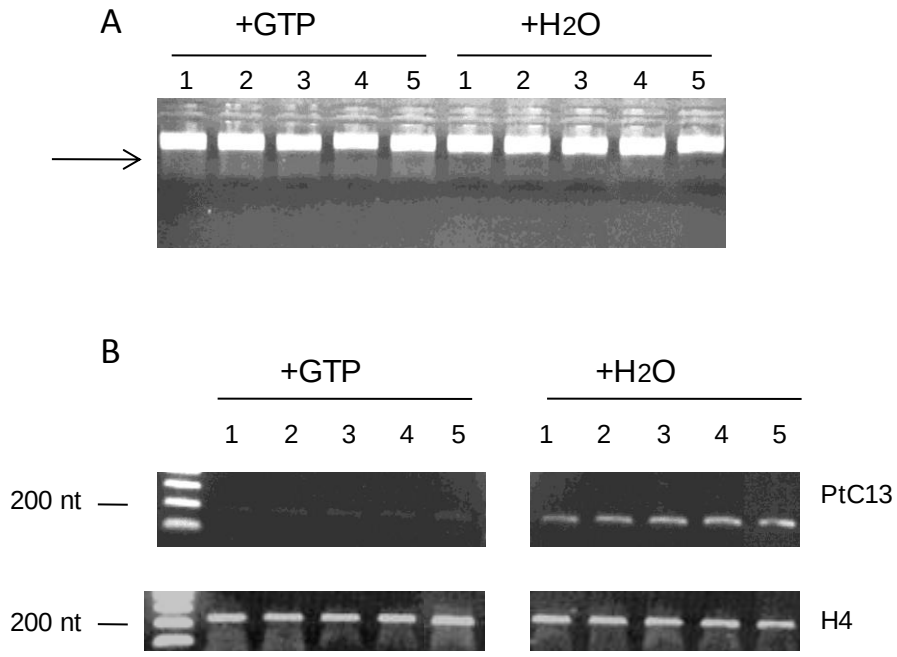
Figure 4.7: Schematic representation of the main steps of the Window McrBC Restriction (WMR) protocol.



In parallel, we started a collaboration with Dr. Michael Sussman and Dr. Franco Cerrina who invented the Maskless Array Synthesizer (MAS) technology and founded the NimbleGen Systems (Roche) company that manufactures custom tiled DNA arrays in a few hours by eliminating the requirement for expensive masks used in traditional DNA chip technology (Singh-Gasson et al., 1999). This technology was used for the design of a *P. tricornutum* tiled array designed for a coverage of one 36mer probe per 45 bases, for both strands. In total, over 1,200,000 oligos were synthesized on an ultra-high density (HD) array (2.1 million features). The empty features on the array were filled with duplicates and random probes.

We prepared WMR samples and their negative control from 5 biological replicates of *P. tricornutum* cultures (Figure 4.8A) and verified by semi-quantitative PCR that McrBC treated samples were depleted of the known methylated fraction (TEs), while amounts of unmethylated loci remained constant (Figure 4.8B). Replicate samples were shipped to Sussman's team in the University of Madison, Wisconsin, where they were subjected to whole-genome amplification to increase amounts of DNA. Samples were labeled using Cy5 or Cy3 dyes, pooled, and hybridized onto the HD array. This was done independently for three of the biological replicates (hyb1, hyb2, and hyb3). The array was then scannd using a 2µm resolution scanner and peaks were identified by dividing Cy3 intensity by Cy5 intensity (control over McrBC-treated). Since at methylated positions McrBC-treated DNA will have less signal, dividing the two channels in this way yields positive ratios.

Figure 4.8: Picture of agarose gels used to assess the efficiency of WMR experiment on 5 biological replicates. (A) DNA windows after McrBC digestion in the presence or absence of GTP. The target indicates the size at which a DNA smear appears. (B) Results of PCR amplification using DNA from 5 biological WMR replicates and negative controls as template and primers corresponding to the PtC13 element and histone H4.



Preliminary analysis of hyb1 suggests that the experiment worked successfully. Indeed, we find that loci corresponding to TEs show positive ratios clearly above background (Figure 4.9). Complete statistical analysis taking into account data from the three replicates, the random probes, and the duplicated probes is being undertaken at the University of Madison. Putative methylated genes will then be analyzed by bisulfite sequencing to confirm their methylation status and to determine sequence context (CpG, CpNpG, or CpNpN). In order to identify manually some putatively methylated loci that do not correspond to TEs, I tried to elaborate a background level for hyb1. For this, I plotted all peak values from hyb1 (Figure 4.10). It appears from this plot that significant peaks may have a value >10. I designed primers for 20 regions that correspond to genes and display peak values >10. Semi-quantitative Mcr-PCR on these genes shows that at least two of them, Pt2_9233, and Pt2_31636, are indeed methylated (Figure 4.11). These correspond to a putative amino acid transporter and a gene of unknown function, respectively. These encouraging results provide the first proof that some genes are also methylated in the *P. tricornutum* genome and promises that interesting results will be obtained from whole-genome analysis.

Figure 4.9 (next page): Preliminary results from hybridization 1 of the WMR tiling experiment. (A) Profile of the peaks obtained at the *Blackbeard* locus (top) whith BKB position shown schematically (bottom). (B) Profile of the peaks obtained for the first 70 kb of P. tricornutum chromosome 1 (up) with scaled TEs annotation track (bottom) (ClassI TEs in brown and ClassII TEs in red).

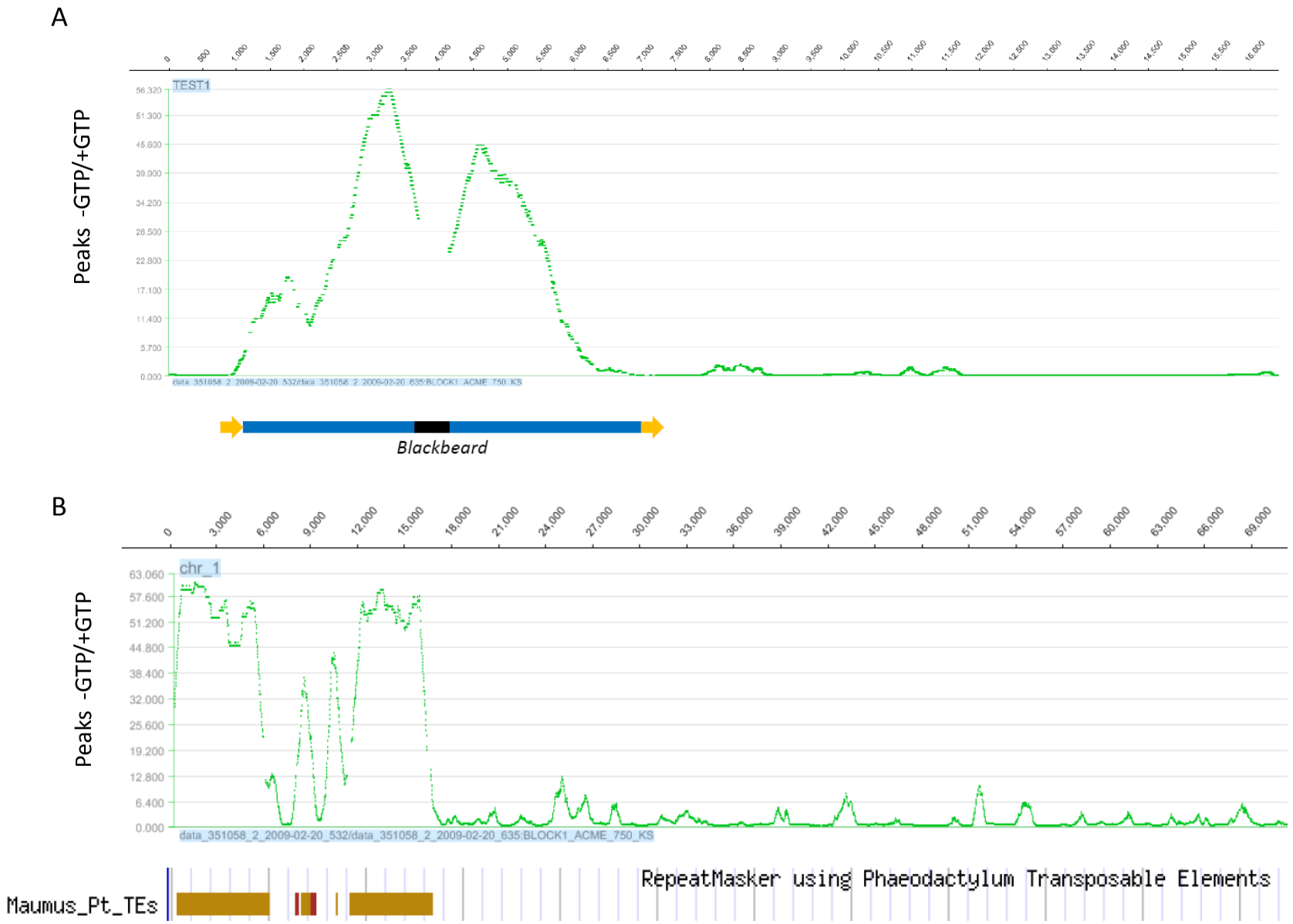

A
Peaks -GTP/+GTP
TEST1
Blackbeard
B
Peaks -GTP/+GTP
chr_1
data_351058_2_2009-02-20_532/data_351058_2_2009-02-20_635:BLOCK1_ACME_750_KS
Maumus_Pt_TEs
RepeatMasker using Phaeodactylum Transposable Elements

Figure 4.10: Plot of the value of the peaks obtained from hybridization 1 of the WMR tiling experiment.

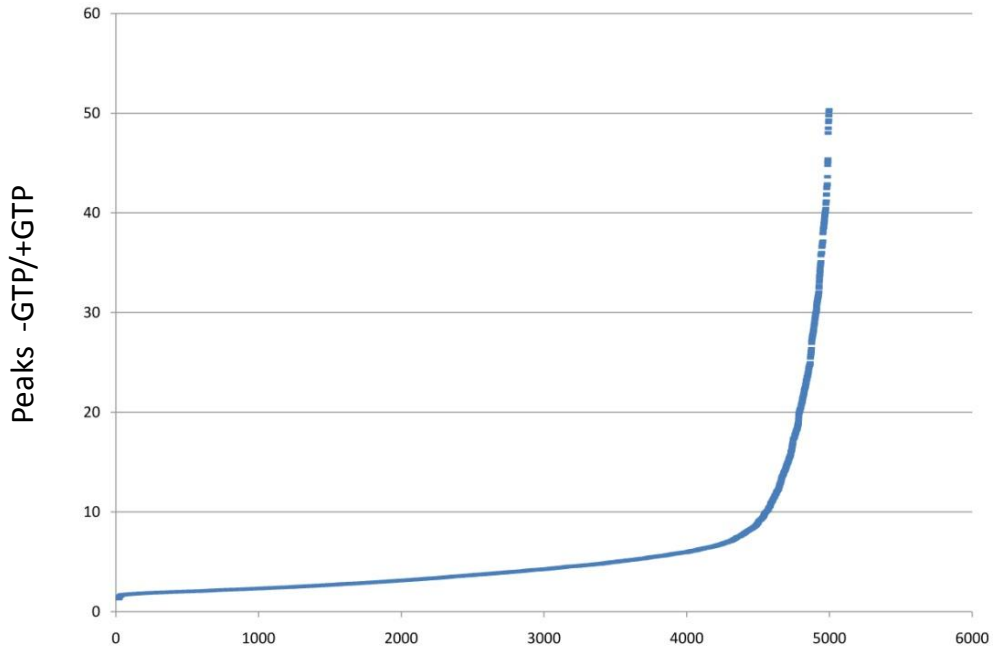
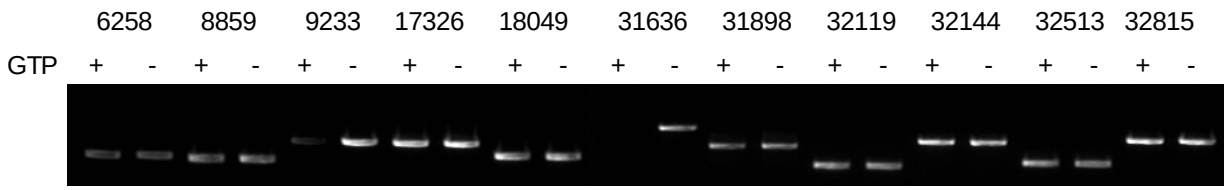


Figure 4.11: Picture of an agarose gel showing the results of PCR amplification using DNA incubated with McrBC in the presence or absence of GTP as template and primers specific of *P. tricornutum* genes identified as putatively methylated.



4.2.3 RNA silencing machinery in *P. tricornutum*

The RNA interference (RNAi) pathway in eukaryotes in involved in transcriptional and translational regulation as well as in the primary defense against viruses and transposable elements. The RNAi process is initiated by short (20-30 nucleotides) double-stranded RNA (dsRNA) molecules called small RNA (sRNA) generated by the endonuclease DICER (Bernstein et al., 2001). One of the two strands of sRNA fragment, known as the guide strand, is then loaded into the RNA-induced silencing complex (RISC) which includes the Argonaute (Ago) protein. Through this association, small RNAs take on unique properties that allow them to regulate diverse biological processes. For example, base-pairing with mRNA strands results in either degradation of the mRNA or suppression of translation. The RISC complex can also be re-located to the nucleus and trigger targeted chromatin condensation. There are three dominant classes of small silencing RNAs in plants and animals that reflect their distinct modes of production: the microRNAs (miRNAs), the small interfering RNA (siRNAs) and the PIWI- interacting RNAs (piRNAs). **ref**

miRNAs are non-protein-coding genes found in the genomes of plants and animals where they help to regulate gene expression, particularly during development (Zhang et al., 2007). miRNAs are transcribed as mRNA-like primary (pri-) miRNAs containing ~70 nucleotide long stem–loop structures. Once excised from the pri- miRNA by the nuclear enzyme Dicer, these stem-loops become precursor (pre-) miRNAs. Pre-miRNAs are exported to the cytoplasm where they are cleaved again by the Dicer endonuclease to yield ~22 nucleotide long mature miRNAs. miRNAs can direct an Argonaute protein to cleave an mRNA target, block mRNA translation or target mRNAs for destruction by standard mRNA turnover mechanisms (Zhao and Srivastava, 2007).

siRNAs were originally observed during transgene- and virus-induced silencing in plants (Mello and Conte, 2004), consistent with a natural role in genome defense. Later,

centromeres, transposons, and other repetitive sequences were uncovered as another source of siRNAs (Lippman and Martienssen, 2004). Functional studies in plants also led to the discovery of trans-acting siRNAs (ta-siRNAs) that are diced from specific genomic transcripts and regulate discrete sets of target genes (Vazquez et al., 2004). More recently, other sources of endogenous siRNAs have been identified including convergent mRNA transcripts and other natural sense-antisense pairs (Golden et al., 2008). In the canonical pathway, the siRNA guide strand directs RISC to complementary RNA targets, which are then degraded. RNA degradation is induced by the Ago protein. In plants, animals, and fungi, the primary dsRNA can induce synthesis of secondary siRNAs through the action of RNA-dependent RNA polymerase (RdRP) enzymes (Baulcombe, 2007). This secondary pool of siRNAs can greatly amplify and sustain the response, and in some organisms, such as plants and nematodes, they can lead to systemic silencing that spreads throughout the organism.

siRNAs are not restricted to posttranscriptional modes of repression. The Ago1-containing effector in fission yeast is referred to as the RNA-induced transcriptional silencing (RITS) complex and is guided to specific chromosomal loci such as centromeric repeats by its bound siRNA). RITS association promotes histone H3 methylation on lysine 9 (H3K9) by histone methyltransferases (HMTs), leading to chromatin compaction (Lippman and Martienssen, 2004). TGS has also been extensively investigated in plants where siRNA can guide the direct methylation of DNA by DNA methyltransferases (DMTs) in addition to histone methylation (Lippman and Martienssen, 2004).

In addition to miRNAs and siRNAs, a third small RNA silencing system has been uncovered that prevents the spreading of selfish genetic elements in animals. Production of the Piwi-associated RNAs (piRNAs) is initiated at a few master control regions within the genome but the nature of the primary piRNA-generating transcript is still unknown. Unlike miRNAs and siRNAs, the production of piRNAs does not require the action of the Dicer

protein. The core piRNA production machinery includes the Piwi protein, that shares similarity with the Ago protein (see later). The “ping pong” model proposes that, similar to the RdRP loop, sense transcripts from TEs and antisense transcripts from piRNA master loci fuel an amplification cycle involving a Piwi-containing RISC complex (Hartig et al., 2007).

A major limitation for the study of diatoms is the lack of tools to generate knockout or knockdown mutants through forward or reverse genetic approaches. The generation of loss-of-function mutants by insertional mutagenesis appears difficult in a diploid organism such as *P. tricornutum* for which a sexual cycle has never been reported. As there is no evidence for homologous recombination events in diatom genomes, it is also unlikely that targeted gene disruption using this process can be developed as a standard approach (Falciatore et al., 1999). The RNA interference pathway is often exploited in experimental biology to down-regulate target genes. Double-stranded RNA is synthesized with a sequence complementary to a gene of interest and introduced into a cell or organism, where it is recognized as exogenous genetic material. Activation of the RNAi pathway causes a drastic decrease in expression of the targeted gene (or gene family). Since RNAi may not totally abolish expression of the gene, this technique is sometimes referred as a "knockdown", to distinguish it from "knockout" procedures (Voorhoeve and Agami, 2003).

Therefore, the possibility of achieving gene silencing in *P. tricornutum* by introducing constructs that express anti-sense or inverted-repeats was investigated by the team of Dr Angela Falciatore in Naples, Italy. The team showed that a GUS reporter gene expressed in a transgenic line can be successfully silenced using anti-sense or inverted-repeat constructs, revealing for the first time the presence of a silencing machinery in diatoms. Additionally, they demonstrate that expression of two endogenous *P. tricornutum* genes, encoding phytochrome (Dph1) and cryptochrome/photolyase family 1 (CPF1), can be modulated using a similar approach. However, the identity of the proteins involved in this process remained

cryptic. Indeed, according to parsimony-based phylogenetic analysis, the most recent common ancestor of all eukaryotes most likely already possessed one Dicer-like protein, one Ago, one Piwi protein, and an RNA-dependent RNA polymerase that may have also played other cellular roles (Cerutti and Casas-Mollano, 2006). However, in their analysis, Cerutti and Casas-Mollano described the absence of genes encoding putative Dicer and RdRP proteins in the *T. pseudonana* genome. In the framework of my PhD I had the opportunity to collaborate with the Naples team to perform an in silico study aiming to identify putative components of the RNAi machinery mentioned above.

4.2.3.1 Identification of diatom genes putatively involved in RNAi-related processes

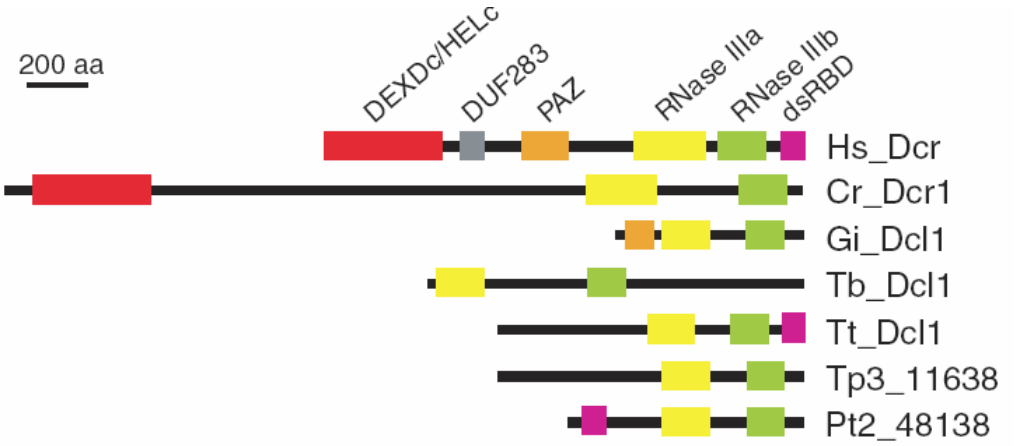
Dicer homologues typically contain an amino-terminal DEADc/HELICASEc domain, followed by a domain of unknown function (DUF283), a PAZ domain, two neighbouring RNase III domains (RNase IIIa and RNase IIIb), and a double-stranded RNA-binding domain (dsRBD) (Figure 4.12). Although previous analysis in the *T. pseudonana* genome indicated the absence of a canonical Dicer, this overall organization does not apply to all organisms. For instance, the Dicer-like protein 1 (DCL1) from the ciliate protozoan *Tetrahymena thermophila* (a chromalveolate) only contains the two RNase III domains and a dsRBD domain (Mochizuki and Gorovsky, 2005). The *Giardia intestinalis* (an excavate) DCL1 (MacRae et al., 2006) consists of a PAZ domain, a divergent DUF283 motif and two adjacent RNase III domains. Remarkably, the highly divergent *Trypanosoma brucei* (another excavate) Dicer-like 1, which contains only two RNase III domains, drives the RNA interference pathway in this organism (Shi et al., 2006). Therefore, to date, the only universal feature of the Dicer family is the presence of two RNase III domains.

I thus searched both diatom genomes for candidate gene models matching these criteria. In the *T. pseudonana* genome I found a gene model encoding a protein with two

RNase III domains (Tp3_11638) and in *P. tricornutum* another (Pt2_48138) containing a dsRBD followed by two RNase III domains (Figure 4.12 and Table 4.2). Therefore, putative diatom Dicer-like proteins are only distantly related to the multi-domain proteins identified in plants and animals, although the conservation in terms of domain structure with other eukaryotes and key amino acid residues (Figure 4.12) may support their role as bona fide Dicers. Interestingly, we identified an additional gene model (Tp3_20605) in the *T. pseudonana* genome that may play a role in RNA silencing (Table 4.2). This protein displays a domain association that we could not find in any other organism, consisting of a DEADc/HELICASEc followed by a C-terminal dsRBD. The DEAD/HELICASEc domain from this protein has homology with the DEAD/HELICASEc domains in a range of Dicer-like proteins.

Figure 4.12 (next page): Diversity of Dicer proteins among eukaryotes. (A) Schematic representation of domains found in various Dicer (Dcr) and Dicer-like (Dcl) proteins. Domain abbreviations: DEXDc/HELIc, DEAD-like helicase domain/helicase C-terminal domain; DUF283, DUF283 domain; PAZ, PAZ domain; RNAseIII (a-b), ribonuclease III domains a and b and dsRBD, double-stranded RNA-binding domain. (B) CLUSTALW multiple sequence alignment of RNase IIIa and RNase IIIb domains from various Dicer-like proteins. Asterisks indicate key residues in the catalytic domain. Species abbreviations: Cr, *Chlamydomonas reinhardtii*; Gi, *Giardia intestinalis*; Hs, *Homo sapiens*; Pt, *Phaeodactylum tricornutum*; Tb, *Trypanosoma brucei*; Tp, *Thalassiosira pseudonana*; Tt, *Tetrahymena thermophila*.

A


200 aa
DEXDc/HELc
DUF283
PAZ
RNase IIIa
RNase IIIb
dsRBD
Hs_Dcr
Cr_Dcr1
Gi_Dcl1
Tb_Dcl1
Tt_Dcl1
Tp3_11638
Pt2_48138

B

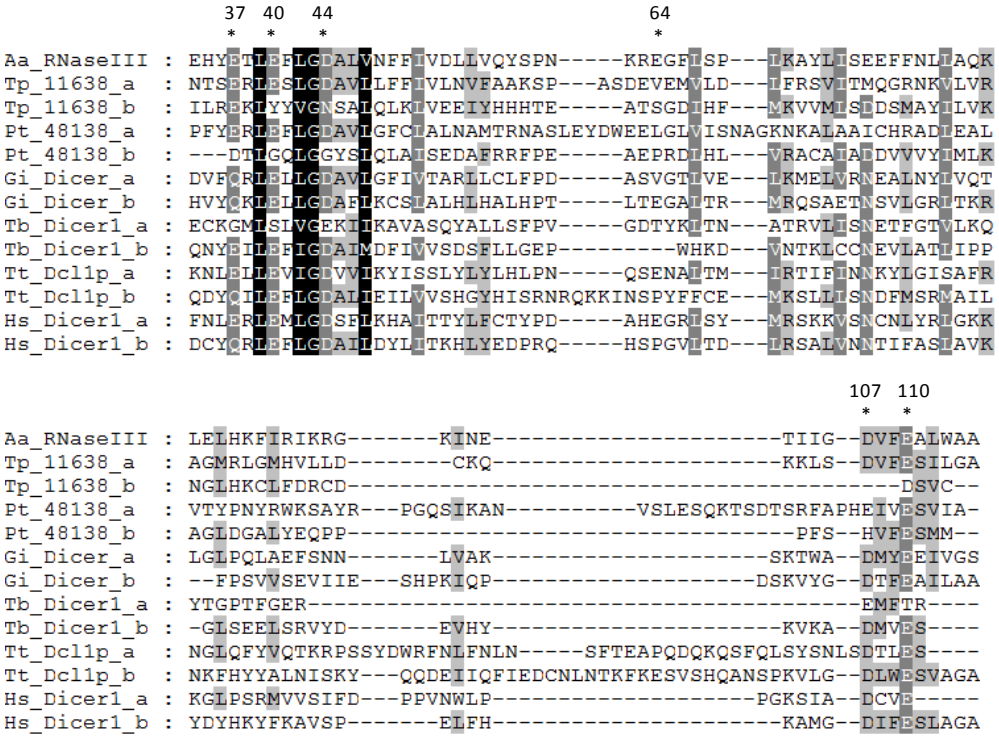

37 40 44 64
Aa_RNaseIII : EHYETLEFLGDALVNFFIVDLLVQYSPN-----KREGFLSP---LKAYLISEEFFNLIAQK
Tp_11638_a : NTSERLESLGDAVLLFFIVLNVFAAKSP---ASDEVEMVLD---LFRSVITMQGRNKVLVR
Tp_11638_b : ILREKLYYVGNSALQLKIVEEIYHHHTE-----ATSGDIHF---MKVVMLSIDSMAYILVK
Pt_48138_a : PFYERLEFLGDAVLGFCLALNAMTRNASLEYDWEELGLVISNAGKNKALAAAICHRADIEAL
Pt_48138_b : ---DTLGQLGGYSLQLALSEDAFRRFPE-----AEPRDIHL---VRACAIALDVVVYIMLK
Gi_Dicer_a : DVFQRLELLGDAVLGFIVTARLLCLFPD-----ASVGTIVE---LKMELVRNEALNYIVQT
Gi_Dicer_b : HVYQKLELLGDAFLKCSLALHLHALHPT-----LTEGAITR---MRQSAETNSVLGRITKR
Tb_Dicer1_a : ECKGMISLVGEKITKAVASQYALLSFPV-----GDTYKITN---ATRVLISNETFGTVLKQ
Tb_Dicer1_b : QNYEILEFIGDAIMDFIVVSDSFLLGEP---------WHKD---VNTKLCCNEVLATLIPP
Tt_Dcl1p_a : KNLELLEVIGDVVIKYISSLYLYLHLPN-----QSENATTM---IRTIFINNKYLGISAFR
Tt_Dcl1p_b : QDYQILEFLGDALIEILVVSHGYHISRNRQKKINSPYFFCE---MKSLLLSNDFMSRMAIL
Hs_Dicer1_a : FNLERLEMLGDSFLKHAITTYLFCTYPD-----AHEGRISY---MRSKKVSNCNLYRLGKK
Hs_Dicer1_b : DCYQRLEFLGDAILDYLITKHLYEDPRQ-----HSPGVLTD---LRSALVNNTIFASLAVK
107 110
Aa_RNaseIII : LELHKFIRIKRG-------KINE----------TIIG--DVFEALWAA
Tp_11638_a : AGMRLGMHVLLD--------CKQ----------KKLS--DVFESILGA
Tp_11638_b : NGLHKCLFDRCD----------DSVC--
Pt_48138_a : VTYPNYRWKSAYR---PGQSIKAN----------VSLESQKTSDTSRFAPHEIVESVIA-
Pt_48138_b : AGLDGALYEQPP----------PFS--HVFESMM--
Gi_Dicer_a : LGLPQLAEFSNN-------LVAK----------SKTWA--DMYEEIVGS
Gi_Dicer_b : --FPSVVSEVIIE---SHPKIQP----------DSKVYG--DTFEAILAA
Tb_Dicer1_a : YTGPTFGER----------EMFTR----
Tb_Dicer1_b : -GLSEELSRVYD-------EVHY----------KVKA--DMVES----
Tt_Dcl1p_a : NGLQFYVQTKRPSSYDWRFNLFNLN-----SFTEAPQDQKQSFQLSYSNLSDTLES----
Tt_Dcl1p_b : NKFHYYALNISKY---QQDEIIQFIEDCNLNTKFKESVSHQANSPKVLG--DLWESVAGA
Hs_Dicer1_a : KGLPSRMVVSIFD---PPVNWLP----------PGKSIA--DCVE-----
Hs_Dicer1_b : YDYHKYFKAVSP-------ELFH----------KAMG--DIFESLAGA

Table 4.2: Protein ID in the diatom genomes, accession number in GenBank, predicted protein domains with Pfam database, and available ESTs information are indicated. NA: Not Applicable.

		P. tricornutum	*T. pseudonana*
Dicer-like	Protein id	Pt2_48138	Tp3_11638
	Accession number	FJ750271	FJ750272
	Pfam (E value)	dsRBD (0.44); RNaseIII (0.0025); RNaseIII (0.33)	RNaseIII (0.0013); RNaseIII (0.0016)
	Best hit in GenBank (E-value)	Dicer-like *Coprinopsis cinerea okayama* (0.012)	RNaseIII *Sulfurovum sp* (0.014)
	EST information	YES	NA
Dicer Related	Protein id	no homolog found	Tp3_20605
	Pfam (E value)		DEAD (4.8e-11); HELICc (2.5e-05); dsRBD (0.0039)
	Best hit in GenBank (E-value)		Dicer Related Helicase *C. elegans (1e-19)*
	EST information		YES
Ago-like	Protein id	Pt2_47611	Tp3_1029
	Accession number	FJ750269	FJ750270
	Pfam (E value)	PAZ (0.13); PIWI (2.4e-12)	PAZ (2.8e-05); PIWI (3.9e-15)
	Best hit in GenBank (E-value)	Ago-like *Chlamydomonas reinhardtii* (5e-21)	Ago-like *Coprinopsis cinerea okayama* (9e-18)
	EST information	YES	YES
RdRP-like	Protein id	Pt2_45417	Tp3_5028
	Accession number	FJ750273	FJ750274
	Pfam (E value)	RdRP (6e-08)	RdRP (0.00015)
	Best hit in GenBank (E-value)	RdRP *Paramecium tetraurelia* (2e-09)	RdRP *Paramecium tetraurelia* (3e-04)
	EST information	YES	YES
	Protein id		Tp3_8685
	Accession number		FJ750275
	Pfam (E value)		RdRP (0.00026)
	Best hit in GenBank (E-value)		RdRP *Zea mais* (5e-06)
	EST information		YES

The Ago proteins, core components of the effector RNA-induced silencing complexes (RISC), are classified into three paralogous groups: the Argonaute-like proteins, similar to *A. thaliana* AGO1, the Piwi-like proteins, related to the *D. melanogaster* PIWI and the recently identified *C. elegans*-specific group 3 Argonautes. Animals have retained members of both the AGO-like and Piwi subfamilies, whereas the different Argonautes identified in green algae and plants all belong to the AGO-like subfamily. The Argonaute proteins involved in RNAi pathways always contain distinct functional domains: a variable N-terminal domain and conserved C-terminal PAZ, MID and PIWI domains (Hutvagner and Simard, 2008; Vaucheret, 2008). The PIWI domain often has endonuclease activity, although this activity is not essential for RNAi regulatory functions. Putative proteins containing these domains were identified both in *T. pseudonana* (Tp3_1029) and *P. tricornutum* (Pt2_47611; Table 4.2). Multiple alignment of the Ago_Piwi domain from diatoms with other Ago-Piwi proteins revealed significant conservation, although the catalytic site for endonuclease activity was only poorly conserved (Figure 4.13). Phylogenetic analysis of the diatom Argonautes with proteins identified in other eukaryotes revealed that they cluster together in a clade only distantly related to the known paralogous groups mentioned above (Figure 4.14), suggesting a possible functional specialization.

We also found genes encoding putative RdRP in *P. tricornutum* (Pt2_45417) and two in *T. pseudonana* (Tp3_ 8685 and Tp3_5028; Table 4.2). Thus, in silico analysis of components of the RNAi pathway indicates that molecular players involved in RNA silencing in other eukaryotes are only poorly conserved in diatoms, and that distantly related proteins may fulfil their function in these organisms. Considering the mixed evolutionary origins of diatom genomes, it is also possible that as yet not characterized proteins, for example of bacterial origin, play a role in RNAi processes in diatoms. Therefore, deciphering the diatom RNAi machinery will now be necessary to characterize the RNAi components proposed here.

Figure 4.13: CLUSTALW multiple sequence alignment of AGO-PIWI domain from various Ago-Piwi-like proteins. Asterisks indicate key residues in the catalytic domain. Species abbreviations: Aa (*Aquifex aeolicus*), At (*Arabidopsis thali*ana), Ce (*Caeonorhabditis elegans*), Dm (*Drosophila melanogaster*), Gi (*Giardia intestinalis*), Hs (*Homo sapiens*), Pt (*Phaeodactylum tricornutum*), Tp (*Thalassiosira pseudonana*)

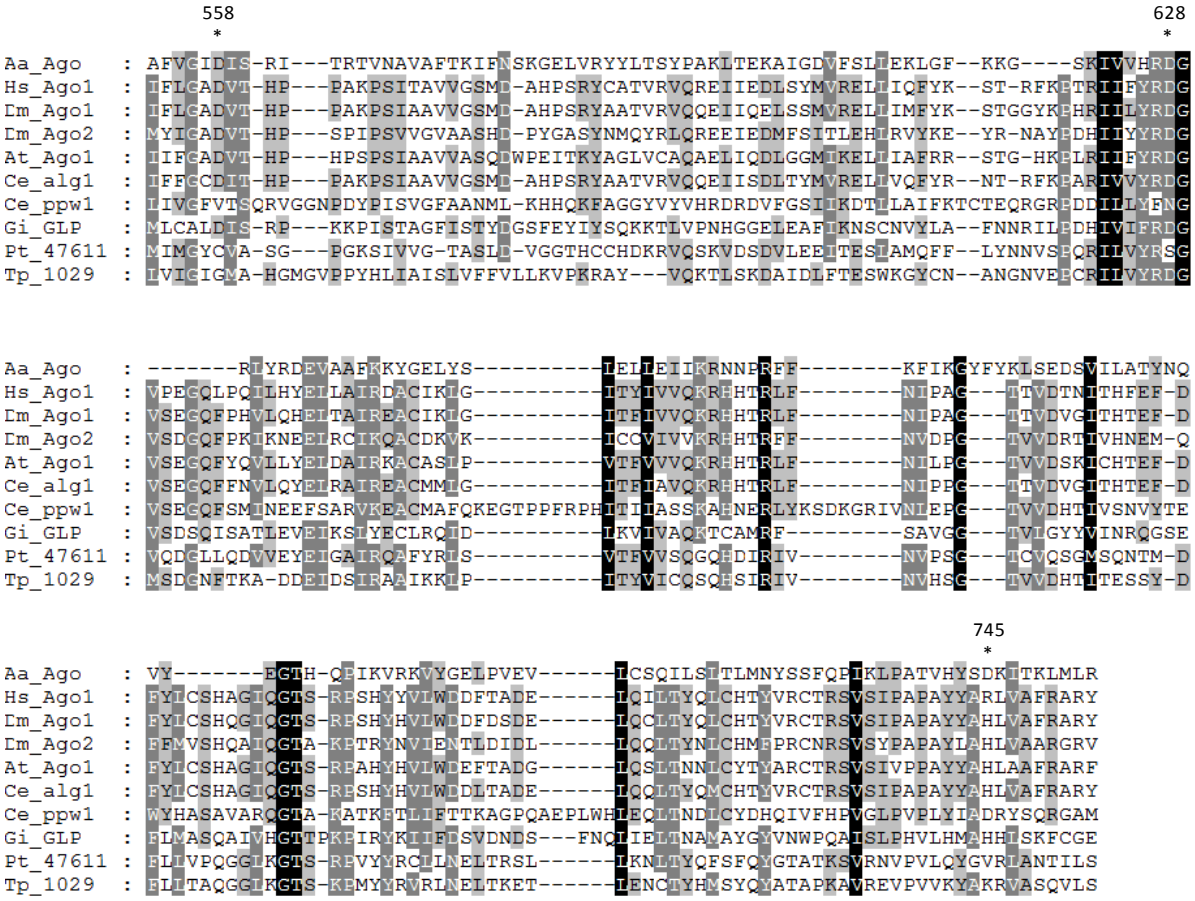
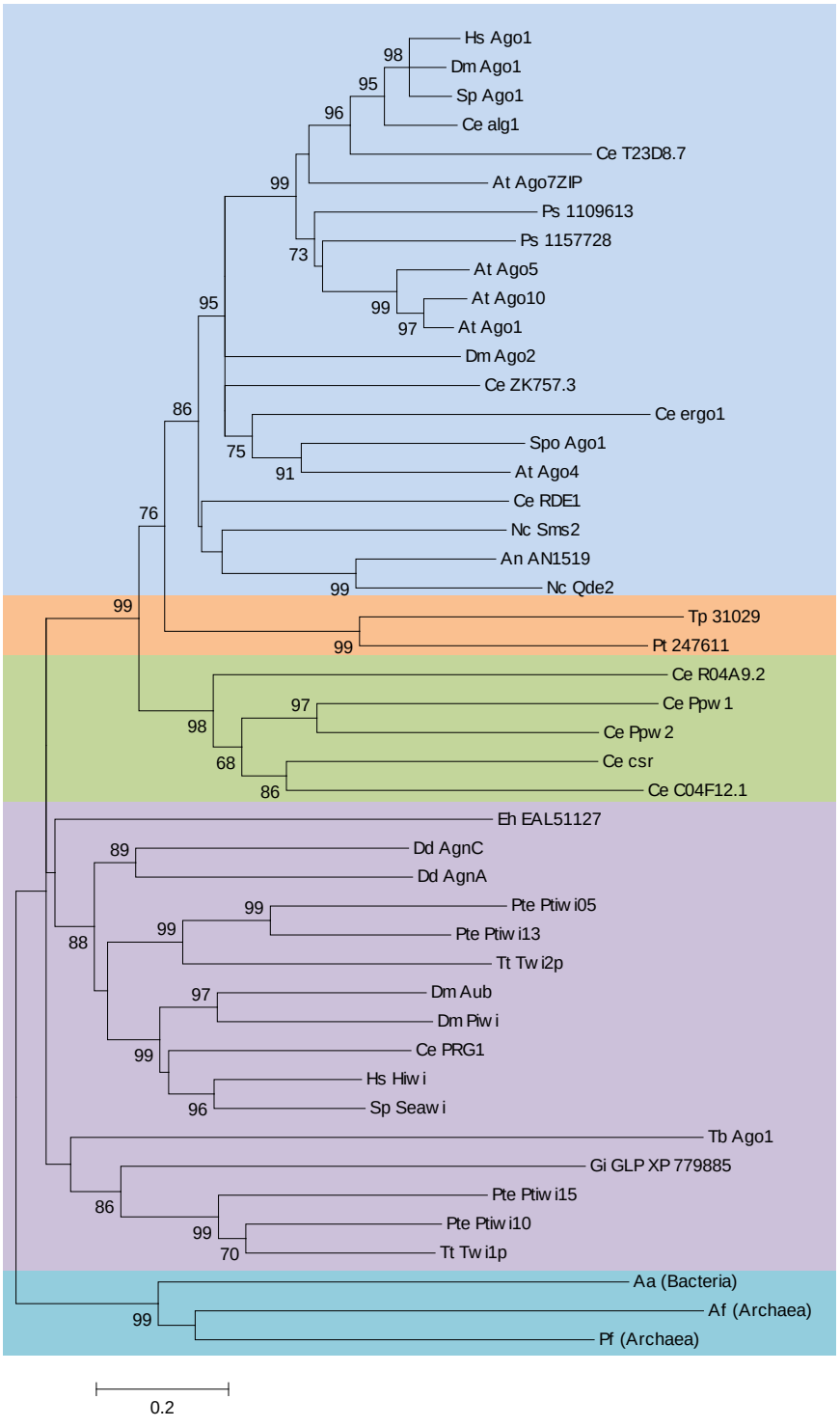


Figure 4.14 (next page): Phylogenetic analysis of the diatom Argonaute-like proteins. Phylogenetic tree constructed using the MEGA 4.0 platform with the Neighbor-Joining method from a MUSCLE alignment of the PAZ and AGO domains from Ago-like, Piwi-like, *C. elegans*-specific group 3 Argonautes, prokaryotic Argonautes, and the putative diatom Ago-Piwi-like proteins identified in this work. The alignment was manually refined to remove gaps and erroneous positions. Numbers indicate interior branch bootstrap values as percentage, based on 1000 pseudoreplicates (only values >60% are shown). Species abbreviations and accession numbers of proteins used to draw the tree: Aa, *A. aeolicus* (AA Ago gi126031218); Af, *Archaeoglobus fulgidus* (Af Ago gi60593831); An, *Aspergillus nidulans* (AN1519, EAA63775); At, *A. thaliana* (At Ago1 AAC18440, At Ago4 NP565633, At Ago5 NP850110, At Ago7 ZIP NP177103, At Ago10 CAA11429); Ce, *C. elegans* (CeC04F12.1 gi17505468, Ce ZK757.3 gi17557077, Ce R04A9.2 gi17569229, Ce PRG1 gi3875393, Ce T23D8.7 gi3880077, Ce RDE1 gi6272678, Ce Alg1 gi25148113, Ce csr1 gi115532838, Ce ergo1 gi25148583, Ce ppw1 NP740835, Ce ppw2 AAF60414); Dd, *Dictyostelium discoideum* (Dd AgnA EAL69296, Dd AgnC EAL71514); Dm, *Drosophila melanogaster* (Dm Ago1 BAA88078, Dm Ago2 Q9VUQ5, Dm Aub CAA64320, Dm Piwi Q9VKM1); Eh, *Entamoeba histolytica* (Eh EAL51127); Gi, *Giardia intestinalis* (Gi GLP XP779885); Hs, *H. sapiens* (Hs Ago1 AAH63275, Hs Hiwi AAC97371); Nc, *Neurospora crassa* (Nc Qde2 AAF43641, Nc Sms2 AAN32951); Pf, *Pyrococcus furiosus* (Pf Ago gi62738878); Pte, *Paramecium tetraurelia* (Pte Ptiwi05 CAI44468, Pte Ptiwi10 CAI39070, Pte Ptiwi13 CAI39067, Pte Ptiwi15 CAI39065); Sp, *Strongylocentrotus purpuratus* (Sp Ago1 XP782278, Sp Seawi AAG42533); Spo, *Saccharomyces pombe* (Ago1 O74957); Tb, *T. brucei* (Ago1 O74957); Tt, *T. thermophila* (Tt Twi1p AAM77972, Tt Twi2p AAQ74967); Ps, *Phytophtora sojae* (1109613, 1157728); Pt, *P. tricornutum* (FJ750269); Tp, *T. pseudonana* (FJ750270).

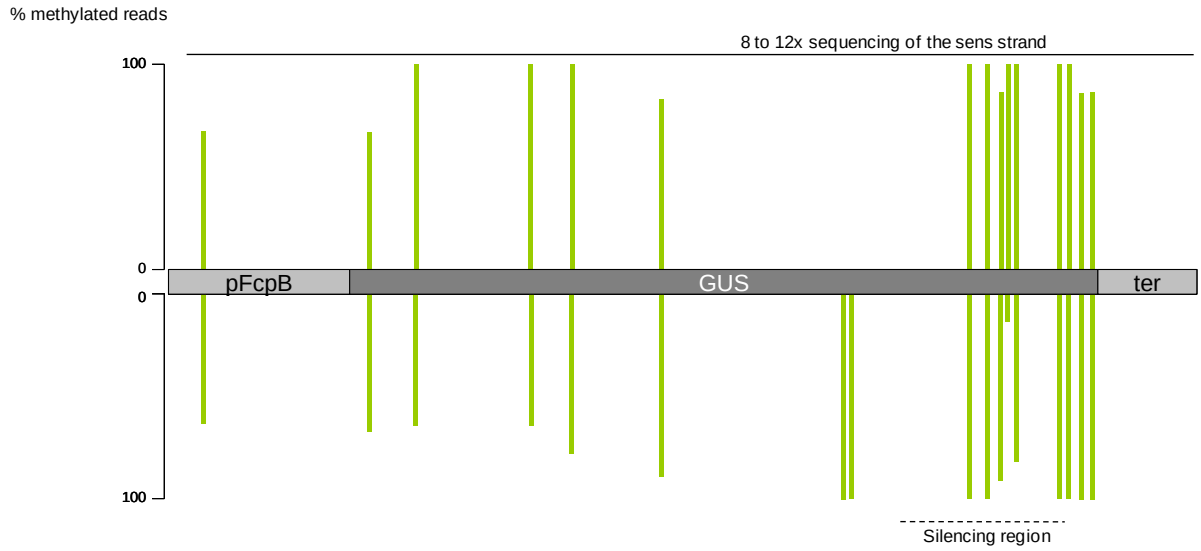

Hs Ago1
Dm Ago1
Sp Ago1
Ce alg1
Ce T23D8.7
At Ago7ZIP
Ps 1109613
Ps 1157728
At Ago5
At Ago10
At Ago1
Dm Ago2
Ce ZK757.3
Ce ergo1
Spo Ago1
At Ago4
Ce RDE1
Nc Sms2
An AN1519
Nc Qde2
Tp 31029
Pt 247611
Ce R04A9.2
Ce Ppw 1
Ce Ppw 2
Ce csr
Ce C04F12.1
Eh EAL51127
Dd AgnC
Dd AgnA
Pte Ptiw i05
Pte Ptiw i13
Tt Tw i2p
Dm Aub
Dm Piw i
Ce PRG1
Hs Hiw i
Sp Seaw i
Tb Ago1
Gi GLP XP 779885
Pte Ptiw i15
Pte Ptiw i10
Tt Tw i1p
Aa (Bacteria)
Af (Archaea)
Pf (Archaea)
0.2
Argonaute
Diatom Argonaute
Group 3 Argonaute
PIWI
Bacteria/Archaea

4.2.3.2 RNA-directed DNA methylation (RdDM) in *P. tricornutum*

As mentioned above, small RNAs can also trigger transcriptional gene silencing associated with DNA methylation by RNA-directed DNA methylation (RdDM) (Matzke et al., 2007). We therefore investigated the appearance of DNA methylation within the GUS gene of silenced clones using bisulfite sequencing to assess and quantify cytosine methylation. The analysis showed DNA methylation of the GUS locus in the silenced line, revealing a total of 17 mC, all in a CG context, and 15 of them being found symmetrically on both sense and anti-sense strands, while two were found only on the anti-sense strand (Figure 4.15). Six of these sites were located in the region base-pairing with the silencing construct. Cytosine methylation was also found outside of this region, with three mCs located immediately downstream and the remaining sites spreading further upstream to the promoter. Introduction of the silencing vector in the GUS expressing line therefore caused de novo DNA methylation within the targeted region as well as methylation spreading along the transcribed region, as also observed in plants (Matzke et al., 2009).

It thus appears that production of dsRNA in *P. tricornutum* feeds backward to alter chromatin structure at the level of DNA regions with which they share sequence identity by the RdDM process. No cytosine methylation at CHG or CHH sites was found by bisulfate sequencing, suggesting that RNAi-induced de novo methylation in *P. tricornutum* may occur only in the CG context. All together, these results indicate that in silenced clones, DNA methylation correlates with transcriptional (or post-transcriptional) silencing of the GUS transgene. More analysis will be required to characterize the small RNAs and effector proteins putatively associated with this phenomenon, for example to understand which DNMTase of those presented above is responsible for the observed sRNA-induced de novo DNA methylation.

Figure 4.15: De novo cytosine methylation in silenced clones. Schematic representation of the methylation profile obtained by bisulfite sequencing of fir-1 clone. Vertical bars show the distribution of mC in the sense and antisense strands. The silencing region is indicated.



4.3 Material and Methods

Sequence identification and analysis

Diatom genomes were screened for the presence of gene models or loci encoding proteins with similarity to proteins and domains reported to be involved in chromatin-based transcriptional regulation. This computational analysis was performed using BLAST and reverse BLAST comparisons as well as search for specific InterPro domains. Analysis of domain composition of the candidate proteins was done by submitting their amino acid sequences to InterProScan (http://www.ebi.ac.uk/InterProScan/index.html). Multiple protein alignments were done using CLUSTALW and the phylogenetic trees were constructed with the Neighbor Joining method using the Mega 4 (Tamura et al., 2007) and the SplitsTree4 platforms (Huson and Bryant, 2006).

Protein Analysis

Histone modifications were analyzed by sodium dodecyl sulfate-polyacrylamide gel electrophoresis and revealed by western blotting using Amersham enhanced chemiluminescence kit (ECL kit, Amersham biosciences, Piscataway, NJ, USA). About 1x107 *P. tricornutum* cells were collected and used for protein extraction. We also extracted proteins from cauliflower to use as positive control. Protein samples were incubated in Laëmmli buffer (1x) for 5 minutes at 95°C, resolved on a 15% SDS-PAGE gels, and transferred to nitrocellulose membranes (Protran BA85, Whatman, GE Healthcare group, Maidstone, UK). The membranes were then incubated at 4°C over night with dilutions (1:5000 or 1:10000) of the following primary antibodies raised in rabbit: anti-monoacetylated H4 (K5), anti-monoacetylated H4 (K8), anti-monoacetylated H4 (K12), and anti-monoacetylated H4 (K16) with respective references AB3064, AB3060, AB3058, and AB3056, Chemicon, Millipore group, Billerica, MA, US, and anti-polyacetylated H3 (K),

anti-dimethylated H3 (K4), anti-dimethylated H3 (K36), anti-dimethylated H3 (K9), and anti-trimethylated H3 (K27) with respective references 06-599, 07-030, 07-369, 07-212, and 07-449, Upstate, Millipore group, Billerica, MA, US, followed by an incubation with horseradish peroxydase (HRP)-conjugated anti-rabbit IgG secondary antibody for 1 hour at room temperature, and visualized with ECL kit.

Chromatin extraction and immunoprecipitation

A chromatin extraction protocol was adapted to *P. tricornutum* from a protocol used in yeast for this purpose (ultimate ChIP yeast assay, Tansley lab protocols, 2002) as follows:

1. Grow a *P. tricornutum* culture to a cell density of about 2.10e6 cells/ml
2. Transfer 10 ml of culture to a 15ml falcon
3. Under hood, add 405µl of 37% formaldehyde dropwise (toxic++)
4. Gently rock or rotate (horizontally at 120 rpm) for 10 minutes at room temperature (RT)
5. Under hood, add 650µl 2M glycine
6. Gently rock or rotate (horizontally at 120 rpm) for 10 minutes at RT
7. Centrifuge 5 minutes at 4,500 rpm, 4°C
8. Under hood, discard supernatant, resuspend cell pellet in 10ml cold PBS
9. Centrifuge 5 minutes at 4,500 rpm, 4°C
10. Discard supernatant, resuspend cell pellet in 10ml cold PBS
11. Centrifuge 5 minutes at 4,500 rpm, 4°C, discard supernatant
12. Resuspend pellet in 1ml lysis buffer + inhibitors
13. Keep tubes on ice for 10 minutes
14. Gently shake for 10 minutes at RT
15. Centrifuge 10 minutes at 10,000 rpm, 4°C

16. Discard supernatant, resuspend in 300µl lysis buffer + inhibitors

17. Store at -20°C or proceed to sonication

For sonication, we use the Bioruptor sonicator (UCD-200, Diagenode, Liège, Belgium. Chromatin is sonicated at full power with 30s/30s (sonic/chill) cycles, 12 times and then centrifuged 2 minutes at 4,000 rpm, 4°C. 10µl of supernatant is then loaded on agarose gel to verify the obtention of DNA with an average size around 700 bp. For Chromatin immunoprecipitation, we used the OneDay ChIP Kit (OD-TM, Diagenode, Liège, Belgium) following manufacturer's recommendations and using 5µl of the antibodies anti-RNA polymerase II and anti-dimethylated H3 (K9) with respective references 05-623 and 07-212, Upstate, Millipore group, Billerica, MA, US). Quantitative PCR was then performed with equal amounts (2 ng) of immunoprecipitated and control (antibody replaced by ChIP buffer) DNA samples using an ABI 7900 Real-Time PCR system and Eurogentec SYBR green I MasterMix Plus with the primers presented in Table 4.3. Relative occupancy for each locus was then calculated as = 2 ^ (Ct_Control – Ct_Antibody), where Ct is the mean cycle threshold value from two technical replicates.

Table 4.3

	Primer name	Sequence 5' 3'
Transposons	PtC13_Fw997	TTGCAAATTTTCAGCAGCAC
	PtC13_Rev1169	AGAAGGCTGGGACACAGAGA
	PtPiggy_Fw547	CAAGCGGAACACTCTGAACA
	PtPiggy_Rev696	AAGACCGCTTCTCCTTCCTC
	BKB_Fw4022	CGAAGCTACTATGCCGGAAG
	BKB_Rev4208	AAGGACACGAGAGTCGAGGA
	BKB_Fw5448	GTCGAACGAGACTTGCATGA
	BKB_Rev5666	CATCGCAAGAAACGAAGACA
	SCF_Fw2508	CAGCCTGAGGCGAAAGATAC
	SCF_Rev2678	TAGTTCTGACATGCGCCAAG
	PtMuDR2_Fw2020	GCATGCAGGAGAGGTGTTTT
	PtMuDR2_Rev2208	ACAATGCTGCTGAGCAAATG
Genes	H4_Fw	AGGTCCTTCGCGACAATATC
	H4_Rev	ACGGAATCACGAATGACGTT
	Fucoxanthin B_Fw	GCCTTCATTGGTTTCTTGGA
	Fucoxanthin B_Rev	TCCCTGGTTAAGCTCAATGG
	ITS2_Fw	GTCTTGAGTCCAACCCGGTA
	ITS2_Rev	GAGGCGCAGATACCCAAATA
	Cryptochrome 1_Fw	CCGTCGACGTATGGAAGTTT
	Cryptochrome 1_Rev	AGCAGCTGCGTATCCAAGAT
	Phytochrome 1_Fw	CGTGCACGAGATTAAAGCAA
	Phytochrome 1_Rev	GCATGCGAATTTGACTGAGA
	Cryptochrome DASH_Fw	CTCCGAACGGCAGTCTACTC
	Cryptochrome DASH_Fw	CGTCCAGTTCGTACATGGTG

Bisulfite sequencing

Genomic DNA from the fir-1 RNAi clone was treated with bisulfite using the MethylCode Bisulfite Conversion Kit, Invitrogen. The converted DNA was subsequently amplified by PCR with AccuPrime™ Taq DNA Polymerase, Invitrogen, using a combination of primers for sense and antisense amplification of the FcpBp-GUS-FcpA3' region (Supplementary Data). PCR products were cloned in pCR2.1 vectors using TA cloning kit (Invitrogen) and transformed into TOP10 E. coli. Positive colonies were identified by PCR, which products

were subsequently sequenced. 8 to 12 reads from each pair of primers were then aligned with the fir-1 construct using CLUSTALW to identify methylated (unconverted) cytosine residues.

Window McrBC Restriction (WMR)

We adapted the McrBC-based protocol to exclude the methylated DNA fraction from Lippman et al. (2005) with the main modification that sonicated DNA if first delimited to a specific size-range (see Figure 4.7).

1. Prepare a tube with 100 ng/µl of genomic DNA and sonicate at an average size of 500-700 nt
2. Load 2µl of sonicated DNA on a 1.5% agarose-TBE gel to verify size
3. Load 50µl of sonicated DNA in a single well of a 1.5% agarose-TBE gel. Run at 50V maximum
4. Let the sample migrate enough to have good resolution, but not too much to avoid cutting a too large agarose slice.
5. Cut sharply the gel slice corresponding to the 500-700 nt DNA fragments. Agarose slice should weight maximum 800 mg. At this point, cut the gel slice in two halves.
6. Purify the DNA using QIAquick Gel Extraction kit according to manufacturer's protocol but melt gel slices in two separate 1.5 ml tubes and bind DNA from both tubes on a single column. Elute DNA with 50µl water.
7. Load 2µl 1.5% agarose-TBE gel to check the size of the cleaned DNA
8. Prepare two tubes with 200 ng of size-filtered sonicated DNA. Label the first tube "McrBC" and add 1µl of McrBC enzyme, 3µl of 10x McrBC buffer, 1µl GTP (100 mM), 1µl BSA (10 mg/ml) and water to 30µl. Label the second tube "control" and add 1µl of McrBC enzyme, 3µl of 10x McrBC buffer, 1µl BSA (10 mg/ml) and water to 30µl.
9. Incubate at 37°C over night and inactivate the enzyme by heating 20 minutes at 65°C.

10. Load McrBC-treated and untreated samples on a 1.5% agarose-TBE gel. Let the sample migrate enough so that DNA fragments cut by McrBC can exit the DNA window. Depending on the methylation level of your DNA, you might see a smear coming down the window of your McrBC-treated sample but not from the control sample.

11. Purify the DNA windows from both samples using QIAquick Gel Extraction kit according to manufacturer's protocol and elute with 50µl water. For PCR or real time PCR, add 200µl of water to each sample and use 1µl per reaction.

4.4 References:

Baulcombe DC: Molecular biology. Amplified silencing. *Science* 2007, 315:199-200.

Bestor T, Laudano A, Mattaliano R, Ingram V: Cloning and sequencing of a cDNA encoding DNA methyltransferase of mouse cells. The carboxyl-terminal domain of the mammalian enzymes is related to bacterial restriction methyltransferases. *J Mol Biol* 1988, 203:971-983.

Cerutti H, Casas-Mollano JA: On the origin and functions of RNA-mediated silencing: from protists to man. *Curr Genet* 2006, 50:81-99.

Chen T, Li E: Structure and function of eukaryotic DNA methyltransferases. *Curr Top Dev Biol* 2004, 60:55-89.

Cheng X, Roberts RJ: AdoMet-dependent methylation, DNA methyltransferases and base flipping. *Nucleic Acids Res* 2001, 29:3784-3795.

Cokus SJ, Feng S, Zhang X, Chen Z, Merriman B, Haudenschild CD, Pradhan S, Nelson SF, Pellegrini M, Jacobsen SE: Shotgun bisulphite sequencing of the Arabidopsis genome reveals DNA methylation patterning. *Nature* 2008, 452:215-219.

Falciatore A, Casotti R, Leblanc C, Abrescia C, Bowler C: Transformation of Nonselectable Reporter Genes in Marine Diatoms. *Mar Biotechnol* (NY) 1999, 1:239-251.

Ferguson-Smith AC, Surani MA: Imprinting and the epigenetic asymmetry between parental genomes. *Science* 2001, 293:1086-1089.

Flanagan JF, Mi LZ, Chruszcz M, Cymborowski M, Clines KL, Kim Y, Minor W, Rastinejad F, Khorasanizadeh S: Double chromodomains cooperate to recognize the methylated histone H3 tail. *Nature* 2005, 438:1181-1185.

Gibbons RJ, McDowell TL, Raman S, O'Rourke DM, Garrick D, Ayyub H, Higgs DR: Mutations in ATRX, encoding a SWI/SNF-like protein, cause diverse changes in the pattern of DNA methylation. *Nat Genet* 2000, 24:368-371.

Golden DE, Gerbasi VR, Sontheimer EJ: An inside job for siRNAs. *Mol Cell* 2008, 31:309-312.

Goll MG, Bestor TH: Eukaryotic cytosine methyltransferases. *Annu Rev Biochem* 2005, 74:481-514.

Goll MG, Kirpekar F, Maggert KA, Yoder JA, Hsieh CL, Zhang X, Golic KG, Jacobsen SE, Bestor TH: Methylation of tRNAAsp by the DNA methyltransferase homolog Dnmt2. *Science* 2006, 311:395-398.

Hartig JV, Tomari Y, Forstemann K: piRNAs--the ancient hunters of genome invaders. *Genes Dev* 2007, 21:1707-1713.

Holliday R, Pugh JE: DNA modification mechanisms and gene activity during development. *Science* 1975, 187:226-232.

Huang Y, Fang J, Bedford MT, Zhang Y, Xu RM: Recognition of histone H3 lysine-4 methylation by the double tudor domain of JMJD2A. *Science* 2006, 312:748-751.

Huson DH, Bryant D: Application of phylogenetic networks in evolutionary studies. *Mol Biol Evol* 2006, 23:254-267.

Hutvagner G, Simard MJ: Argonaute proteins: key players in RNA silencing. *Nat Rev Mol Cell Biol* 2008, 9:22-32.

Jaenisch R, Bird A: Epigenetic regulation of gene expression: how the genome integrates intrinsic and environmental signals. *Nat Genet* 2003, 33 Suppl:245-254.

Jarvis EE, Dunahay TG, Brown LM: DNA nucleoside composition and methylation in several species of microalgae. *Journal of Phycology* 1992, 28:356-362.

Jeddeloh JA, Stokes TL, Richards EJ: Maintenance of genomic methylation requires a SWI2/SNF2-like protein. *Nat Genet* 1999, 22:94-97.

Jones PA, Laird PW: Cancer epigenetics comes of age. *Nat Genet* 1999, 21:163-167.

Jones PA, Takai D: The role of DNA methylation in mammalian epigenetics. *Science* 2001, 293:1068-1070.

Jung J, Mysliwiec MR, Lee Y: Roles of JUMONJI in mouse embryonic development. *Dev Dyn* 2005, 232:21-32.

Kim J, Daniel J, Espejo A, Lake A, Krishna M, Xia L, Zhang Y, Bedford MT: Tudor, MBT and chromo domains gauge the degree of lysine methylation. *EMBO Rep* 2006, 7:397-403.

Kouzarides T: Chromatin modifications and their function. *Cell* 2007, 128:693-705.

Kumar S, Cheng X, Klimasauskas S, Mi S, Posfai J, Roberts RJ, Wilson GG: The DNA (cytosine-5) methyltransferases. *Nucleic Acids Res* 1994, 22:1-10.

Li E, Bestor TH, Jaenisch R: Targeted mutation of the DNA methyltransferase gene results in embryonic lethality. *Cell* 1992, 69:915-926.

Lippman Z, Gendrel AV, Colot V, Martienssen R: Profiling DNA methylation patterns using genomic tiling microarrays. *Nat Methods* 2005, 2:219-224.

Lippman Z, Martienssen R: The role of RNA interference in heterochromatic silencing. *Nature* 2004, 431:364-370.

Macrae IJ, Zhou K, Li F, Repic A, Brooks AN, Cande WZ, Adams PD, Doudna JA: Structural basis for double-stranded RNA processing by Dicer. *Science* 2006, 311:195-198.

Maloisel L, Rossignol JL: Suppression of crossing-over by DNA methylation in Ascobolus. *Genes Dev* 1998, 12:1381-1389.

Matzke M, Kanno T, Daxinger L, Huettel B, Matzke AJ: RNA-mediated chromatin-based silencing in plants. *Curr Opin Cell Biol* 2009.

Matzke M, Kanno T, Huettel B, Daxinger L, Matzke AJ: Targets of RNA-directed DNA methylation. *Curr Opin Plant Biol* 2007, 10:512-519.

Mello CC, Conte D, Jr.: Revealing the world of RNA interference. *Nature* 2004, 431:338-342.

Miura A, Yonebayashi S, Watanabe K, Toyama T, Shimada H, Kakutani T: Mobilization of transposons by a mutation abolishing full DNA methylation in Arabidopsis. *Nature* 2001, 411:212-214.

Mochizuki K, Gorovsky MA: A Dicer-like protein in Tetrahymena has distinct functions in genome rearrangement, chromosome segregation, and meiotic prophase. *Genes Dev* 2005, 19:77-89.

Myant K, Stancheva I: LSH cooperates with DNA methyltransferases to repress transcription. *Mol Cell Biol* 2008, 28:215-226.

Nowack MK, Shirzadi R, Dissmeyer N, Dolf A, Endl E, Grini PE, Schnittger A: Bypassing genomic imprinting allows seed development. *Nature* 2007, 447:312-315.

Nowak SJ, Corces VG: Phosphorylation of histone H3: a balancing act between chromosome condensation and transcriptional activation. *Trends Genet* 2004, 20:214-220.

Ponger L, Li WH: Evolutionary diversification of DNA methyltransferases in eukaryotic genomes. *Mol Biol Evol* 2005, 22:1119-1128.

Riggs AD: X inactivation, differentiation, and DNA methylation. *Cytogenet Cell Genet* 1975, 14:9-25.

Ruthenburg AJ, Li H, Patel DJ, Allis CD: Multivalent engagement of chromatin modifications by linked binding modules. *Nat Rev Mol Cell Biol* 2007, 8:983-994.

Selker EU: Epigenetic phenomena in filamentous fungi: useful paradigms or repeat-induced confusion? *Trends Genet* 1997, 13:296-301.

Selker EU, Tountas NA, Cross SH, Margolin BS, Murphy JG, Bird AP, Freitag M: The methylated component of the Neurospora crassa genome. *Nature* 2003, 422:893-897.

Shen L, Kondo Y, Guo Y, Zhang J, Zhang L, Ahmed S, Shu J, Chen X, Waterland RA, Issa JP: Genome-wide profiling of DNA methylation reveals a class of normally methylated CpG island promoters. *PLoS Genet* 2007, 3:2023-2036.

Shi H, Tschudi C, Ullu E: An unusual Dicer-like1 protein fuels the RNA interference pathway in Trypanosoma brucei. *RNA* 2006, 12:2063-2072.

Shi Y, Lan F, Matson C, Mulligan P, Whetstine JR, Cole PA, Casero RA: Histone demethylation mediated by the nuclear amine oxidase homolog LSD1. *Cell* 2004, 119:941-953.

Singh-Gasson S, Green RD, Yue Y, Nelson C, Blattner F, Sussman MR, Cerrina F: Maskless fabrication of light-directed oligonucleotide microarrays using a digital micromirror array. *Nat Biotechnol* 1999, 17:974-978.

Sterner DE, Berger SL: Acetylation of histones and transcription-related factors. Microbiol *Mol Biol Rev* 2000, 64:435-459.

Trojer P, Li G, Sims RJ, 3rd, Vaquero A, Kalakonda N, Boccuni P, Lee D, Erdjument-Bromage H, Tempst P, Nimer SD, et al: L3MBTL1, a histone-methylation-dependent chromatin lock. *Cell* 2007, 129:915-928.

Vaucheret H: Plant ARGONAUTES. *Trends Plant Sci* 2008, 13:350-358.

Vaughn MW, Tanurdzic M, Lippman Z, Jiang H, Carrasquillo R, Rabinowicz PD, Dedhia N, McCombie WR, Agier N, Bulski A, et al: Epigenetic natural variation in Arabidopsis thaliana. *PLoS Biol* 2007, 5:e174.

Vazquez F, Vaucheret H, Rajagopalan R, Lepers C, Gasciolli V, Mallory AC, Hilbert JL, Bartel DP, Crete P: Endogenous trans-acting siRNAs regulate the accumulation of Arabidopsis mRNAs. *Mol Cell* 2004, 16:69-79.

Voorhoeve PM, Agami R: Knockdown stands up. *Trends Biotechnol* 2003, 21:2-4.

Wang Y, Wysocka J, Sayegh J, Lee YH, Perlin JR, Leonelli L, Sonbuchner LS, McDonald CH, Cook RG, Dou Y, et al: Human PAD4 regulates histone arginine methylation levels via demethylimination. *Science* 2004, 306:279-283.

Zhang Y, Reinberg D: Transcription regulation by histone methylation: interplay between different covalent modifications of the core histone tails. *Genes Dev* 2001, 15:2343-2360.

Zhao Y, Srivastava D: A developmental view of microRNA function. *Trends Biochem Sci* 2007, 32:189-197.

Zilberman D, Gehring M, Tran RK, Ballinger T, Henikoff S: Genome-wide analysis of Arabidopsis thaliana DNA methylation uncovers an interdependence between methylation and transcription. *Nat Genet* 2007, 39:61-69.

Zofall M, Grewal SI: Swi6/HP1 recruits a JmjC domain protein to facilitate transcription of heterochromatic repeats. *Mol Cell* 2006, 22:681-692.

Chapter V:

Identification and analysis of transposable elements in the genome of the brown alga *Ectocarpus siliculosus*

5.1 Introduction

Following the analysis of transposable elements in the *P. tricornutum* and *T. pseudonana* genomes, I was interested to investigate the genome sequences of closely-related organisms. A First was to determine whether CoDiI and CoDiII-like elements are indeed diatom-specific. I was also interested to examine the presence/absence and abundance of different classes of TEs and compare it with what was observed in diatoms.

Brown algae (Phaeophyceae) constitute a large group of mostly marine multicellular algae, including many types of seaweed. Phaeophyceae belong to the stramenopile (heterokont) lineage and are believed to have separated from diatoms around 250 Mya. They are unique among heterokonts in their ability to develop multicellular forms with differentiated tissues. The giant kelps, for example, may reach 60 m in length and exhibit a similar complexity to some land plants. The developmental complexity of brown algae is of interest because, although multicellularity has evolved many times in diverse lineages of the tree of life, very few of these lineages have produced organisms with complex body plans. Brown algae are usually the dominant organisms in shoreline marine ecosystems in terms of biomass, often forming extensive forests. Industrial production of brown algae is a growing industry, with most of the biomass being produced by aquaculture. The major uses are as human and animal food, but there is growing interest to use brown algae for biofuel production, for which they would have the advantage of high productivity without competing with terrestrial crops for farmland.

Genomic and genetic approaches applied to brown algae have long been impractical because of their large genomes and because it is difficult to complete their life cycles in the laboratory. To address this problem, a survey of brown algal species was carried out in the Station Biologique of Roscoff, France (Peters et al., 2004a), which led to the choice of the filamentous alga, *Ectocarpus siliculosus* (Figure 5.1), which is widely distributed in temperate

seas, as a model species for this group. This choice was based on several characteristics including small size, the fact that the entire life cycle can be completed in Petri dishes (as with many brown algae, the *E. siliculosus* life cycle is divided into the diploid sporophyte and the haploid gametophyte phases) (Müller et a.l, 1998), the ease with which genetic crosses can be carried out (Peters et al., 2004b), and the relatively small size of the genome (about 200 Mb).

The draft sequence of the *E. siliculosus* nuclear genome was recently generated in Genoscope by whole-genome shotgun sequencing corresponding to more than 10-fold coverage. The assembled sequence of the nuclear genome consists of 1569 scaffolds that were subsequently used by the platform of Bioinformatics and Evolutionary Genomics in the University of Gent, Belgium, for automated gene prediction and annotation. A consortium involving researchers belonging to different laboratories from around the world was established in September 2007 in order to complete as much as possible manual annotation and computer-based analyses aimed at understanding what this genome is made of and how it evolved. In the framework of this genome annotation consortium, I was nominated together with Hadi Quesneville to take charge of the identification and annotation of the repeated fraction of the genome. This aspect is of particular importance in the case of such middle-sized genomes since repeated sequences con constitute as much as 20% of genomes of similar size such as that of *Drosophila* (Quesneville et al., 2003). In addition to being important to help understanding genome composition and evolution, the identification of repeated sequences also allows to detect putative erroneous gene models built on open reading frames corresponding to transposable elements and introducing a bias in the composition of the gene pool. With a 200 Mb genome to examine, the need for powerful genome-wide computational methods to indentify all repeated sequences is required.

Figure 5.1 : An *Ectocarpus* sporophyte (picture courtesy of Delphine Scornet, UMR7139, Roscoff, France).

5.2 Results and Discussion

5.2.1 Identification of repeated sequences in the *E. siliculosus* genome

With the *E. siliculosus* genome as input, TEdenovo produced a total of 3,792 consensus sequences. 1,432 of these were classified as putative TEs of a specific family and the remaining were not annotated (NoCat). At this step, the TEdenovo output was compared to the *E. siliculosus* gene models and we found that over 3,200 predicted genes matched to repeated sequences. This data together with integration of improved intron initiation/termination signals was used to reiterate gene prediction in Gent.

At the end of the manual screening of the clusters, the *E. siliculosus* TE library consisted in 219 distinct sequences, including 31 NoCat. We tried to annotate or clean the initial set of consensus sequences from the TEdenovo output using this library. To do this, we compared our library of TEs to the TEdenovo output by BLAST and removed all hit sequences from the latter set (with e value < -15). All the remaining sequences were compared to the Repbase, Genbank, and tRNA library datasets by BLAST which enabled identification of a few other TEs that were removed from the TEdenovo output, annotated, and pasted into the TE library. Following this, the TEdenovo set was composed of 721 sequences with no similarity to elements from our library or any known TEs.

To validate this strategy, I did a series of small scale tests on a genome sample composed of 20 scaffolds (29 Mb). Using the initial TEdenovo output as library (3792 sequences) to run the RepeatMasker programme, I found that these consensus sequences make up ~23.2% of the test sample. I then used the *E. siliculosus* TE library to run RepeatMasker and found that the sequences comprising this set of sequences make up ~15.7% of the same sample. Using the trimmed TEdenovo output (721 sequences) to run the RepeatMasker programme, I found that this set of sequences make up ~7.7% of the sample.

Thus, in total, the TE library plus the trimmed TEdenovo set of sequences cover ~23.4% of the sample.

We considered this result sufficiently consistent with the initial 23.2% (at least in terms of minimal redundancy) and validated this split into two libraries: one called EsFull made of manually curated sequences including some NoCat, and the second called Tedenovo_minus_EsFULL consisting of unknown sequences. The EsFULL library was made of different LTR-retrotransposons including Ty1/copia, Ty3/gypsy, and DIRS/Ngaro-like elements, non-LTR retrotransposons and non autonomous Class I elements (TRIMs/LARDs), as well as subclass I DNA transposons including Harbinger, JERKY, and POGO-like elements, subclass II elements (Table 5.1). This library also contains putative TEs, the sequence of the *E. siliculosus* double-stranded DNA virus EsV-1, whose full length is ~335 kb (Delaroque and Boland, 2008), and various tandem repeats. During the establishment of this library, we identified over ten different consensus sequences that clearly corresponded to various genes including those coding for fucoxanthin, cell wall adhesion proteins, and ubiquitin. These sequences were re-inserted into the gene model repertoire.

Table 5.1: Number of sequences annotated for different types of TEs

Class	Subclass	Type	Number of sequences
Class I	LTR retrotransposons	LTR GYPSY	21
		LTR COPIA	23
		LTR DIRS	12
	Non-LTR retrotransposons	LINE	5
	Non-autonomous	TRIM/LARD	12
Class II	Subclass I	TIR	20
		TIR putative	3
		non autonomous TIR	34
	Subclass II	HELITRON	6
Others	Virus	*E. siliculosus* virus 1	1
	Structural feature	TE putative	25
		Tandem repeat	26
		Unclassified repeat	31
Total			219

5.2.2 Masking the *E. siliculosus* genome

We then entered the second phase of the REPET pipeline procedure with the TEannot sub-pipeline. This step is used to map a library of repeats on the genome using different BLAST-based algorithms including Blaster, Censor, and RepeatMasker, in order to optimize the masking as much as possible (Quesneville et al., 2005). TEannot also launches several mini-satellite detection programmes including TRF, Mreps, and RepeatMasker-SSR. The results are then combined under different file formats for further analysis. We found that our libraries masked a total of ~ 45 Mb, thus about 22.7% of the genome, which is very close to the value obtained with RepeatMasker in the test sample consisting of 20 scaffolds.

We then analyzed the contribution to the genome of the different classes of repeats identified. Note that from now, we considered the sequences annotated "putative TEs" as unclassified repeats, because the former implies the latter. In total, we found that the TEannot output was made of 41.3% of class I elements and 12.3% of class II elements, and that the remaining 46.4% corresponded to other sequences including 1.2% from Esv-1. It appears that LTR-RTs constitute the most abundant subclass, representing over 26% of the *E. siliculosus* repeat complement. This includes 10.6% from Ty1/copia-like elements, which are the most profuse, 9.6% from Ty3/gypsy-like elements, and 6.4% from DIRS/Ngaro-like elements. LINE-like non-LTR-RTs appear to contribute 9% of the repeat complement while putatively autonomous elements with TIRs are less abundant than the putative non-autonomous copies (Figure 5.2 and Table 5.2).

Figure 5.2: Pie chart representing the contribution of the different classes of repeated sequences identified to the *Ecotocarpus* repeat complement.

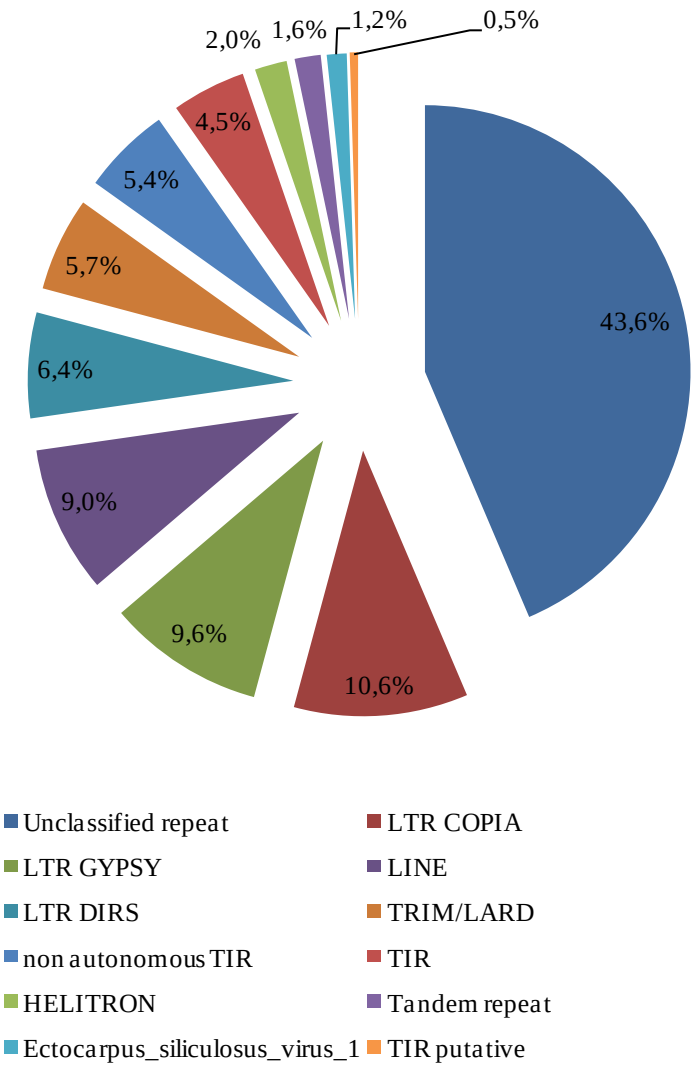


Table 5.2: Contribution of the different types of repeats found to the *Ectocarpus* repeat complement.

Class	Type	coverage (kb)	% of output
Class I	LTR GYPSY	4,348	9.6%
	LTR COPIA	4,817	10.6%
	LTR DIRS	2,918	6.4%
	LINE	4,075	9.0%
	TRIM/LARD	2,600	5.7%
	Total	**18,758**	**41.3%**
Class II	TIR	2,042	4.5%
	TIR putative	228	0.5%
	non autonomous TIR	2,445	5.4%
	HELITRON	892	2.0%
	Total	**5,607**	**12.3%**
Others	*E. siliculosus* virus 1	549	1.2%
	Tandem repeat	721	1.6%
	Unclassified repeat	19,817	43.6%
	Total	**21,087**	**46.4%**

We then examined the TEannot output at the family level. Table 5.3 shows the 10 most abundant families in the genome. Although LINE elements constitute only 9% of the repeat complement, three families of LINE elements: EsRTE2, EsRTE3, and EsRTE4 are among the ten most abundant repeated sequences in the genome. We also found that one family of each class of LTR-RT identified appears in this list: EsNgaro1, EsGypsy2, and EsCopia2 as well as a sequence related to the large retrotransposon derivative EsLARD4_1. Interestingly, the three remaining top 10 sequences are unclassified repeats. Two of them, including top 1, were initially annotated as “Sower-related”. In fact there are several elements related to “Sower”. Sower originally got its name from its apparent highly abundant distribution, sown all over the genome. It appears that Sower integrated within different repeated sequences: the Sower-

related unclassified repeats. Sower is usually flanked by unrelated sequences, suggesting its physical autonomy with respect to the Sower-related unclassified repeats. Sower is 676 nucleotides long, starting and ending with a repeat of ~10 cytosines. Its 5' extremity consists of one, two, or three (full length) repeats of a ~67bp motif starting by a stretch of cytosines, and a 3' end also consisting in a stretch of cytosines (Figure 5.3A). Analysis of several copies of Sower failed to identify target site duplication and any conserved open reading frame. RNA secondary structure prediction for Sower resulted in a stable (free energy of – 387 kcal/mol) rod-like shape branched with numerous hairpins (Figure 5.3B). We used BLASTN (word size of 16) to compare Sower to Genbank. Interestingly, we found significant similarity (78% identity over ~150 nucleotides) with a sequence annotated as a microsatellite from *Undaria pinnatifida* which also belongs to Phaeophyceae. This sequence indeed contains 9 dinucleotide (CA) repeats. However, when aligning the *U. pinnatifida* with a set of Sower sequences, we observed that instead, significant similarity was found in the 3' region of both sequences (Figure 5.4). The nature of the Sower repeat and the way it amplifies is thus still poorly understood, although it covers about 2.5 Mb (1.2 %) of the genome. As such, it is likely to constitute an important potential source of recombination in the *E.siliculosus* genome.

Table 5.3: Presence of the most abundant TE families in the *Ectocarpus* genome

Rank	Family	Type	coverage (kb)
1	EsSower_related_1	Unclassified repeat	1,954
2	EsRTE3	LINE	1,760
3	EsRTE2	LINE	1,450
4	EsNoCat	Unclassified repeat	1,314
5	EsSower_related_2	Unclassified repeat	1,135
6	EsNgaro1	LTR DIRS/Ngaro	1,011
7	EsGypsy2	LTR GYPSY	763
8	EsLARD4_1_related	TRIM/LARD	761
9	EsCopia2	LTR COPIA	709
10	EsRTE4	LINE	688

Figure 5.3: Schematic representation of the Sower element. (A) Sower with simple, double, and triple 67 bp repeat. The 5' repeat units are indicated as R1, R2, and R3. Stretches of cytosines are shown in bold red (B) Schematic view of Sower secondary RNA structure.

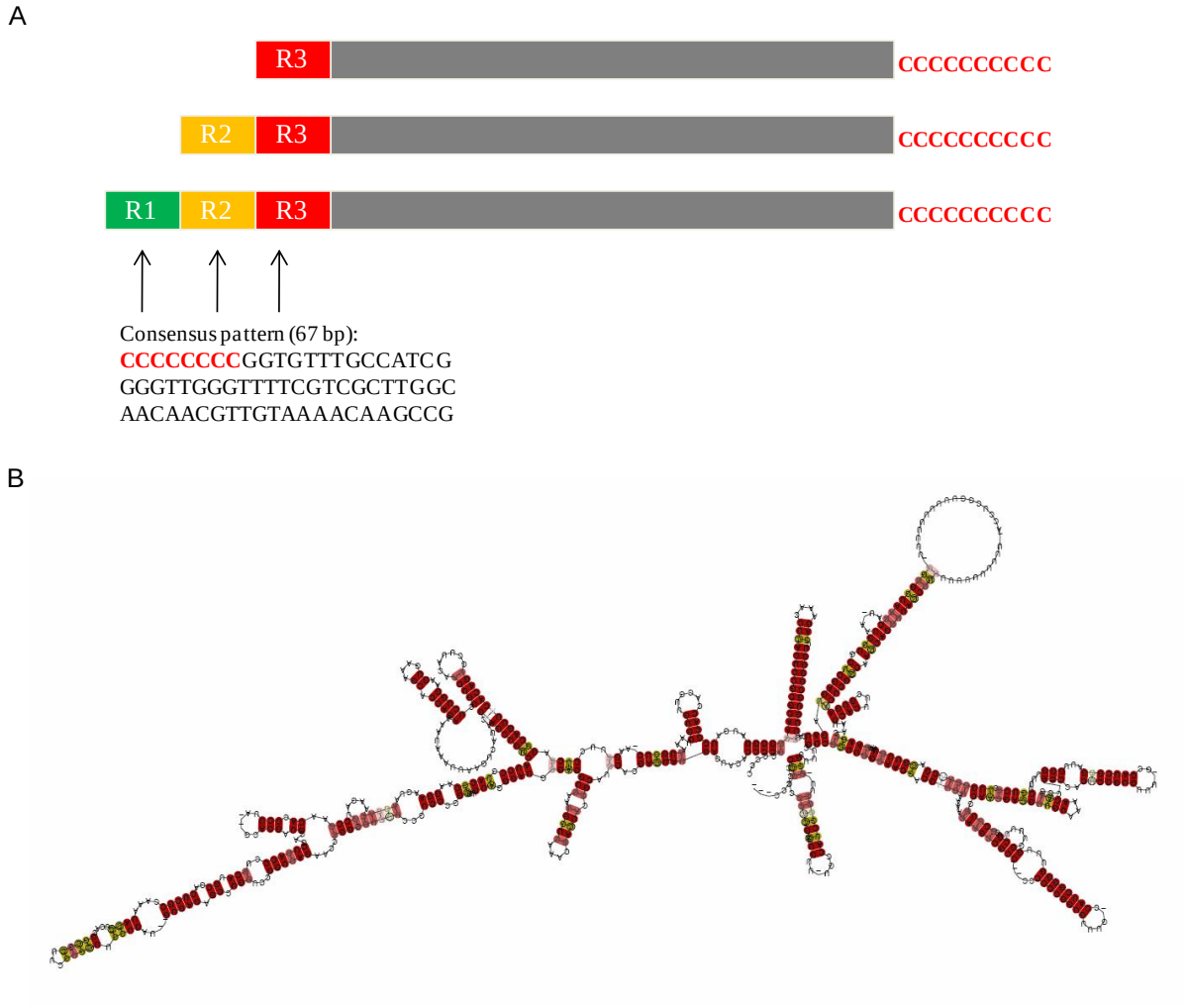


Figure 5.4: CLUSTALW multiple alignement of the nucleotide sequences from three copies of the *E. siliculosus* Sower element and a similar sequence found in *U. pinnatifida.*

5.2.3 Phylogeny of *E. siliculosus* LTR-retrotransposons

Relevant to the previous work analyzing the phylogeny of LTR-RT from diatom and marine metagenomes (Maumus et al., 2009), we examined the relationships between those from *E. siliculosus* and the various lineages mentioned above. To do this, we identified and extracted RT domains from the various *E. siliculosus* LTR-RTs identified and used a relevant subset to construct a phylogenetic tree from a multiple alignment with markers from the various RT lineages mentioned above, including CoDiI, CoDiII, and the RAS lineage (Figure 5.5). The integration of these sequences in our reference RT tree further diversifies the previous Ty1, Copia, Ty3/Gypsy, DIRS, BEL, CoDiI, CoDiII, and RAS classification of LTR-RTs. We observed that EsCopia-like elements fit into two distinct clusters: EsC1 and EsC2. EsC1 appears more closely related to the Copia lineage and EsC2 forms a new distinct lineage. It also appears that the addition of the *E. siliculosus* sequences caused the separation of RT domains from CoDi6 elements into a distinct lineage which appears to be diatom-specific. In this tree, we conserve the RAS, Ty1, CoDiI, CoDiII, and BEL lineages. We also noticed that the DIRS/Ngaro-like sequences from *E. siliculosus* cluster within the DIRS lineage. EsGypsy-like elements are found in four different clusters: EsG1 elements fit within the Ty3/Gypsy lineage, EsG2 appears closely related to the Ty3/Gypsy lineage, EsG3 is more closely related to the Caulimoviridae lineage, and the element EsGypsy14 branches at the root of the GyDi lineage. We thus recognize in this tree further modifications of the LTR-RT phylogeny with the branching of EsC2 and CoDi6 elements into two new distinct lineages (Figure 5.5). These were called Brownies (Phaeo means brown in Greek) and CoDi III, respectively. Moreover, we observe a further phylogenetic remodeling with the accentuated split of Ty3/Gypsy-like which appears to cluster into two distinct lineages: on one side elements “true” Ty3/Gypsy-like elements which include EsG1 and EsG2, and on the other side Caulimoviridae together

with some so-called Ty3/Gypsy-like elements such as Cer1 and Skipper, and EsG3 elements which in fact appear to be more closely related to Caulimoviridae.

We then wanted to compare the diversity of RT domains found in *E. siliculosus* LTR-RTs to the genome of the Pelogophyceae *A. anophagefferens*. For this, we queried RT domains belonging to the different clusters of *E. siliculosus* LTR-RT versus the *A. anophagefferens* genome using TBLASTN and recovered the amino acid sequences from the best hit in each case. We recovered 3 Copia-like and 2 Gypsy-like RT sequences and integrated these into the multiple aligment used previously to construct a new phylogenetic tree (Figure 5.6). We found the the *A. anophagefferens* Copia-like elements belong to the Brownies lineage and that the Gypsy-like sequences branch close to the GyDi lineage. Therefore, Brownie elements likely evolved within the brown algal lineage after the separation with diatoms.

Figure 5.5: Phylogenetic tree showing the relationships between the reverse transcriptase domains from E. siliculosus LRT-RT elements, the major LTR-RT lineages, and retroviruses. The tree was constructed with the NJ method using the SplitsTree4 software (Huson and Bryant, 2006).

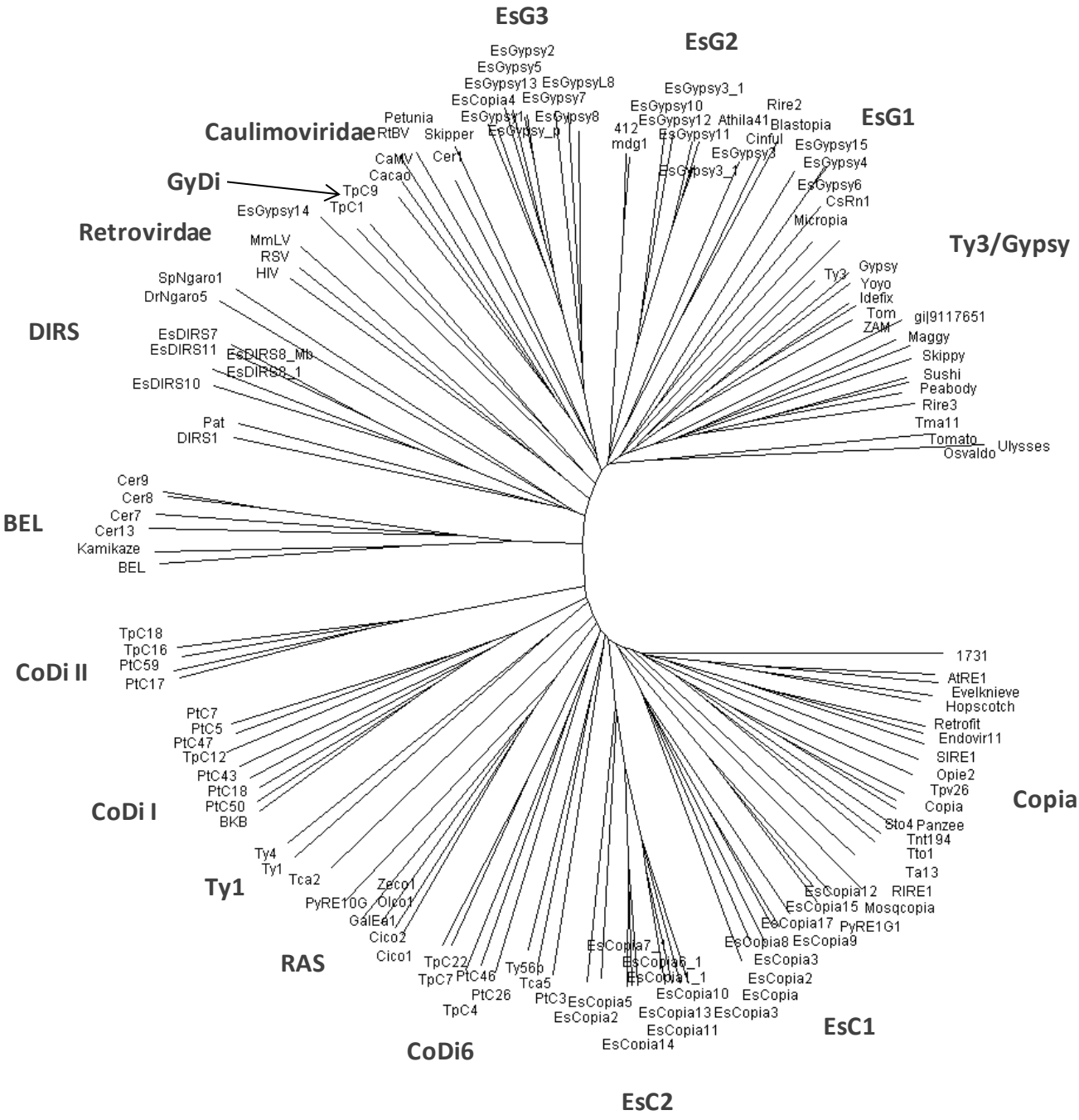
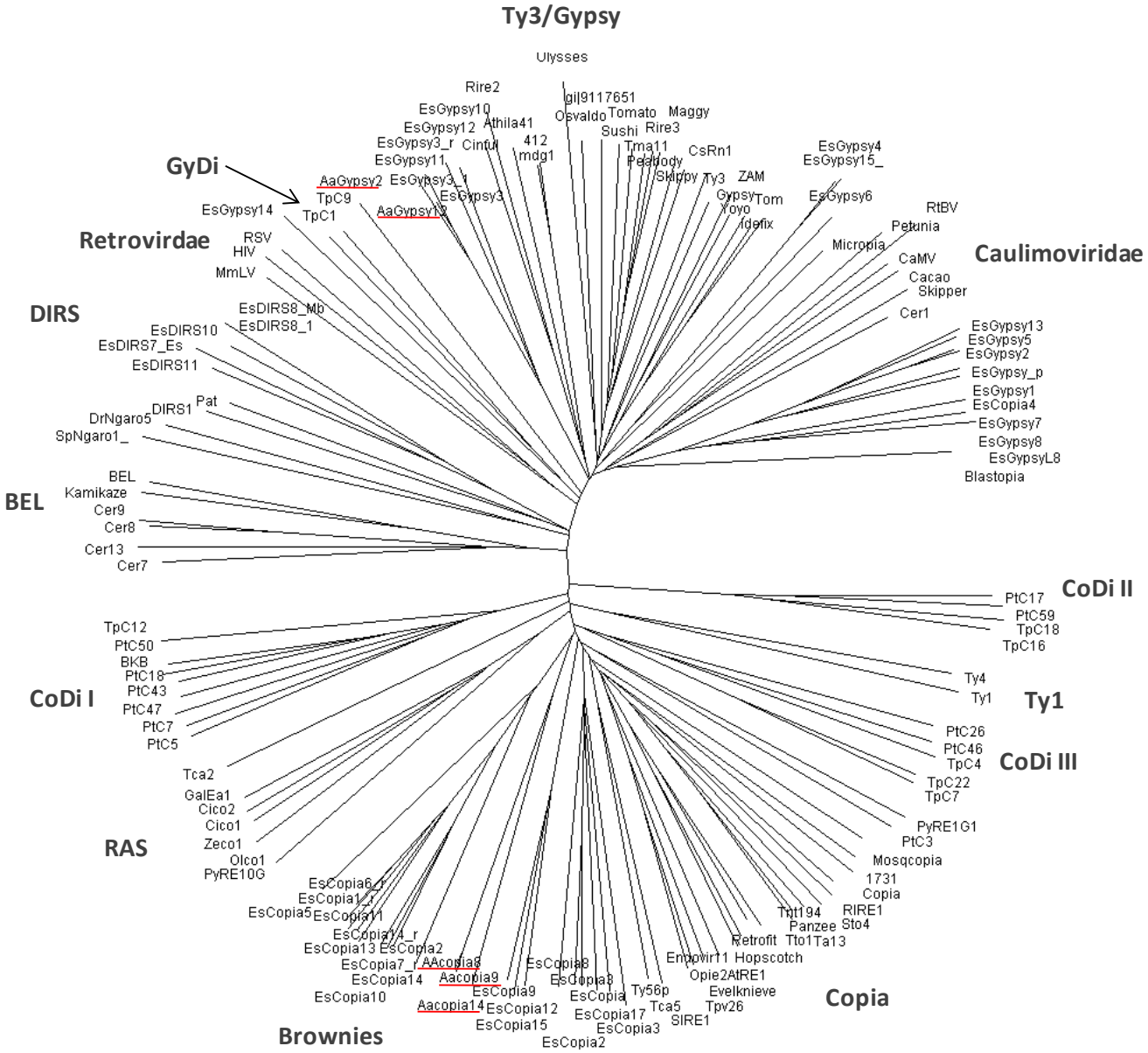


Figure 5.6: Phylogenetic tree showing the relationships between the reverse transcriptase domains from LRT-RT elements from Phaeophyceae, the major LTR-RT lineages, and retroviruses. The tree was constructed with the NJ method using the SplitsTree4 software (Huson and Bryant, 2006). The *A. anophagefferens* sequences are underlined in red.



5.2.4 Expression analysis

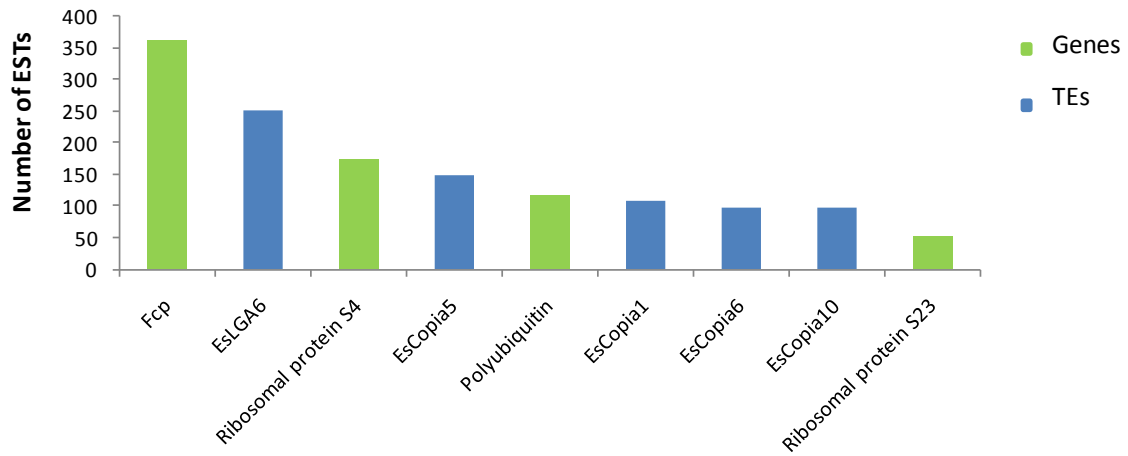
We conducted an expression analysis for the 98 best characterized (fully classified) TE families found in the Ectocarpus genome by BLASTN comparison to the Ectocarpus EST dataset obtained from tissues grown under normal laboratory conditions. We found that 5 TE families presented a significantly high number of corresponding ESTs: EsLGA6 (see below), EsCopia5, EsCopia1, EsCopia6, and EsCopia10 (Table 1). We also observed that loci corresponding to elements from 4 out of these 5 TE families were present among the 100 most highly expressed loci in the genome after transcription array analysis (Table 5.4). For comparison, we counted the number of ESTs for some genes present among the transcription array top100 which correspond mostly to house-keeping genes. We found a number of ESTs for these genes that was comparable to the ones found for the 5 TE families with abundant ESTs (Figure 5.7). Therefore the expression levels observed for these 5 TE families are similar to the ones observed for highly expressed genes leading us to think that such expression levels are high relative to Ectocarpus pool of mRNA. This is surprising because TEs are usually maintained silent under non-stressful conditions; some families being transcriptionally activated in response to specific stress (Capy et al., Heredity 2000). In some rare cases, TEs have been found to be transcriptionally active under non-stressful conditions but during a particular stage of the life cycle such as the Ty1 element which is expressed during the haploid phase in yeast (Elder et al., Cold Spring Harb Symp Quant Biol., 1981) or germ line-specific expression of intracisternal A-particle in mice (Dupressoir and Heidmann, Mol Cell Biol., 1996.). Here we observe high expression levels for 5 TE families under non-stressful (laboratory) conditions based on EST dataset prepared using tissues harvested during the dominant Ectocarpus life stage. We therefore think that the expression levels for these 5 TE families are unexpectedly high. Incidentally, these five TE had a recent peak of activity (peak age =0) according to copy age analysis (conducted by Hadi Quesneville). Also,

although we could not assign the ESTs found to specific elements because of the lack of enough sequence divergence, all these families contain a pool of potentially active non-disrupted elements. All together these 5 TE families likely represent good candidates for being still active today.

Table 5.4: Consitutively transcribed TE families

Family	Category	Number of ESTs	Among the top100 on tiling array	EST Frequency	Coverage (kb)
EsLGA6	TRIM/LARD	250	yes	3,02E-03	122,80
EsCopia5	LTR COPIA	149	yes	1,80E-03	188,84
EsCopia1	LTR COPIA	108	no	1,30E-03	306,22
EsCopia6	LTR COPIA	98	yes	1,18E-03	329,24
EsCopia10	LTR COPIA	97	yes	1,17E-03	365,28

Figure 5.7: Constitutive expression of 5 TE families. The y axis indicates the number of ESTs corresponding to the 5 most highly expressed TE families (blue bars) and to 4 highly expressed genes (green bars).



Interestingly, the most highly expressed TE family corresponds to a member of the new type of LTR-retrotransposons that we called large GAG-related elements (LGA) (see below). The other four highly expressed TE families all correspond to Ty1/copia-like LTR-retrotransposons (Table 5.4). Since the lack of sequence divergence hampers detailed expression analysis for individual copies, we addressed whether higher EST abundance for these TE families could reflect leaky basal transcription of the most abundant TE families in the genome. We plotted for each TE family assessed the number of EST found against its genome coverage and found that although there is indeed a slight trend towards higher expression levels of more abundant families, the 5 highly expressed TE families remarkably do not match the trend, especially the EsLGA6 and EsCopia5 families (Figure 5.8). Accordingly, the 5 highly expressed TE families have the highest "number of ESTs/coverage" ratio among the 98 families tested, which is predominantly high for the EsLGA6 and EsCopia5 families (Figure 5.9). Note that we measured EST number/TE family respect to TE family coverage and not respect to the number of elements in this analysis not to introduce any bias about what templates are used for transcription: full length or truncated copies. We therefore thought that genome coverage/TE family could best reflect the amount of potential DNA templates per family and would best challenge the even leaky transcription hypothesis.

Figure 5.8: Five significantly transcribed TE families. Each diamond represents one of 98 TE families which coordinates correspond to its number of ESTs on the y and its genome coverage (in kb) on the x axis.

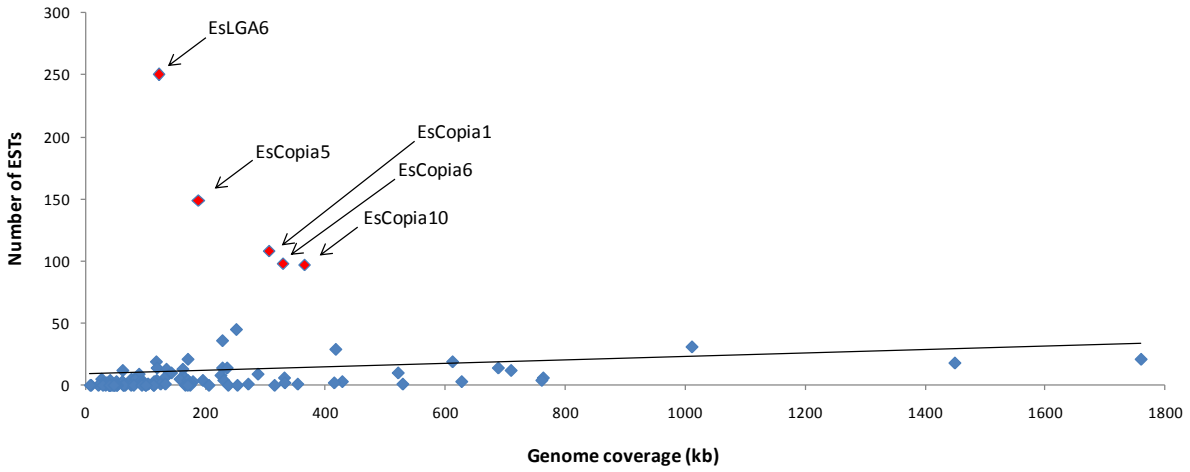
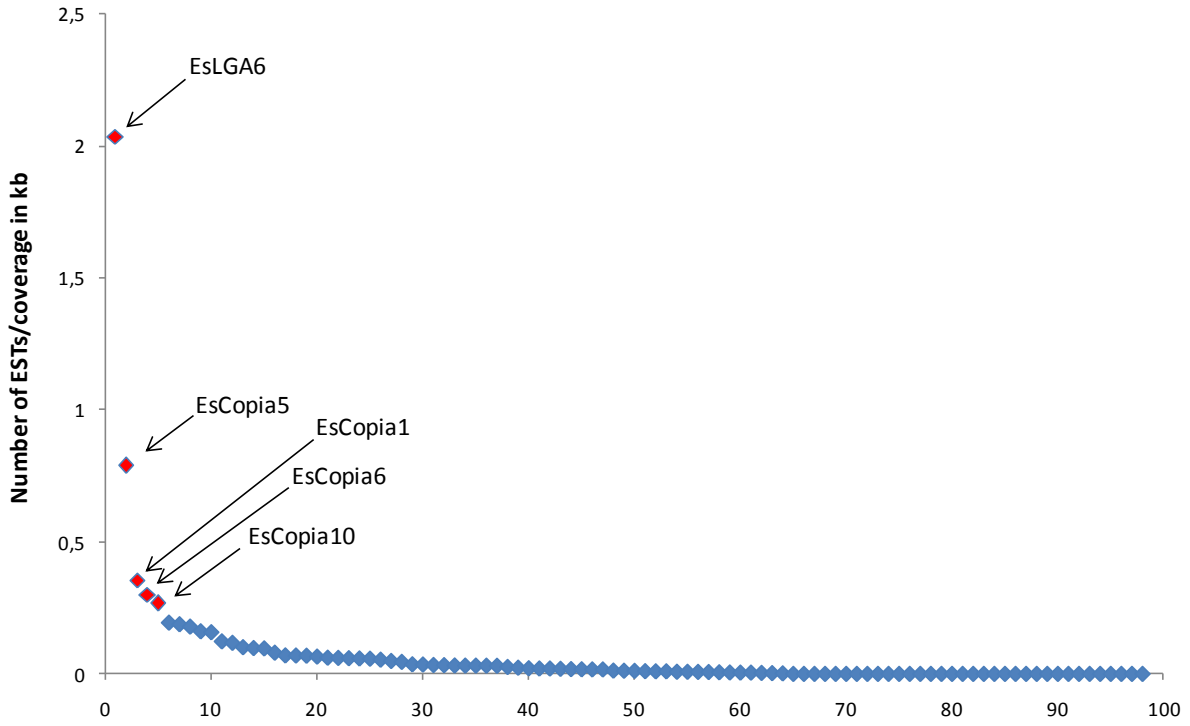
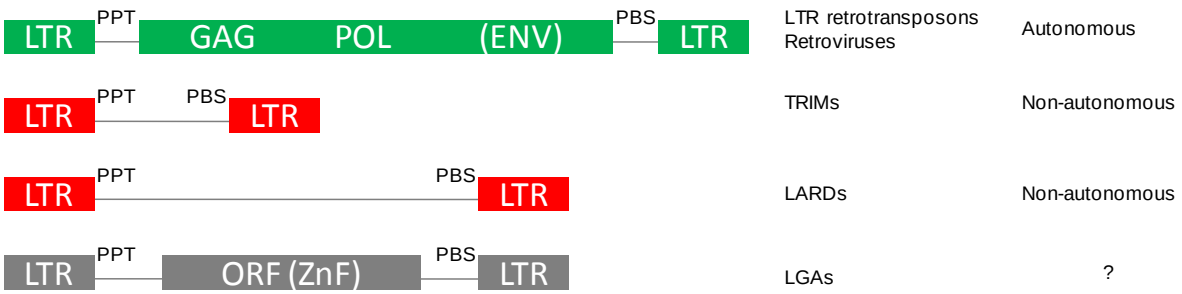


Figure 5.9: Expression/coverage ratio. Diamonds represent 98 TE families that are plot according to their ratio for transcription level (ESTs)/genome coverage.



ES LGA6 (also called EsLARD6) is not a typical large retrotransposon derivative, nor a terminal repeat retrotransposon in miniature (TRIM). TRIMs are composed of 100-250 bp terminal direct repeats which appear to be LTRs and the priming motifs (PBS and PPT) commonly found internal to the LTRS in full length LTR-RT and retroviruses (Witte et al., 2001). However, aside from the priming sites and a small intervening segment, the TRIM elements almost lack internal domains. LARDS were recently described in barley and other species of the Triticeae (Kalendar et al., 2004). Members of this group are similar to TRIMS in that they possess LTRs and priming motifs, and lack protein-coding domain. However, LARDS are unique in the presence of a large internal domain that is highly conserved despite its lack of coding capacity. EsLARD6 (3368 bp) is similar to LARDs in that its extremities consist in 100% identical 365 bp-long LTRs, it possesses priming sites, and contains a large internal domain. Nonetheless, EsLARD6 is peculiar because the internal sequence contains an open reading frame encoding a protein of 403 amino acids. Domain searches using InterProScan revealed that this protein contains a zinc finger (ZnF) at its C-terminus, which is commonly found within the GAG protein from LRT-RTs and retroviruses. BLAST comparison to Genbank and Repbase revealed that this protein has weak similarity with GAG from plant copia-like elements and human immunodeficiency virus (HIV), especially at the level of the ZnF (about 20 residues). The remainder of the protein did not show similarity to any known protein or TE. Thus, EsLARD6 is an element with LTRs and priming sites, is flanked by a 4 nucleotide target site duplication, and contains a large internal domain which encodes a protein containing a GAG-related ZnF. These elements are unlikely to be autonomous since they do not show similarity to any enzymes involved in the transposition process such as the reverse transcriptase. However, we cannot exclude that this ORF may encode an enzyme with an as yet unknown transposition process. This new class of elements was called LGA elements for LTR Gag–related elements (Figure 5.10).

Figure 5.10: Schematic representation of the structure of autonomous and non-autonomous LTR elements, including the newly identified LTR Gag–related elements (LGAs).

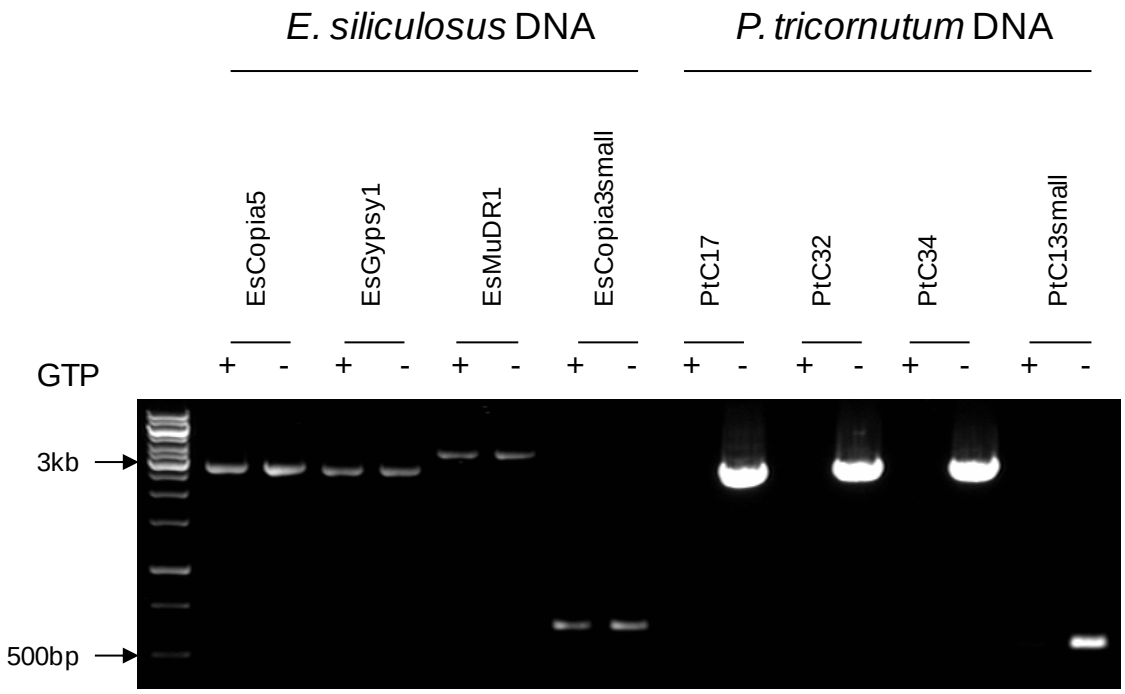


5.2.5 DNA Methylation in *Ectocarpus siliculosus*

I was intrigued by the high expression levels observed for several TEs. Indeed, the starting RNA material used for both the construction of the *E. siliculosus* EST library and the hybridization of the expression array was extracted from non-stressed cultures in which conditions TEs are usually maintained silent through chromatin-based processes. This silencing commonly involves DNA methylation of transposable elements as observed in diatoms. TE activation is instead usually accompanied by hypomethylation of the locus of interest. Considering the unusual TE expression levels in *E. siliculosus*, I wondered whether this may be due to abnormal DNA methylation in this species.

Prior to sequencing, HPLC analysis of the base composition of the *E. siliculosus* nuclear genome did not detect methylated cytosines. Since HPLC may not detect traces of methylation in the genome, I wanted to verify experimentally if TEs, which are commonly the most highly methylted sequences in a genome, were marked by DNA methylation in this species. I obtained DNA extracted from *E. siliculosus* cultures grown under normal conditions from Mark Cock's laboratory in Roscoff, France. The DNA was digested with McrBC and I proceeded to use Mcr-PCR to determine whether DNA had been cut at transposon loci. To make sure that a negative result would not reflect inefficient experimental conditions, I incubated in the same tube the same quantity of *P. tricornutum* DNA as positive control for the McrBC digestion reaction. We assessed the methylation status of EsCopia5 which has a high EST frequency and is highly represented among the top 100 expression array peaks, and EsGypsy1, EsMuDR1, and EsCopia3 which have low expression levels with 21, 14, and 4 ESTs, respectively. Interestingly, we observed that highly expressed elements as well as almost silent elements are not sensitive to McrBC digestion (Figure 5.11), indicating the absence of methylated cytosines in their sequences.

Figure 5.11: Picture of agarose gel showing the incapacity of the McrBC enconuclease to cut at TEs loci in *E. siliculosus* compared to *P. tricornutum* loci.



This data is consistent with the apparent lack of putative DNA methyltransferase in the E. siliculosus gene repertoire. It thus appears that, as a few other eukaryotes (Unikonts) such as C. elegans, Ectocarpus lacks this component of chromatin-based silencing machinery. Ectocarpus is therefore the first non-Unikont eukaryote found to lack DNA methylation which is very interesting in a evolutionary perspective since we can expect that the chromatin-based processes involved may be quite different to those reported in other eukaryotes reported to lack DNA methylation. Deep sequencing of Ectocarpus small RNAs in the framework of the Ectocarpus genome consortium proved the presence of a functional sRNA pathway in this species and the fact that most of them are base-pairing with TEs suggests that TEs regulation is nonetheless likely to be chromatin-based. Therefore, the study of the epigenetic mechanisms involved in regulating chromatin compaction and TE silencing in Ectocarpus could be of particular interest.

5.3 Material and Methods:

Mcr-PCR

For McrBC digestion, 500 ng of *E. siliculosus* DNA together with 500 ng of *P. tricornutum* DNA was incubated for 4 hours at 37°C with 50 units McrBC endonuclease supplemented with 100 μg/ml bovine serum albumin and 1 mM guanosine triphosphate. Negative controls were obtained with the same experimental procedure but replacing guanosine triphosphate with water. The enzyme was subsequently inactivated by incubation at 65°C for 10 minutes. Digestion efficiency of different transposable elements loci was measured by semi-quantitative PCR. For Ectocarpus, we used the primers FwEsCopia5 5'-GCACACTAAACGCGGTACAA-3', RevEsCopia5 5'-CTGCAACACCGTTGTATTGG-3', FwEsGypsy1 5'-CATGCGGAGGTTACTTCCAT-3', RevEsGypsy1 5'-AGAGCAGGCACTAACCGAAA-3', FwEsMuDR1 5'-ACAGGGTAAATGCCAAGTCG-3', RevEsMuDR1 5'-CTGACGCACACACCTGAACT-3', FwEsCopia3 5'-CGGTGCTCACGACTGTCTTA-3', and RevEsCopia3 5'-TTCTTCCATTCGGAAGTTGG-3'.

Identification of Repeated Sequences

We used the REPET pipeline (Quesneville et al., 2005) to identify repeated sequences in the *E. siliculosus* genome. Hadi Quesneville is the head of URGI (Unité de Recherche Génomique Info) in Versailles, France. His team has developed over the last years a powerful bioinformatics pipeline aimed at identifying all repeated sequences in eukaryotic genomes. This pipeline, called REPET, is made of two sub-pipelines, TEdenovo and TEannot, which are dedicated to the detection, annotation and analysis of repeats in genomic sequences and

specifically designed for transposable elements. TEdenovo identifies consensus TEs and classifies them. It works with five successive steps:

Step 1- Find redundancy by alignment of the input genome on itself and selection of high scoring pairs (HSPs)

Step 2- Assignement of groups containing similar HSPs using three different methods: PILER, GROUPER, and RECON according to sequence similarity and length.

Step 3- Build a consensus sequence for each group derived from a multiple alignment of HSPs.

Step 4- Detect structural features such as LTRs or TIRs for each consensus and compare it to known TEs from the Repbase library

Step 5- Classify each consensus according to sequence and structural features and generate an annotated output library of repeated sequences.

Manual Annotation

Automatic TE annotation has several problems. Although significant similarity to conserved catalytic domains is a good indication to affiliate a consensus to a specific type of TEs, assumptions based on structural similarities such as the presence of a poly-A tail are more speculative. The method therefore detects false positives as well as false negatives, after closer examination. Furthermore the length of a consensus may not correspond to the full length of a TE due to non-optimal programme settings for each particular case. Finally, a consensus can correspond to a family of recently duplicated genes and these must be identified and specified in order to restore corresponding gene models instead of having them

annotated as “repeated sequences”. For all these reasons the TEdenovo output must be subjected to further examination and curation.

It is obviously not feasible to manually annotate all single 3,792 consensus sequences. In order to facilitate the annotation of these consensus sequences, the REPET team elaborated 197 clusters of consensus presenting sequence similarities using the program GROUPER. Note that at this stage, each consensus can be affiliated to one or several groups if it contains similarities with several sets of consensus sequences. I was initially supposed to verify and annotate each of these clusters. To do this, I designed a manual annotation strategy and proceeded with in this way, assuming that all sequences grouped in cluster were related. This manual curation of the clusters required the following steps:

Step 1- Select the longest sequence found in a cluster because it is most likely to represent the full length representative consensus. Make sure that this consensus indeed constitutes the full length of the repeat of interest by three successive steps:

Perform BLASTN analysis of the consensus sequence on the genome

Extract the genomic positions corresponding to the five best hits together with 5 kb of 5’ and 3’ flanking sequences.

Make a multiple alignment of the sequences recovered and determine the full length of the sequence of interest on the basis of the block of conserved nucleotides. Extract the full length sequence.

Step 2- Compare the sequence by BLAST to the *E. siliculosus* repeats library which includes a small library of TEs from the *E. siliculosus* genome that I previously generated manually using the methods I had applied to diatom genomes. Go to next step if there was no significant similarity in sequence and length with the library.

Step 3- Search for structural features in selected sequence by two different means:

Compare the sequence by BLASTN versus itself to identify putative LTRs, TIRs, palindromes, or short repeats.

Compare the sequence by BLASTN against a library of tRNA that was elaborated using tRNA scan-SE (http://lowelab.ucsc.edu/tRNAscan-SE/), an online tool for identifying tRNAs in query sequences.

Step 4- Search for similarities to known genes or TEs in two different ways:

Translate the consensus in the six frames (with minimum open reading frame length of 150 aa) and compare it to Genbank by BLASTP.

Compare the sequence directly using TBLASTX against the Repbase library.

Step 5- Integrate the results of the steps above and annotate the sequence accordingly as a transposable element if it shows sequence similarity to known TEs, as a putative TE if it contains at least one structural feature, as an unknown repeated sequence if all above steps yielded no results, or as a putative gene if it shows similarity to genes annotated in Genbank.

The REPET pipeline is still undergoing improvements and it is frequently updated thanks to the comments of the curators. As expected in such pioneering situations I noticed several errors and inconsistencies in several steps of the TEdenovo and GROUPER phases during manual annotation. I will describe a few of such examples to justify why and how our initial TE annotation strategy evolved. For example, cluster number 108 is made of two consensus sequences, both annotated by TEdenovo as NoCat. The first one called Recon_552_20 is almost identical to an element that I had manually annotated as a putative DIRS/Ngaro-like element and the second (called Recon_533_14) shows no similarity at all to this element. Surprisingly when comparing this Recon_533_14 (> 5kb) by BLAST against the genome, the

best hit is only 70 bp long with 93% sequence identity to the query. I called this a ghost consensus that was probably simply invented by the computer because of an unknown problem in the script it was processing. It also appears that another cluster (cluster 188) comprises a consensus called Recon_533_14 but the sequence in this case is completely different to the one found in cluster 108 and in fact is highly similar to the DIRS/Ngaro-like element from the latter and Recon_552_20. Further, it appears that among nine other consensus sequences in cluster 188, eight show no significant similarity to the DIRS/Ngaro-like element and could not be found by BLAST search in the genome, thus constituting additional ghost consensus sequences. Another major problem with the grouping step was that one of the clusters, cluster 146, contained over 2,500 consensus sequences. As a consequence, only 1,227 consensus sequences were assigned a cluster other than cluster 146. In view of these inconsistencies and uncertain groupings, we decided not to consider the cluster classification and to adapt the annotation strategy to the situation.

We used CLUSTALW (Higgins and Sharp, 1998) or MUSCLE (Edgar, 2004) for multiple sequence alignments. Phylogenetic trees were constructed using the Splitstree program (Huson and Bryant, 2006) using the neighbor joining method. Structural features from repeated sequences were examined using the dot plot on line program (http://www.vivo.colostate.edu/molkit/dnadot/) and the BLAST2seq program (http://www.animalgenome.org/blast/wblast2.html). Protein domains were searched using InterProScan online tool (http://www.ebi.ac.uk/Tools/InterProScan/index.html). For epression analysis, we compared the *Ectocarpus* TE library to the *Ectocarpus* "non-stressed" cDNA library by BLASTN. Significant hits (evalue < E-20 and > 95% identity) were selected and each hit was attributed to only one TE family (highest identity). Frequency was then calculated for each TE family as the number of corresponding ESTs divided by the total number of ESTs in the library. The loci corresponding to the top 100 positive peaks identified

from the analysis of the *Ectocarpus* tiling expression array were extracted and compared to the *Ectocarpus* TE library using BLASTN. Significant hits (evalue < E-20 and > 95% identity) were attributed to only one TE family (highest identity).

Expression analysis

The manually annotated library of Ectocarpus TE sequences contains 98 fully annotated and putatively autonomous elements with coding capacity that each represents a TE family. We have queried these 98 sequences against the "non-stressed" Ectocarpus EST dataset (82872 sequences) using BLASTN with filter for low complexity regions. Results were filtered for expectation value <10-25, and identity >95%. Redundancy in the hits due to ESTs matching more than one element and/or more than once the same element was removed by keeping the hit with highest identity to query. We then counted the number of ESTs corresponding to each of the sequences queried. The EST frequency was calculated for each TE family as the number of ESTs found for a TE family divided by the size of the EST dataset (n=82872) (Supplementary Table 1). In order to evaluate these numbers respect to the pool of Ectocarpus mRNA present in the EST dataset, we sought to compare them to the number of ESTs corresponding to highly expressed genes. We therefore analyzed the levels of EST sequences corresponding to genes that are among the top 100 most transcribed genes after transcription array analysis. We selected annotated genes for polyubiquitin (Esi0010_0201), ribosomal protein s4 (Esi0162_0045), ribosomal protein s23 (Esi0274_0035) genes, and several sequences coding for fucoxanthin chlorophyll a/c-binding protein (Fcp) (Esi0559_0003, Esi0982_0001, Esi0092_0045, Esi0458_0016, Esi0203_0022, Esi0492_0004, Esi0458_0015) since several slightly divergent are highly expressed. DNA sequences were compared to the "non-stressed" Ectocarpus EST dataset as described for TEs. We then counted the number of

ESTs corresponding to each of the sequences queried, pooling the results for the Fcp gene family.

5.4 References:

Delaroque N, Boland W: The genome of the brown alga Ectocarpus siliculosus contains a series of viral DNA pieces, suggesting an ancient association with large dsDNA viruses. *BMC Evol Biol* 2008, 8:110.

Edgar RC: MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research* 2004, 32:1792-1797.

Higgins DG, Sharp PM: CLUSTAL: a package for performing multiple sequence alignment on a microcomputer. *Gene* 1988, 73:237-244.

Huson DH, Bryant D: Application of phylogenetic networks in evolutionary studies. *Mol Biol Evol* 2006, 23:254-267.

Kalendar R, Vicient CM, Peleg O, Anamthawat-Jonsson K, Bolshoy A, Schulman AH: Large retrotransposon derivatives: Abundant, conserved but nonautonomous retroelements of barley and related genomes. *Genetics* 2004, 166:1437-1450.

Muller DG, Kapp M, Knippers R: Viruses in marine brown algae. In *Advances in Virus Research, Vol 50. Volume* 50. San Diego: Academic Press Inc; 1998: 49-67: *Advances in Virus Research*].

Peters AF, Marie D, Scornet D, Kloareg B, Cock JM: Proposal of Ectocarpus siliculosus (Ectocarpales, Phaeophyceae) as a model organism for brown algal genetics and genomics. *Journal of Phycology* 2004, 40:1079-1088.

Peters AF, Scornet D, Muller DG, Kloareg B, Cock JM: Inheritance of organelles in artificial hybrids of the isogamous multicellular chromist alga Ectocarpus siliculosus (Phaeophyceae). *European Journal of Phycology* 2004, 39:235-242.

Quesneville H, Bergman CM, Andrieu O, Autard D, Nouaud D, Ashburner M, Anxolabehere D: Combined evidence annotation of transposable elements in genome sequences. *Plos Computational Biology* 2005, 1:166-175.

Quesneville H, Nouaud D, Anxolabehere D: Detection of new transposable element families in Drosophila melanogaster and Anopheles gambiae genomes. *J Mol Evol* 2003, 57 Suppl 1:S50-59.

Witte CP, Le QH, Bureau T, Kumar A: Terminal-repeat retrotransposons in miniature (TRIM) are involved in restructuring plant genomes. *Proceedings of the National Academy of Sciences of the United States of America* 2001, 98:13778-13783.

Chapter VI:

Conclusions and perspectives

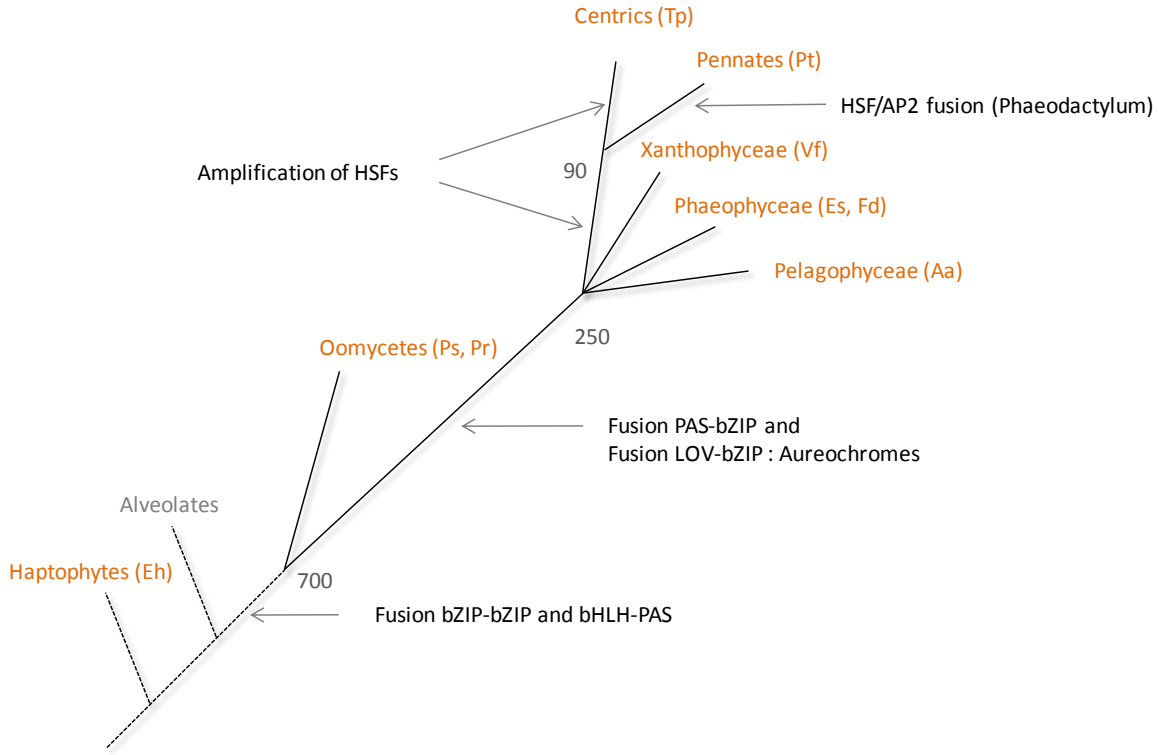
6.1 Conclusions and perspectives

The study of diatoms is a fascinating field trying to unravel the divergences that could help explaining their success in the contemporary oceans. Secondary endosymbiosis for example was essential to the expansion of the chromalveolate group as it has been accompanied by the acquisition of a plastid and by the transfer of a multitude of genes selected from the red-algal endosymbiont. However, phylogenomic analyses of diatom proteomes are still ongoing to trace back the origin of each diatom gene. In addition to the identification of many genes of probable bacterial origin in diatoms (Bowler et al., 2008), Moustafa and collaborators (2009) recently provided evidence for a prasinophyte-like endosymbiont in the common ancestor of chromalveolates as supported by the fact the 70% of diatom genes of Plantae origin are of green lineage provenance and that such genes are also found in the genome of other stramenopiles. Therefore the authors propose that chromalveolates were "green" prior to "red" and are the product of serial secondary endosymbiosis first with a green algae, followed by a second one with a red algae that conserved the genomic footprints of the previous but displaced the green plastid. Hence, diatom and chromalveolate proteomes definitely appear to result of a melting pot of genes of various origins. Beyond the origins, the management and harmonization of their regulation is likely to be crucial in defining diatom's ecological success.

In order to decipher these aspects, I focused much of my PhD trying to identify in diatoms some core regulators and mechanisms known to be involved in the regulation of gene expression. For example, I have shown in Chapter II that the HSF class of transcription factors (TFs) has undergone a significant expansion since the divergence with other stramenopile lineages. The abundance of HSFs in diatoms likely reflects important biological functions in regulating the transcription of genes which *cis*-elements evolved in parallel. When we look at the evolution of the TF complement at a broader scale we notice that,

similar to the acquisition of genes by HGT or EGT, the appearance of new gene fusions is a slow and cumulative (but selective) process. Therefore, the composition of TF complements diverges gradually and consists of both ancestrally shared and lineage-specific peculiarities which *in fine* determine species-specific regulatory capabilities (Figure 6.1). As shown in Chapter 2, the resulting TF complements can further evolve specific responses in a process likely coordinated with the evolution of new DNA binding specificities and the building of regulatory networks.

Figure 6.1: Schematic representation of the time of appearance of new TF features along the evolution of the stramenopile and haptophyte groups (united as Chromista). The groups studied in this work are shown in orange. Numbers indicate approximate divergence times in millions of years. Species abbreviation: Aa (*A. anophegefferens*), Eh (*E. huxleyi*), Es (*E. siliculosus*), Pr (*P. ramorum*), Ps (*P. sojae*), Pt (*P. tricornutum*), Tp (*T. pseudonana*), Vf (*V. frigida*, Fd (*F. distichus*).



Apart from transcriptional regulation that is largely governed by TFs, several other layers of control regulate gene expression. In most eukaryotes, the modification of histones and DNA (cytosine methylation), as well as the RNA interference (RNAi) machinery constitute such mechanisms unified under the term "epigenetic". I have shown in Chapter IV that the *P. tricornutum* genome encodes a collection of proteins known to be responsible for histone modifications that probably reflect the existence of complex epigenetic regulation or "histone code" in this organism. I also verified that such marks could be identified in a *P. tricornutum* protein extract and adapted a chromatin immunoprecipitation protocol for this species. Using this protocol will allow deciphering the distribution of different histone marks along the diatom chromosomes and such studies will be undergoing in the home laboratory by a PhD student.

I also identified several DNA methyltransferases (DNMTases) in diatom genomes including two which are not typically found in most eukaryotes. In addition, I have shown that transposable elements (TEs) are methylated in *P. tricornutum* and that some genes also appear to be methylated, suggesting that they are regulated at the epigenetic level. Analysis of the *P. tricornutum* Window McrBC Restriction (WMR) array is still undergoing and should reveal a precise whole-genome DNA methylation pattern for this species. Bisulfite sequencing of putatively methylated genes will tell what contexts (CpG, CpNpG, CpNpN) are the target of methylation in this species.

The identification of a functional RNAi machinery in *P. tricornutum* now enables to knock down the expression of endogenous genes. For instance, the gene silencing method set up by De Riso and collaborators (2009) could allow generating mutants down-regulated for the different DNMTases found. This, combined with WMR and tiling array, could help understanding their functional specificities. Similarly, I have designed an experiment aiming to help deciphering whether the putative diatom Argonaute (AGO) and Dicer proteins

identified are indeed involved in epigenetic regulation. The procedure is based on the following observations: (i) *Blackbeard* (*BKB*) and *Surcouf* (*SCF*) elements are up-regulated in response to specific stresses; (ii) TEs are methylated in the *P. tricornutum* genome and *BKB* activation is accompanied by hypomethylation; (iii) down-regulation of the *GUS* gene through the RNAi machinery is accompanied by DNA methylation. A postdoc student from the home laboratory has been generating *P. tricornutum* lines transformed with a silencing construct targeting the *Ago* or *Dicer* genes. He will now assess whether *BKB* and *SCF* are constitutively up-regulated and/or hypomethylated in the transgenic lines respect to wild type. For instance, if the down-regulation of the putative Dicer protein causes the activation and/or hypomethylation of TEs, this would bring an indirect proof that it is indeed involved in the production of small RNAs which themselves direct DNA methylation.

The study of the RNAi machinery in *P. tricornutum* is also continuing in the laboratory of Angela Falciatore. Indeed, the team has recently performed the isolation and deep-sequencing of *P. tricornutum* small RNAs derived from cells grown under different conditions (personal communication). This experiment will enable to map the small RNAs from each condition on the genome and help understanding their functions. Thus several genome-wide epigenetic characterizations will soon come for *P. tricornutum* regarding small RNAs, DNA methylation, and histone marks. In order to unify these data into a global picture and help their visualization, I will generate a custom genome browser (Figure 6.2 and Figure 6.3) dedicated to the study of the *P. tricornutum* epigenome in which will also appear tracks displaying TEs mapping and gene models. Indeed, it is obvious today that studying any biological process, researchers must consider how it is regulated at the epigenetic level.

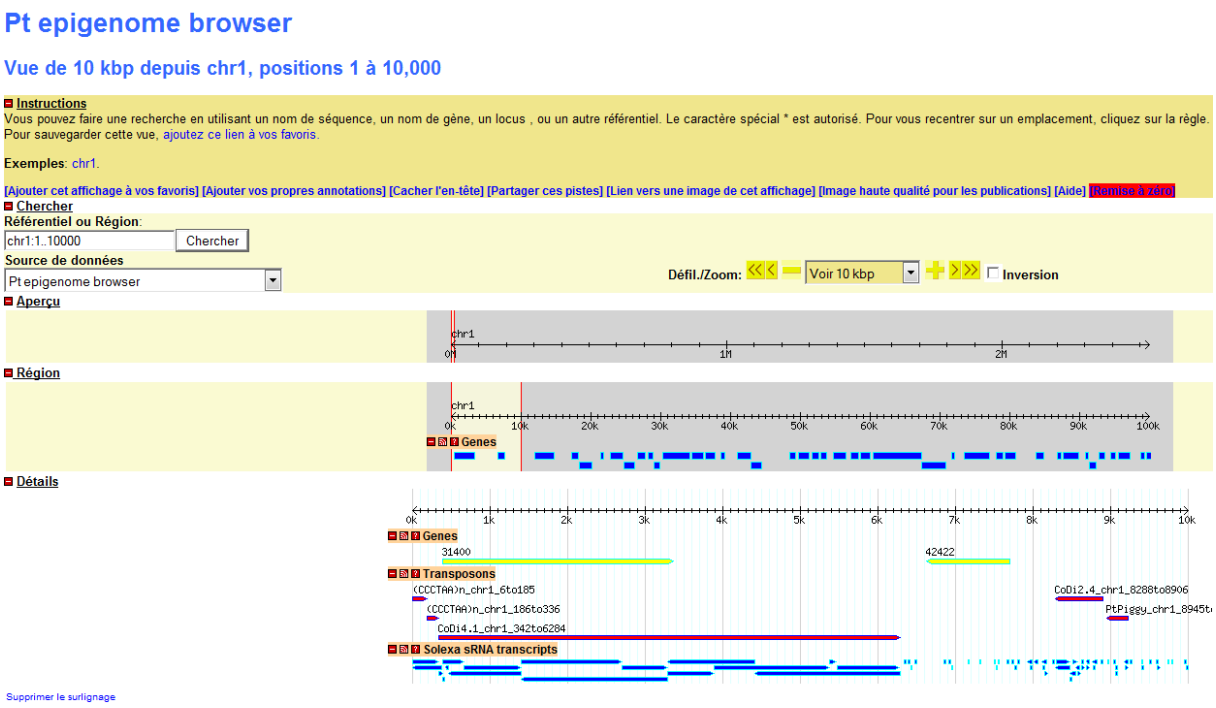
Finally, although I already discussed many points at the end of Chapter III, I would like to emphasize some of them. First, the haplotype specificity of all LTR-RT insertions

assessed is a remarkable feature of the *P. tricornutum* genome. This may reflect the fact that the insertions that occurred after the loss of sexual reproduction in *P. tricornutum* (if it did loose it) remained in a heterozygous state. Considering the abundance of LTR-RTs in *P. tricornutum*, the divergence they generate between haplotypes is likely biologically significant and may still be accentuating with *de no*vo insertions. Therefore, it seems that at most if not all insertion loci, only one allele of the surrounding genes may be impacted. This also questions whether genes originating from recent HGT also remained haplotype-specific.

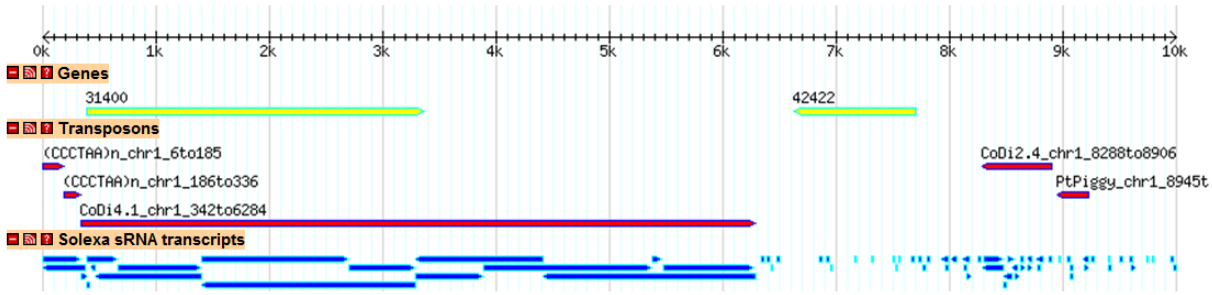
I have identified several new lineages of LTR-RTs from diatoms, brown algae, and also from metagenomic sequences. As we have seen in chapter III, in the case of the environmental samples examined, the bigger the filter, the more frequent are the RT sequences. This is for at least two reasons: first, the smaller filters examined are likely to catch mainly bacteria. Second, in many single-celled eukaryotes, genome size is positively correlated with cell size. For example, *P. tricornutum* and *T. pseudonana* which are about 10 µm in size have small genomes of about 30 Mb. In contrast, genome size of the much larger diatom Coscinodiscus (about half a millimeter in size) has been estimated to be around 50 Gb (Connolly et al., 2008). Since TEs can be major contributors of genome fattening, the TE content of such cells could be similar to what is observed in maize. Therefore, TE sequences are expected to constitute a major fraction of sequences coming from bigger filters. The GOS expedition has collected many samples corresponding to such filters and only a small fraction had been sequenced at the time I did my analysis. Therefore, sequencing efforts on such samples will be of high interest to continue analyzing the abundance and diversity of TEs from microplankters. Similarly, some middle-sized diatom genomes are now being sequenced and it will be of interest to measure how much of their extra size can be attributed to TEs.

Figure 6.2: The Pt epigenome browser. (A) Screenshot overview of the Pt epigenome browser. (B) Screenshot of track details for positions 1 to 10.000 of chromosome 1 displaying genes, TEs, and sRNA transcripts.

A



B



References:

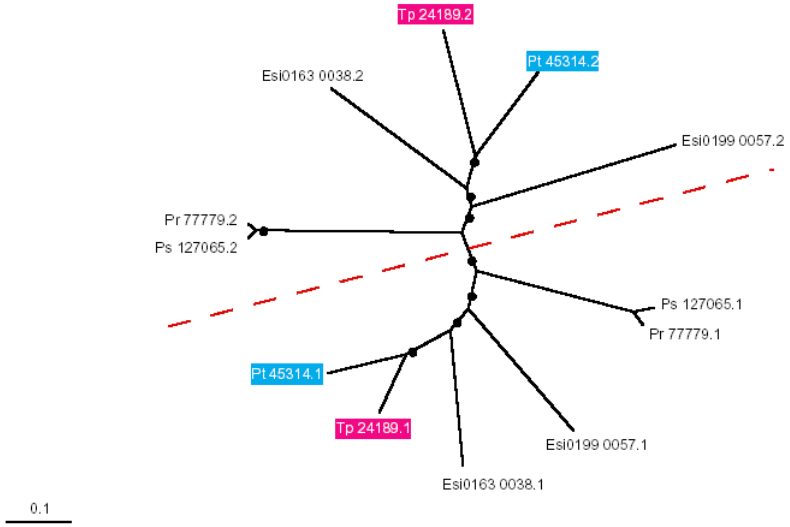
1. Bowler C, Allen AE, Badger JH, Grimwood J, Jabbari K, Kuo A, Maheswari U, Martens C, **Maumus F**, Otillar RP, et al: The Phaeodactylum genome reveals the evolutionary history of diatom genomes. Nature 2008, 456:239-244.

2. Connolly JA, Oliver MJ, Beaulieu JM, Knight CA, Tomanek L, Moline MA: Correlated evolution of genome size and cell volume in diatoms (Bacillariophyceae). Journal of Phycology 2008, 44:124-131.

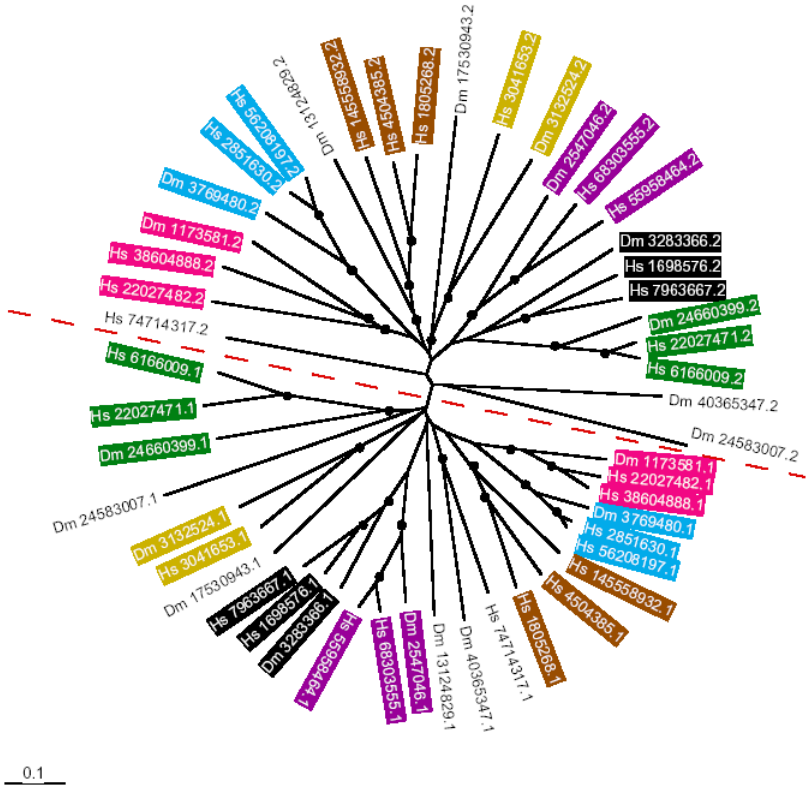
3. De Riso V, Raniello R, **Maumus F**, Rogato A, Bowler C, Falciatore A: Gene silencing in the marine diatom Phaeodactylum tricornutum. Nucleic Acids Res 2009.

4. Moustafa A, Beszteri B, Maier UG, Bowler C, Valentin K, Bhattacharya D: Genomic footprints of a cryptic plastid endosymbiosis in diatoms. Science 2009, 324:1724-1726.

Supplementary Figure 1.1



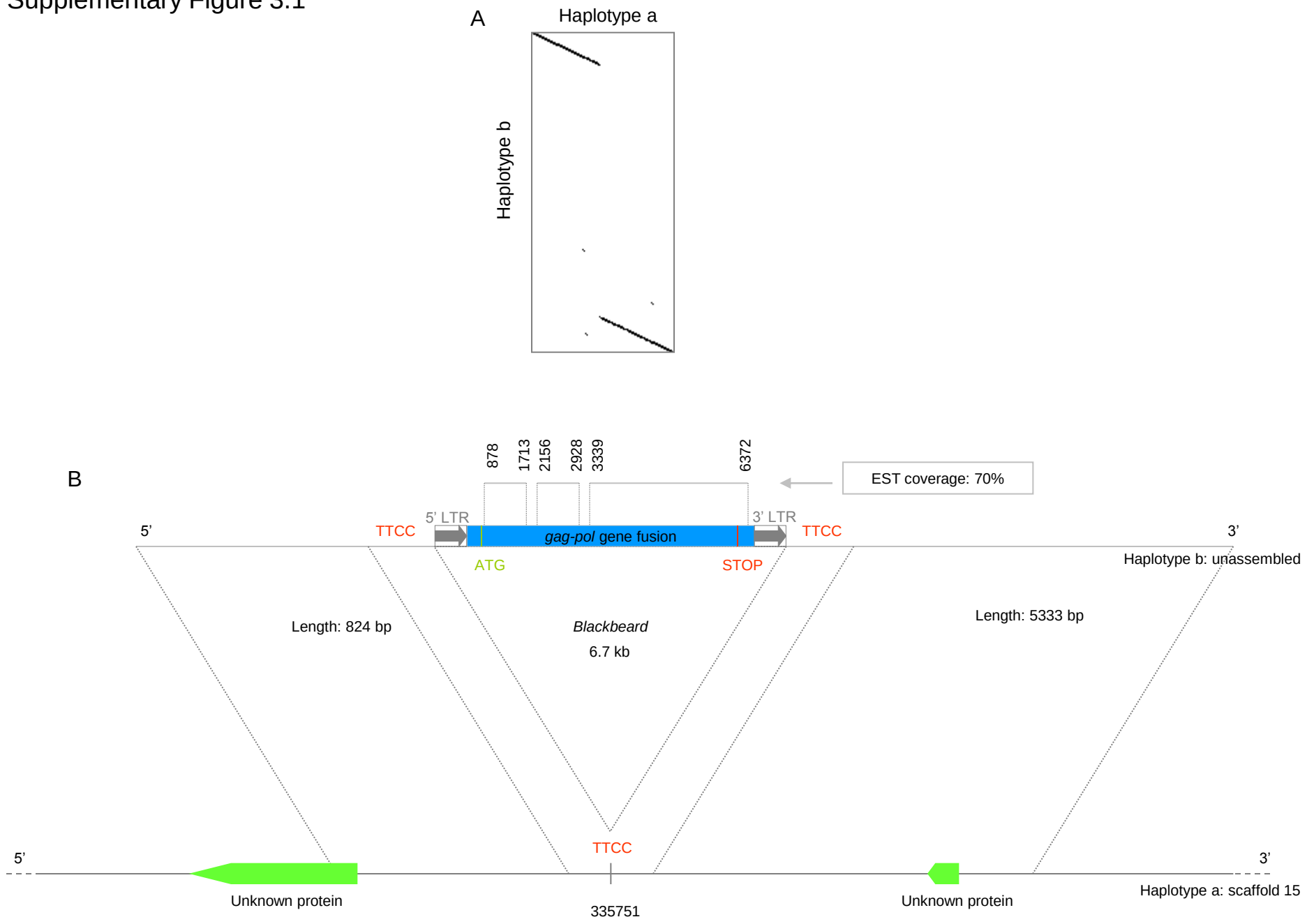
a) Stramenopile proteins containing two bZIP domains



b) PAS domains in human and Drosophila proteins containing bHLH and PAS

Supplementary Fig. 4. Two bZIP and two PAS domains in selected proteins

Supplementary Figure 3.1



Supplementary Figure 3.2

Supplementary Table 3.1

Classification and structural features of putatively active LTR-RTs in diatom genomes

Group	Family	Name	Size (bp)	ORF1 size (aa)	ORF2 size (aa) *	Frame GAG/POL	LTR size (bp)	LTR id (%)	Terminal inverted repeats	Scaffold**	Target site duplication	YVDD motif ***	GenBank accession number
CoDi1	CoDi 1.1	PtC5	6560		1902	in frame	468	100	TG…CA	1	ACCCT…ACCCT	EYVCVYVDDIACVSRNP-KAFFD	EU432476
		PtC20	6559		1902	in frame	468	100	TG…CA	9	TGGCC…TGGCC	EYVCVYVDDIACVSRNP-KAFFD	
		PtC42	6559		1902	in frame	468	100	TG…CA	18	GGCCA…GGCCA	EYVCVYVDDIACVSRNP-KAFFD	
CoDi2	CoDi 2.1	PtC10	7643		1929	in frame	638	100	TGT…ACA	3	AAATG…AAATG	EYIAVYVDDIAVAAHDP-GSIVE	EU432477
	CoDi 2.2	PtC30	7689		1922	in frame	721	100	TGTTACGG…CCGTAACA	13	GTCAC…GTCAC	EYIAVWVDDIAVAAHSP-GDIID	EU432478
		PtC49	7687		1922	in frame	720	100	TGTTACGG…CCGTAACA	19	CTTCT…CTTCT	EYIAVWVDDIAVAAHSP-GDIID	
	CoDi 2.3	PtC25	7394		1912	in frame	616	99	TGT…ACA	11	not found	EYIAVYVDDIAVAAHDP-GAIIH	EU432479
		PtC75	7394		1912	in frame	616	99.7	TGT…ACA	31	not found	EYIAVYVDDIAVAAHDP-GAIIH	
	CoDi 2.4	PtC7	7396		1914	in frame	614	99.8	TGT…ACA	1	CATAA…CATAA	EYIAVYVDDIAVAAQDP-EGIIR	EU432480
		PtC21	7394		1913	in frame	614	100	TGT…ACA	9	AATAC…AATAC	EYIAVYVDDIAVAAQDP-EGIIR	
		PtC73	7395		1914	in frame	614	99.7	TGT…ACA	31	GCAGG…GCAGG	EYIAVYVDDIAVAAQDP-EGIIR	
CoDi3	CoDi 3.1	PtC8	7450		1979	in frame	696	100	TGATAA…TTATCA	2	GTCG…GTCG	EYIISYVDDLTVVSADT-KAILK	EU432481
		PtC33	7450		1979	in frame	696	100	TGATAA…TTATCA	14	CGGA…CGGA	EYIISYVDDLTVVSADT-KAILK	
		PtC53	7450		1976	in frame	696	100	TGATAA…TTATCA	22	CCAG…CCAG	EYIISYVDDLTVVSADT-KAILK	
	CoDi 3.2	PtC43 (SCF)	7680		1939	in frame	844	100	TG…TA?	18	AAAG…AAAG	EYIISYVDDLTVVSENT-RAILE	EU363805
		PtC79	7670		1939	in frame	833	97.2	TG…TA?	BD.8.17	not found	EYIISYVDDLTVVSENT-RAILE	
CoDi4	CoDi 4.1	PtC18	7254	561	1420	stop codon	300	99.8	TGATT…AATCA	7	TTTG…TTTG	EYVLVYSDDILVVSHDP-KRTMD	EU432482
		PtC56	7250	561	1420	stop codon	299	100	TGATT…AATCA	24	TCGC…TCGC	EYVLVYSDDILVVSHDP-KRTMD	
	CoDi 4.2	PtC63	6667		1828	in frame	163	100	TAGAAAA…TTTTCTA	27	GTGT…GTGT	EYVLIYTDDFLCISSDP-KSILD	EU363804
		PtC22	6667		1866	in frame	163	100	TAGAAAA…TTTTCTA	10	TATT…TATT	EYVLIYTDDFLCISSDP-KSILD	
		PtC70	6668		1828	in frame	163	100	TAGAAAA…TTTTCTA	30	not found	EYVLIYTDDFLCISSDP-KSILD	
		BKB	6669		1853	in frame	163	100	TAGAAAA…TTTTCTA	15	TTCC…TTCC	EYVLIYTDDFLCISSDP-RNILD	
	CoDi 4.3	PtC66	6720	502	1349	stop codon	162	100	TAGAAAA…TTTTCTA	29	ATAT…ATAT	EYVLIYTDDFLCISHDP-KSILD	EU432483
		PtC82	6720	502	1349	stop codon	162	100	TAGAAAA…TTTTCTA	BD.31.35	CAAC…CAAC	EYVLIYTDDFLCISHDP-KSILD	
	CoDi 4.4	PtC50	6753	479	1385	stop codon	153	94	TG…CA	20	TGAC…TGAC	EYVLIYVDDILAVSEKP-DRIME	EU432484
	CoDi 4.5	TpC12	6715	683	1402	stop codon	235	99.6	TGATAA…TTATCA	7	ATAA…ATAA	SYILCYVDDILVIHHQP-EDVIK	EU432485
		TpC17	6715	683	1402	stop codon	235	100	TG…CA	12	GAGA…GAGA	SYILCYVDDILVIHHQP-EDVIK	
CoDi5	CoDi 5.1	PtC13	5274	400	1291	stop codon	203	100	TG…CA	6	A(c)AAAAAC…A(t)AAAAAC	IQFTIVVDDFGVKYTGT-EHAQH	EU432486
		PtC17	5274	400	1291	stop codon	203	100	TG…CA	7	CCACC…(ata)CCACC	IQFTIVVDDFGVKYTGN-EHAQH	
		PtC31	5274	394	1291	stop codon	204	99.5	TG…CA	14	CATTT…(a)CATTT	IQFTIVVDDFGVKYTGT-EHAQH	
		PtC32	5274	400	1291	stop codon	203	100	TG…CA	14	GTCCT(tat)…GTCCT	IQFTIVVDDFGVKYTGT-EHAQH	
		PtC34	5241	400	1280	stop codon	203	100	TG…CA	15	AAATA…(ata)AAATA	IQFTIVVDDFGVKYTGT-EHAQH	
		PtC35	5262	400	1287	stop codon	203	99.5	TG…CA	15	ACAAC(tat)…ACAAC	IQFTIVVDDFGVKYTGT-EHAQH	
		PtC39	5274	400	1291	stop codon	203	97	TG…CA	17	GTCAC(tat)…GTCAC	IQFTIVVDDFGVKYTGT-EHAQH	
		PtC45	5274	400	1291	stop codon	203	99.5	TG…CA	19	not found	IQFTIVVDDFGVKYTGN-EHAQH	
		PtC51	5274	400	1291	stop codon	203	100	TG…CA	20	GTTACT(at)…GTTACT	IQFTIVVDDFGVKYTGN-EHAQH	
		PtC67	5264	394	1230	stop codon	202	100	TG…CA	29	not found	IQFTIVVDDFGVKYTGT-EHAQH	
	CoDi 5.2	PtC59	5670	140	1372	stop codon	308	99.7	TG…CA	24	CT(t)AT…CTAT	ISFSIVVDDFGVKYVGK-EHAEH	EU432487
	CoDi 5.3	PtC60	5553	349	1263	stop codon	239	97.5	TG…CA	26	not found	ISFSIVVDDFGVKYVGR-EHAEH	EU432488
	CoDi 5.4	TpC16	5376	413	1227	stop codon	164	100	TG…CA	10	TCATC…TCATC	ITFSIVVDDFGIKTVGL-SHAKH	EU432489
		TpC24	5396	412	1235	stop codon	164	96.4	TG…CA	22	not found	ITFSIVVDDFGIKTVGL-SHAKH	
	CoDi 5.5	TpC18	5758	397	1388	frameshift?	246	100	TGCT…AGCA	12	ACAAC…ACAAC	VCFTIVVDDFCVKYIGK-QHAEH	EU432490
	CoDi 5.6	TpC25	6174	548	1401	stop codon	203	93.6	TG…CA	23	not found	IMFTIVVDDFGVKYVGK-EHADH	EU432491
CoDi6	CoDi 6.1	TpC4	6957	535	1440	stop codon	431	98.8	TGTG…CACA	3	TACAT…TACAT	VVFIVYVDD-GV-FLGR-CDRQL	EU432492
		TpC19	6941	535	1442	stop codon	420	100	TGTG…CACA	14	GTGCA…GTGCA	VVFIVYVDD-GV-FLGR-CDRQL	
	CoDi 6.2	TpC7	5241		1511	in frame	242	98.8	TGTT…AACA	6	GTTGCC…GTTGCC	CFVGIYVDDNLIIG-HP-ELVDD	EU432493
		TpC15	5241		1511	in frame	242	100	TGTT…AACA	9	AATTGA…AATTGA	CFVGIYVDDNLIIG-HP-ELVDD	
	CoDi 6.3	TpC22	5182		1496	in frame	161	98.1	TG…CA	22	not found	VFIAIYVDDNLIVG-HP-KAIEC	EU432494
	CoDi 6.4	PtC3	5237		1447	in frame	413	100	TGT…ACA	1	GAACGG…GAACGG	VLMMVHVDDCKITG-SR-EKVMW	EU432495
	CoDi 6.5	PtC23	5349		1566	in frame	348	99.3	TGT…ACA	10	GACG…GACG	LVAMTHVDDTIIFG-PK-QEIEW	EU432496
	CoDi 6.6	PtC26	8062	535	1720	frameshift?	279	99.6	TGT…ACA	12	CACCT…CACCT	CIFLFWVDDAIICSHDD-STIVS	EU432497
	CoDi 6.7	PtC46	6644	468	1411	stop codon	490	99.7	TGT…ACA	19	AATGT…AATGT	SVYVMYTDDSIIAGPDL-EELEQ	EU432498
CoDi7	CoDi7.1	PtC47	5954		1796	in frame	257	100	TG…CA	19	not found	EYVCTHVDDFMIVSKTP-EKIME	EU432499
GyDi1	GyDi1.1	TpG1	5260	407	1193	stop codon	197	100	TG…CA	1	GTCCT…GTCCT	-EAEVYLDDVGAFSN-TWAQHI-	EU432500
GyDi2	GyDi2.1	TpG9	6759	332	1239	stop codon	644	99.9	TGTC…GACA	6	CTGAA…CTGAA	-YVRTYLDDLLILSKGTFDDHL-	EU432501

*: ORF2 refers to the major ORF found between LTR pairs which can constitute either the POL ORF or a GAG-POL fusion [69].

**: BD (bottom drawer) refers to elements found in the non assembled set of contigs from the *P. tricornutum* genome.

***: This column presents the amino acid sequence of the YVDD box found in the fifth conserved region of the RT domain as previously defined [11].

Supplementary Table 3.2

Distribution across *P. tricornutum* accessions and coordinates in Pt1 genome of several bands extracted from the SSAP experiment
S: band was exctracted from this accession; Y: band of identical size detected

Accession#	1	2	3	9	5	10	6	7	8	11	12	13	4	Scaffold	Position
BKB															
7					Y	S								9	496675
8							S	Y	Y					24	81317
9							S	Y	Y					5	789278
12					S	Y								5	213378
14													S	in CoDi4.3	
15													S	14	651334
17													S	18	611025
18													S	in CoDi6.7	
19							S	Y	Y					11	12197
21							S	Y	Y					18	615410
22										Y	Y	S		28	310243
23										Y	Y	S		16	462113
24												S		26	94227
25												S		in CoDi5.1	
26										Y	Y	S		in tandem repeat	
SCF															
29													S	24	160246
32													S	3	264470
33							S	Y	Y				Y	in CoDi2.1	
36							S	Y	Y					in CoDi5.1	
39					Y	S								in CoDi2.2	
41					Y	S								in PiggyBac	
42					Y	S								5	533524
46										Y	S	Y		in CoDi1.1	
PtC34															
51													S	2	235182
53													S	14	26911
55					S	Y								2	1006068
57							S	Y	Y					4	1036780
59							S	Y	Y					12	2791
60							S	Y	Y					5	504951
64					Y	S								29	1557