Charaterization of the Phaeodactylum tricornutum epigenome

Xin Lin

UNIVERSITE PARIS-SUD

Ecole doctorale Sciences du végétal
IBENS UMR 8197

DISCIPLINE Biogenetic

THÈSE DE DOCTORAT

soutenue le 18/10/2012

par

Xin LIN

Charaterization of the *Phaeodactylum tricornutum* epigenome

Directeur de thèse :	Chris Bowler	Directeur de Recherche (IBENS-Paris UMR8197)
Co-directeur de thèse :	Leila Tirichine	Ingénieur de recherche (IBENS-Paris UMR8197)

Composition du jury :

Président du jury :	Zhou Dao-xiu	Professeur (Université de Paris-Sud 11 UMR8618)
Rapporteurs :	Thomas Mock	Principal investigator (University of East Anglia)
	Philippe Gallusci	Maitre de conférences (INRA Bordeaux-Aquitaine UMR 1332)
Examinateurs :	Allison Mallory	Chargé de recherche (Institut Jean-Pierre Bourgin, UMR1318)

Acknowledgements

I am very lucky to have this opportunity to study in Paris. I received help and support from the different people I met during this period. I had good times and hard times, but never bad times during my PhD study. The wonderful memories during these years will be my precious treasure for my whole life.

First of all, I want to express my deepest gratitude to my supervisors Chris Bowler and Leila Tirichine. I am very thankful to Chris Bowler who gave me the opportunity to come to Paris and work in excellent conditions. Chris is busy but he always tries his best to find the time to listen to me and advise me. The freedom and the “pressure” he gave allowed me to learn how to think independently and acquire scientific knowledge. The trust he showed encouraged me to conquer difficulties. Without his criticisms and encouragement I cannot finish my PhD. I will remember what he said to me: “no pains no gains”. I have no words to express my gratitude to Leila. Her enthusiasm to science and her “never give up” spirit inspired me every day in the lab. She helped me solve my day-to-day problems and encouraged me when I was frustrated. We shared happiness and sorrow. We always tried to find solutions together when we had difficulties. I thank her for her patience as well as for the time she spent teaching me and guiding me.

I want to thank many members from Chris’ lab. Firstly I would like to thank the members from diatom group: Alaguraj Veluchamy, Judit Prihoda, Atsuko Tanaka, Joe Morrisse, Shruti Mishra, Yann Thomas, Hanhua Hu, Florian Maumus, Agnès Meichenin and other members who have left the lab. In particular, I thank Alaguraj who did most of bioinformatic analyses for my thesis. Without his hard work I cannot finish my PhD. During this work I got help from many colleagues: Frédy Barneche, François Roudier, Stéphanie Drevense, Clara Bourbousse, Anne-Sophie Fiorucci, Alexis Sarazin, Gérald Zabulon et al. I sincerely thank who helped me for their advices, scientific discussions and the support they gave me.

During the past four years, though I am far away from my home country, I did not feel lonely in Paris because my friends. In this respect, I also want to thank my friends in Paris for their supports: Hongyi Li, Cong Zhou, Fei Teng, Jie Pan, Fanglei Wang, Zheng Yang et al.

Lastly but most importantly, I would like to thank my parents and grandparents for their unconditional love and supports. And also thanks to the most important finding I had during

this thesis in Paris -my husband Xiaoliang Fan, for all his love and comprehension. I want to delicate this work to them.

Table of Contents

Chapter I **1**

Introduction **1**

Table of contents **3**

1.1 Epigenetics and epigenomes **4**

1.2 DNA methylation **7**

1.2.1 DNA methylation detection 7

1.2.2 DNA methylation patterns 10

1.2.3 The DNA methylation machinery 10

1.2.4 Functions of DNA methylation 11

1.2.5 Evolution of DNA methylation 12

1. 3. Histone modifications **15**

1.3.1 Histone modification detection 15

1.3.2 The functions of histone modification 16

1.3.3 The Histone modification machinery 19

1.4 RNA interference **19**

1.5. Interactions between epigenetic components **21**

1.6. Epigenetics and the environment **23**

1.7 Diatoms and epigenetics **25**

1.7.1 DNA methylation and its machinery in diatoms 29

1.7.2 Histone modification in diatoms 35

1.7.3 RNA interference in diatoms 38

1.8 PhD thesis description **40**

1.9 References **42**

Chapter II **53**

Insights into the Role of Methylation in Diatoms by Genome-Wide Profiling in *Phaeodactylum tricornutum* **53**

2.1 Abstract **57**

2.2 Introduction **58**

2.3 Results **59**

2.3.1 Whole genome methylation landscape 59

2.3.2 HMRs in TEs and other repeat loci 60

2.3.3 Gene methylation profiles 65

2.3.4 Genomic distribution of methylated genes 66

2.3.5 Methylation, gene expression, and gene product function 70

2.3.6 Methylation and non-autonomous Class II TEs 73

2.4 Discussion **74**

2.5 Materials and Methods **77**

2.5.1 Culture conditions 77

2.5.2 DNA preparation, microarray hybridization, and validation 78

2.5.3 RNA-seq preparation 78

2.5.4 Identification of methylated regions, distribution and expression analysis 79

2.7 References **80**

2.8 Supporting information **84**

2.8.1 Supporting Supplementry Figures 84

2.8.2 Detailed Methods 99

2.8.3 Supplementary References 102

Chapter III: **105**

Genome wide analysis of the histone marks H3K4me2, H3K9me2 and H3K27me3 in the model diatom *Phaeodactylum tricornutum* **105**

3.1 Abstract **108**

3.2 Introduction **108**

3.3 Results **110**

3.3.1 Identification of genomic regions associated with H3K4me2, H3K9me2 and H3K27me3 110

3.3.2 Genome wide distribution of H3K4me2, H3K9me2 and H3K27me3 on genes.... 117

3.3.3 Genome wide distribution of H3K4me2, H3K9me2 and H3K27me3 on TEs 123

3.3.4 Combinatorial marking of H3K4me2, H3K9me2 and H3K27me3 on the same genomic regions 125

3.3.5 Combinatorial effect of histone modifications and DNA methylation on gene expression 127

3.4 Discussion 132

3.4.1 H3K4me2 mainly marks genes in *P. tricornutum* 132

3.4.2 H3K9me2 mainly marks TEs in *P. tricornutum* 133

3.4.3 The distribution of H3K27me3 in *P. tricornutum* is unorthodox 134

3.4.4 Methylated TEs and heavily methylated genes tend to be co-marked by H3K27me3 and H3K9me2 139

3.5 Perspectives 141

3.6 Material and Methods 142

3.6.1 Growth conditions 142

3.6.2 Peptide competition assay for antibody specificity test 142

3.6.3 Chromatin immunoprecipitation and sequencing 143

3.6.4 RNA sequencing 143

3.6.5 Computational analysis of histone modifications *P. tricornutum* by ChIP-sequencing 144

3.7 References 145

3.8 Supplementary information 157

3.8.1 Supplementary Figures 157

3.8.1 Supplementary tables 163

3.9 Annex-Chromatin immunoprecipitation protocol 166

Chromatin immunoprecipitation coupled to detection by quantitative real time PCR to study in vivo protein DNA interactions in two model diatoms *Phaeodactylum tricornutum* and *Thalasiossira pseudonana* 166

Chapter IV 187

Putative epigenetic components and reverse genetic approach for dissecting epigenetic machineries in *Phaeodactylum tricornutum* **187**

4.1 Introduction **191**

4.1.1 C5-DNA methylation and its machinery 191

4.1.2 N6- adenine methylation 197

4.1.3 C5-DNA demethylation machinery 198

4.1.4 SET domain containing proteins for histone lysine methylation 199

4.1.4.1 Polycomb group proteins 199

4.2 Results **204**

4.2.1 Putative C5-Methyl transferases in diatoms 204

4.2.2 Genome wide DNA methylation quantification of wild type and C5-MTases mutants 206

4.2.3 DNA demethylation machinery in diatoms 212

4.2.4 Identification of proteins containing SET domains for histone methylation in *P. tricornutum* 215

4.2.5 Distribution of PRC2 core subunits in diatoms and other eukaryotic algae 217

4.2.6 Phylogentic analyses of PRC2 component E(z) 219

4.2.7 Phylogentic analyses of PRC2 components ESC 219

4.2.8 Phylogentic analyses of PRC2 components Suz(12) 220

4.2.9 Distribution of PRC1 core subunits in diatoms and other eukaryotic algae 223

4.2.10 Knockdown of E(z) homolog 32817 in*P. tricornutum* 226

4.2.11 Effect of EZH down regulation on global post-translational histones modifications 226

4.2.12 Light microscopy observations of Ez antisencse strains 228

4.3 Discussion **230**

4.3.1 C5-MTase knockdown in *P. tricornutum* 230

4.3.2 Distribution of PcG components in algae and E(z) knockdown in *P. tricornutum* 230

4.4 Materials and methods **233**

4.4.1 Cell culture 233

4.4.2 Antisense vector construction 233
4.4.3 Genetic transformation of *P. tricornutum* 234
4.4.4 DNA methylation quantification 234
4.4.5 Analysis of DNA methylation by McrBC-qPCR 235
4.4.6 Western blotting 238
4.4.7 Phylogenetic analyse 238
4.5 References 239
Chapter V 249
Discussion and perspectives 249

Chapter I

Introduction

Table of contents

Chapter I **1**

Introduction **1**

Table of contents **3**

1.1 Epigenetics and epigenomes **4**

1.2 DNA methylation **7**

1.2.1 DNA methylation detection 7

1.2.2 DNA methylation patterns 10

1.2.3 The DNA methylation machinery 10

1.2.4 Functions of DNA methylation 11

1.2.5 Evolution of DNA methylation 12

1. 3. Histone modifications **15**

1.3.1 Histone modification detection 15

1.3.2 The functions of histone modification 16

1.3.3 The Histone modification machinery 19

1.4 RNA interference **19**

1.5. Interactions between epigenetic components **21**

1.6. Epigenetics and the environment **23**

1.7 Diatoms and epigenetics **25**

1.7.1 DNA methylation and its machinery in diatoms 29

1.7.2 Histone modification in diatoms 35

1.7.3 RNA interference in diatoms 38

1.8 PhD thesis description **40**

1.9 References **42**

1.1 Epigenetics and epigenomes

Deoxyribonucleic acid (DNA) is a nucleic acid that contains the genetic information used in the functioning of biological processes inliving organisms and sequencing of entire genomes has revealed a tremendous amount of information about genes promoters, transposable elements (TEs）. In eukaryotic organisms, DNA and its associated proteins form a condensed structure called chromatin which is an instructive DNA scaffold in the nucleus. The chromatin is constituted by the basic unit of DNA packaging termed the nucleosome. The nucleosome core particle contains approximately 147 base pairs of DNA wrapped in around 2 copies each of the core histones H2A, H2B, H3 and H4 like thread wrapped around a spool (Richmond, Sargent, Richmond, Luger, & Ma, 1997). Up to 80 bp linker DNA with H1 and H5 linker histones connect the core particles.

Chromatin has many roles, e.g., in preventing DNA damage or orchestrating, DNA replication and gene regulation. The properties of chromatin vary between different regions of the genome, phases of life cycle, and cell types in the same organism. The way DNA is packaged determines its function. A tightly packed form of DNA is called heterochromatin which includes constitutive and facultative heterochromatin. Constitutive heterochromatin, usually repetitive and permanently transcriptionally silenced, form centromeres and telomeres whereas facultative heterochromatin is not repetitive though it also has a compact structure like constitutive heterochromatin and it can lose its condensed structure and become transcriptionally active under certain conditions. A lightly packed form of DNA is called euchromatin within which genes are transcriptionally active (Ris & Kubai, 1970).

Chromatin states define gene expression states at a whole genome level. So that genomes are programmed to express an appropriate set of genes at a specific time point in response to certain environmental stimuli. The chromatin control of genomes is not at the DNA sequence level: the cells carry the same DNA sequence whereas different chromatin structures and properties affectthe read out of the underlying DNA sequence. Moreover, the chromatin mediated control of genomes can be transient or stably transmitted through cell divisions or even generations.

Chromatin states are modulable because by chromatin modifications the core histone tails are subject to different post-translational modifications. So far more than 60 histone modifications have been identified which make a great contribution to chromatin modifications (Bannister & Kouzarides, 2011; Kouzarides, 2007). DNA sequence can also be subject to DNA methylation,which is involved in enzymatic transfer of a methyl group on the fifth position of the cytosine pyrimidine ring or the number 6 nitrogen of the adenine purine ring. DNA methylation on cytosines and adenines is widely distributed in both prokaryotes and eukaryotes. Modification mainly or exclusively on cytosines (5mC) can be inherited through cell division.

Epigenetics refers to the study of the change of heritable chromatin states without changes of underlying DNA sequence. The chromatin state is controlled by two major epigenetic elements: DNA methylation and histone modifications. The sum information of the genome-wide patterns of DNA and chromatin modifications is termed the epigenome (**Figure 1.1**). These two major epigenomic elements have profound effects on chromatin state, which regulates processes such as gene expression.

Unlike the genome, the epigenome is variable between cells, different developmental stages and in response to different stimuli. Recently, some studies suggest that not only the two classical epigenetic elements mentioned above are involved in genome organization and development, but other aspects of epigenetic regulation, such as small RNA has also been shown to play a central role in gene regulation and become fasting growing epigenetic research field (Costa, 2008; Kerppola, 2009). Furthermore, in 2011 it was discovered that mRNA can be subject to N6-methyladenosine methylation and it is proposed that it has a critical role in human energy homeostasis (Jia et al., 2011). This has opened up the field of RNA epigenetics.

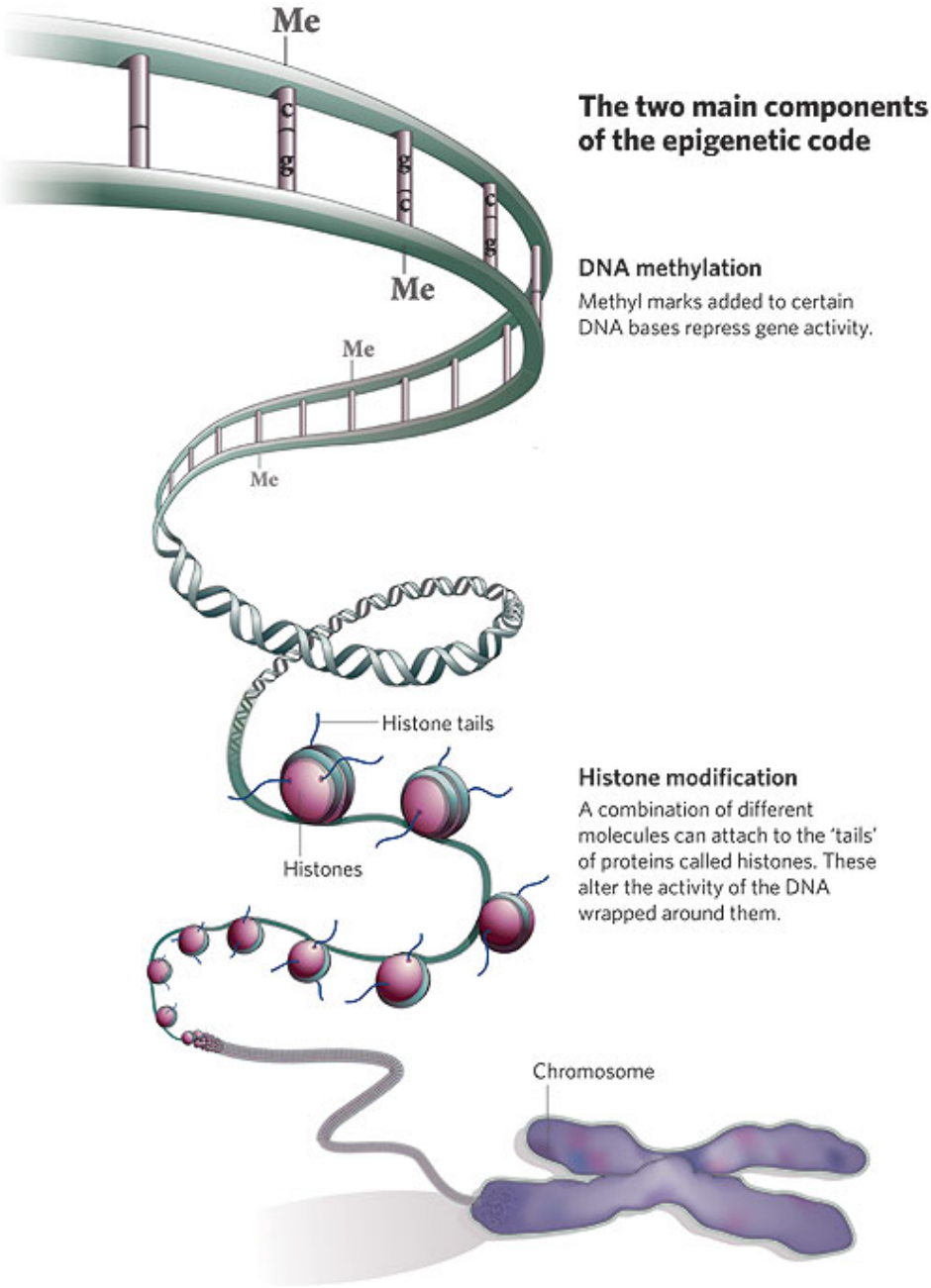


Figure 1.1 The two main components of epigenetics.

1.2 DNA methylation

Recent progress in genomic and biochemical studies in different organisms has elucidated the provenance of DNA cytosine methylation. DNA methylation appears to have emerged first in bacterial restriction–modification (R–M) systems. DNA methylation is so far the best characterized epigenetic mark. It is a biochemical process in which a methyl group is added to the cytosine pyrimidine ring at position 5 (5meC) or a nitrogen of the adenine purine ring at position 6 (N6mA). The former is common to all three super kingdoms while the latter has been characterized in prokaryotes and certain eukaryotic lineages (Aravind, Abhiman, & Iyer, 2011). Another related methylation is on the NH2 group attached to the 4^{th} position of cytosine residues (N4mC), which is found only in prokaryotes. In this introduction, I mainly focus on 5mC methylation and its further modification. DNA methylation is widespread among protists, plants, fungi and animals (Feng & Jacobsen, 2011; Zemach, McDaniel, Silva, & Zilberman, 2010a). It is however absent or poor in some species such as the budding yeast *Saccharomyces cerevisiae*, the fruit fly *Drosophila melanogaster*, the nematode worm *Caenorhabditis elegans* and the brown alga *Ectocarpus siliculosus* (Cock et al., 2010; Goll & Bestor, 2005; Tweedie et al., 1999).

1.2.1 DNA methylation detection

In recent years, a variety of methods for detection of genome-wide cytosine methylation have been developed. Widely used methods include restriction enzyme digestion (such as McrBC digestion) of methylated DNA followed by high density oligonucleotide hybridization microarray or sequencing, and immunoprecipitation of methylated DNA followed by microarray hybridization or sequencing. But these approaches have their drawbacks such as the restriction enzyme bias, difficulties in detection of enriched regions and, more importantly, their limited resolution.

Bisulfite sequencing (sodium bisulfite conversion followed by DNA sequencing) has now become the gold standard method for methylation detection. Sodium bisulfite can convert unmethylated cytosine to uracil while methylated cytosine bases remain unchanged. This method can provide single base resolution methylation profiles. Although bisulfite sequencing is considered the gold standard for DNA methylation studies, it cannot distinguish between 5meC and 5hmC which has been recognized as another important DNA modification recently

(Ficz et al., 2011; Pastor et al., 2011; Williams, Christensen, & Helin, 2011). This means that previous studies of methylation on genomes that contain different cytosine modifications have probably overestimated m5C methylation. The presence of 5hmC was recently reported in brain, neuron and embryonic stem (ES) cells (Ficz et al., 2011). The function of 5hmc is not yet clear. Some studies suggest its involvement in cell differentiation (Ficz et al., 2011). This epigenetic mark which has been ignored for a long time is showing its importance nowadays.

The pioneering new generation of sequencing approaches including single molecule real time (SMRT) sequencing technology holds the promise for future DNA methylation studies. SMRT detects single molecules of DNA being sequenced without the need of cloning or PCR amplification, both of which are processes that can introduce biases. More importantly, SMRT can discriminate 5meC and 5hmC. Selecting this approach has therefore become crucial for DNA methylation studies that focus on the dynamics of 5meC and 5hmC. SMRT can combine advantages of "specificity" and "single base resolution" for DNA methylome studies. Although currently the accuracy of the data generated by SMRT is not satisfactory, it is likely to replace bisulfite sequencing in the foreseeable future.

These approaches have provided a comprehensive picture of conserved and divergent features of DNA methylation patterns in a diverse range of species representing the three kingdoms of life (Feng, Cokus, Zhang, Chen, Bostick, Goll, Hetzel, Jain, Strauss, Halpern, Ukomadu, Sadler, Pradhan, Pellegrini and Jacobsen 2010, Yan *et al.* 2010, Zemach, McDaniel, Silva and Zilberman 2010). Nonetheless, these studies focus mainly on two eukaryotic groups, Unikont and Archaeplastida (**Figure 1.2**). Other organisms such as representatives of marine phytoplankton are waiting for similar analysis which will provide further insights into such a complex regulation mechanism.

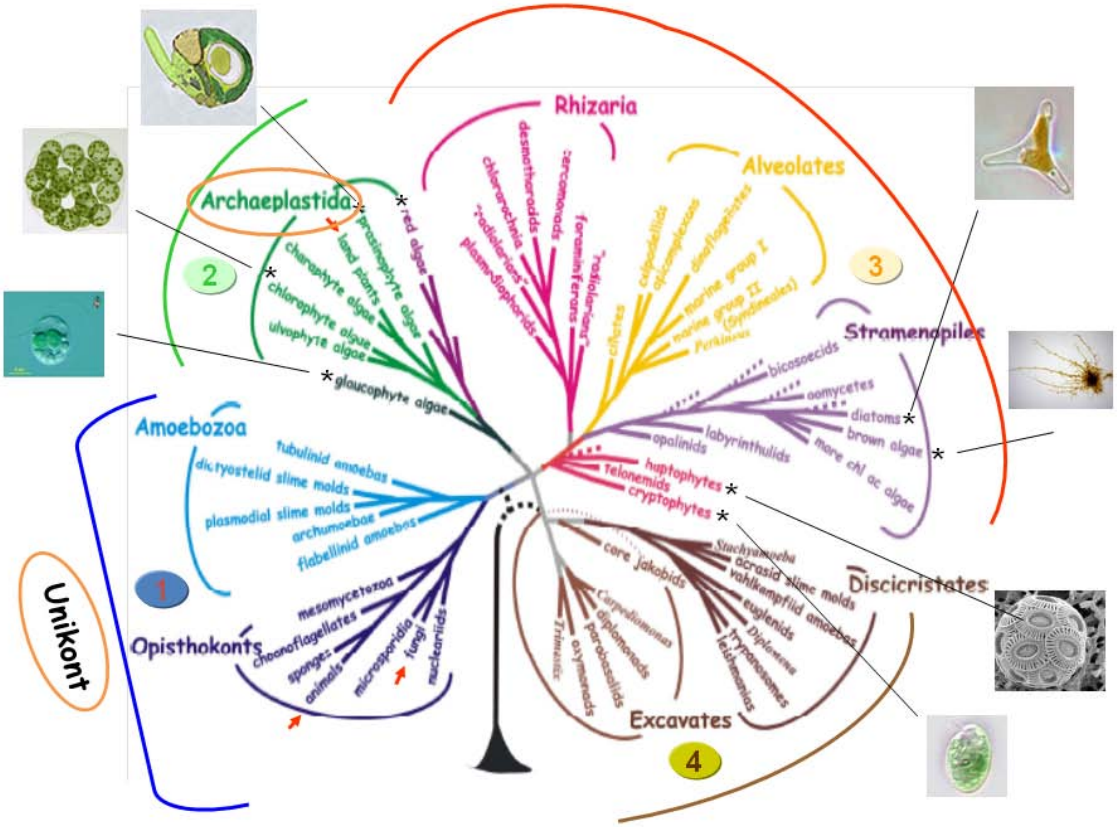


Figure 1.2 Eukaryote phylogenetic tree derived from different molecular phylogenetic and ultrastructural studies (adapted from Baldauf 2008). There is today a consensus for assigning known eukaryotes to one of the major groups in the tree of life. Stars indicate algal species for which genome sequences or ESTs are available (*Volvox carteri*, a representative of Chlorophytes, *Micromonas* for Prasinophytes, *Chroomonas* for chryptophytes, *Emiliania huxleyi* for Haptophytes, *Ectocarpus* for Brown algae and *Phaeodactylum* for diatoms). Numbers 1 to 4 represent the different groups of Eukaryotes. Red arrows point to the plant, animal and fungal kingdoms. Unikont and Archaeplastida are in orange circles (taken from Tirichine & Bowler, 2011).

1.2.2 DNA methylation patterns

DNA methylation studies in mammals have shown that 5 meC is found throughout the genome with the exception of short unmethylated regions called CpG islands (Bird, 2002). CpG islands are genomic regions that contain a high frequency of CpG sites which are often located at the promoter regions of housekeeping genes. Approximately 70% of annotated gene promoters are associated with CpG islands in vertebrates. CpG islands are involved in gene transcription. For example, CpG promoter silencing can be achieved by DNA methylation (Deaton & Bird, 2011). Compared to mammals with global methylation, the most frequent distribution pattern of DNA methylation in invertebrates is mosaic methylation, defined as both heavily methylated regions interspersed by methylation free domains. In fungi, 5mC is only on repetitive sequences (Ooi & Bestor, 2008), whereas the model plant species *A. thaliana* contains not only extensive methylation on TEs and repeat sequences but also around one third of genes are methylated (Lister, et al 2008).

DNA methylation can be established upon non-methylated sites, where it is called de novo methylation. Symmetric methylated sites become hemi methylated after DNA replication. DNA methylation can also be established upon hemi-methylated sites after DNA replication called maintenance DNA methylation. There are huge differences in DNA methylation which is called propagation in plants and mammals. In plants, the pre-existing DNA methylation mark in the previous generation can propagate across generations, whereas mammals erase the mark during zygote formation and re-establish it through cell division during development. The context of DNA methylation is also different in plant and mammals. In plants, notably *A. thaliana*, methylation occurs in all DNA sequence contexts i.e., symmetrical, CG or CHG sites (in which H is A, T or C) and asymmetrical CHH sites. In mammals, it occurs almost exclusively in the symmetric CpG dinucleotide context, with however a small amount of non CpG methylation, observed in embryonic stem cells (Lister, Pelizzola, Dowen et al, 2009).

1.2.3 The DNA methylation machinery

Cytosine DNA methylation is catalyzed by a family of conserved enzymes known as cytosine DNA methyltransferases (C5-MTases). In mammals, DNA methylation is coordinated by a family of DNA methyltransferases DNMT1, 3A, 3B, and DNA methyltransferase 3 Like that lacks catalytic activity but is essential for the function of DNMT3A and DNMT3B (Chedin,

Lieber, & Hsieh, 2002; Hsieh, 1999; Okano, Bell, Haber, & Li, 1999). DNMT1 is the most abundant DNA methyltransferase thought to be responsible for maintenance of methylation allowing thus its faithful propagation during cell division. In contrast, DNMT3A and 3B are *de novo* methyltransferases (Okano et al., 1999) which establish the methylation of previously unmethylated sequences. Another member of animal DNA methyltransferases is Dnmt2 which has highly conserved catalytic motifs with homologues among protists, fungi and plants. However, this protein has no reported DNA methyltransferase activity, which was attributed to the insertion of a serine residue into a critical prolinecysteine dipeptide that is essential for DNA methyltransferase activity in other enzymes. It was demonstrated subsequently that Dnmt2 uses a DNA methyltransferase mechanism for RNA methylation (Schaefer et al., 2010; Schaefer, Pollex, Hanna, & Lyko, 2009)

In plants, three C5-MTases have been characterized: DNA Methyltransferase 1, MET1 (known as DMT1), Chromomethylase 3 (CMT3) and Domains Rearranged Methyltransferase 2 (DRM2). MET1, the homolog of DNMT1, is also responsible for maintenance of DNA methylation. CMT3 is a plant specific C5-Mtase responsible for CHG methylation. *De novo* methylation is catalysed by DRM2, a homologue of DNMT3 (Cao *et al.* 2003). DRM2 is responsible for three contexts of DNA cytosine methylation CG, CHG and CHH, but its role in CHH methylation is the most prominent.

1.2.4 Functions of DNA methylation

In different organisms, 5meC is a conserved epigenetic mechanism crucial for a number of developmental processes such as regulation of imprinted genes, X-chromosome inactivation, silencing of repetitive elements including viral DNA and transposable elements (TEs) and regulation of gene expression (Law & Jacobsen, 2010; Sharp et al., 2011; Suzuki & Bird, 2008).

DNA methylation functions primarily in epigenetic silencing such as inactivation of TEs. However, recent discoveries about gene methylation imply that DNA methylation has more complicated functions such as preventing cryptic gene transcription and regulation of exon splicing (Foret et al., 2012). In plants, *de novo* methylation is guided by small RNA, known as RNA-dependent DNA methylation (Aufsatz, Mette, van der Winden, Matzke, & Matzke, 2002). *De novo* DNA methylation can also spread through the existing loci (Ahmed, Sarazin,

Bowler, Colot, & Quesneville, 2011). After DNA replication, symmetric methylated sites become hemi methylated. In this way, the DNA methylation maintenance machinery can recognize the sites and add a methyl group to the corresponding newly synthesized unmethylated strands. Recently DNA methylation studies have expanded from model systems to different organisms from various lineages.

1.2.5 Evolution of DNA methylation

Feng et al (Feng & Jacobsen, 2011) and Zemach et al (Zemach et al., 2010a) expanded the list of methylomes to 20 additional species utilizing bisulfite sequencing. Their research revealed the phylogenetic relationship of different organisms in the context of DNA methylation and the common characteristics of DNA methylation. Most methylated cytosines are found in repetitive elements, proposed to reflect the ancient conserved role of cytosine methylation against foreign DNA sequences. Cytosine methylation is concentrated mainly on TEs, except in invertebrates, which have more cytosine methylation in active genes. Gene body methylation was found conserved among all bisulfite sequenced organisms (**Figure 1.3**). Interestingly, gene body methylation to some extent parabolicly correlates with gene expression: moderately expressed genes having the highest levels of gene body methylation (Zemach et al., 2010a). Furthermore, gene body methylation tends to be more within exons than introns. Methylation levels were typically found to drop at Coding DNA sequence (CDS) start sites, increases in the body of genes and falls towards at the end of CDS. These findings imply general roles of DNA methylation besides TE silencing such as transcriptional elongation, termination, and perhaps alternative splicing (Feng et al., 2010). The negative correlation between gene expression and methylation was observed at the 3’ end of genes in the rice genome which suggests that the transcription termination site is important for gene expression. The inverse correlation of gene expression and promoter methylation was observed in all the species except the invertebrates (Zemach et al., 2010 Suzuki & Bird, 2008). In silk worm, CG methylation enriched in gene bodies is positively correlated with gene expression levels, suggesting that it has a positive role in gene transcription (Xiang et al., 2010). In honey bee, at least 560 differentially methylated ubiquitously expressed genes are involved in generating molecular brain diversity in Queen and workers, however it is not clear how methylation is linked with the gene regulatory network (Frank Lyko et al., 2010). Generally highly developed multicellular organisms tend to be more methylated, perhaps

because of DNA methylation control on TEs and on genes involved in cell development and differentiation during embryogenesis.

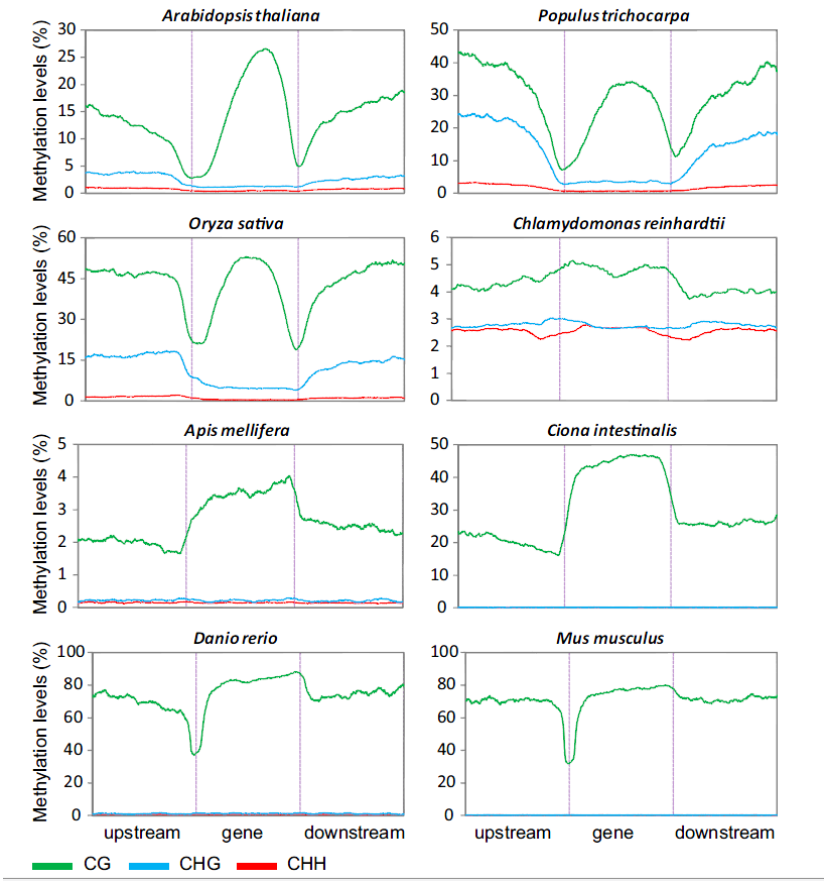


Figure 1.3 Distribution of methylation along protein-coding genes in 8 organisms. Upstream and downstream regions are the same length as the gene (taken from Feng et al., 2010).

1. 3. Histone modifications

There are a large number of different histone post-translational modifications (PTMs) including methylation, acetylation, phosphorylation, ubiquitination, sumoylation, deimination, β-N-acetylglucosamination, ADP ribosylation, ubiquitylation, histone tail clipping and histone proline isomerization (Bannister & Kouzarides, 2011). However, these modifications are usually limited to several amino acids, for example lysine acetylation, lysine and arginine methylation, serine and threonine phosphorylation, and lysine ubiquitination and sumoylation. Histone modifications are generally represented by the following system: 1. the name of the histone (e.g., H3), 2. The single letter amino acid abbreviation (e.g., K for lysine) plus the number of position in the protein sequence, 3. The type of modification (e.g., me: methyl, P: phosphate, ac: acetylation) plus the number of groups present.

1.3.1 Histone modification detection

Different histone modification detection approaches are available, with mass spectrometry and chromatin immunoprecipitation-based techniques being the most popular. Mass spectrometry, an analytical technique that measures the mass-to-charge ratio of charged particles, can quantitatively and qualitatively detect histone modifications. High sensitivity and high mass accuracy make this technique very well suited for characterization of histone modifications (Sidoli, Cheng, & Jensen, 2012). In this method, histone proteins first have to be isolated. The following steps are separation of isotypes, enrichment of modified peptides and mass spectrometry analysis. Though mass spectrometry analysis allows for a global view of the abundance of different histone modifications, a more detailed and “local” histone modification information has to be generated by ChIP-Chip or ChIP-seq. ChIP-Chip combines chromatin immunoprecipitation with microarray technology. ChIP-seq is a more advanced technique which combines chromatin immunoprecipitation and high throughput sequencing. ChIP-Chip and ChIP-seq enable detection of histone modification patterns at a give locus. In these method, a chromatin extract is prepared (not histone proteins) and antibodies specific for each histone modification are used for immunoprecipitation of chromatin that carries the modification in question (Bernstein, Humphrey, Liu, & Schreiber, 2004). Both mass spectrometry and chromatin immunoprecipitation have been used for analyzing histone modifications in different organisms (Garcia, Shabanowitz, & Hunt, 2007;

Jayani, Ramanujam, & Galande, 2010). The combination of ChIP-seq and mass spectrometry is very promising for gaining deeper insight into histone modifications in different organisms.

1.3.2 The functions of histone modification

The array of modifications on histones provides enormous potential for regulating transcription in different cell types and responding to different environmental stimuli. It has been proposed that histone modifications exert their effects either positively or negatively on gene expression through at least two distinct mechanisms. The first mechanism involves the modification of the electrostatic charge of histone which may influence the structure of chromatin over long and short distances. The second mechanism involves these modifications negatively or positively regulating the binding of effector molecules by changing the binding sites for molecule recognition and combination.

Specific histone modifications are associated with various chromatin mediated processes such as heterochromatin formation and regulation of transcription. The general functions of different histone modifications are summarized in **Table 1.1** and **Figure 1.4.** However, in different organisms, the same "code" correlates with different chromatin states. For example, H3K9me3 tends to mark active genes in *A. thalinana* whereas in *Drososphila* and mammals this mark is deposited in heterochromatic regions (Roudier et al., 2011a; Wang, Schones, & Zhao, 2009).

The H3K4me2, H3K9me2 and H3K27me3 histone modification marks have been extensively studied in different organisms. H3K4me2 associates with genes without any discrimination in expression level in different organisms. H3K9me2 and H3K27me3 are usually associated with TEs in different organisms (Barth & Imhof, 2010; Lennartsson & Ekwall, 2009).

Table 1.1 Histone modifications and their proposed functions (taken from PhD thesis of Ikhlak Ahmed).

Histone	Residue	Type of modification	Proposed Function
H1	S27	Phosphorylation	Transcriptional activation
	K26	Methylation	Transcriptional silencing
H2A	K4 (S.cerevisiae), K5 (mammals), K7 (S.cerevisiae)	Acetylation	Transcriptional activation
	S1,T119 (D.melanogaster)	Phosphorylation	Mitosis
	S122 (S.cerevisiae), S129 (S.cerevisiae), S139 (mammalian H2AX)	Phosphorylation	DNA repair
	K119 (mammals)	Ubiquitination	Transcriptional silencing
	K126 (S.cerevisiae)	Sumoylation	Transcriptional silencing
	K9, K13	Biotinylation	Unknown
H2B	K5, K11(S.cerevisiae), K12 (mammals), K15 (mammals), K16 (S.cerevisiae), K20	Acetylation	Transcriptional activation
	S10 (S.cerevisiae), S14 (Vertebrates)	Phosphorylation	Apoptosis
	S33 (D.melanogaster)	Phosphorylation	Transcriptional activation
	K34 (D.melanogaster), K119 (S. pombe), K120 (mammals), K123 (S.cerevisiae), K143 (Arabidopsis)	Ubiquitination	Transcriptional activation
	K6 or K7 (S. cerevisiae)	Sumoylation	Transcriptional silencing
H3	K4, K9, K14, K18, K23, K27, K56 (S.cerevisiae)	Acetylation	Transcriptional activation
	K4, R17	Methylation	Transcriptional activation
	K36, K79	Methylation	Transcriptional activation (elongation)
	K9, K27, R8	Methylation	Transcriptional silencing
	K9me3 (tri-methylation in Arabidopsis)	Methylation	Transcriptional activation
	T3, S10, T11 (mammals), S28 (mammals)	Phosphorylation	Mitosis
	K4, K9, K18	Biotinylation	Transcriptional activation
H4	K5, K12	Acetylation	Histone deposition
	K8, K16	Acetylation	Transcriptional activation
	K91 (S.cerevisiae)	Acetylation	Chromatin assembly
	R3,	Methylation	Transcriptional activation
	K20	Methylation	Transcriptional silencing
	K59	Methylation	Transcriptional silencing
	S1	Phosphorylation	Mitosis, Chromatin assembly, DNA repair
	K12	Biotinylation	DNA damage response

H1: Linker histone; K: Lysine; R: Arginine; S: Serine; T: Threonine

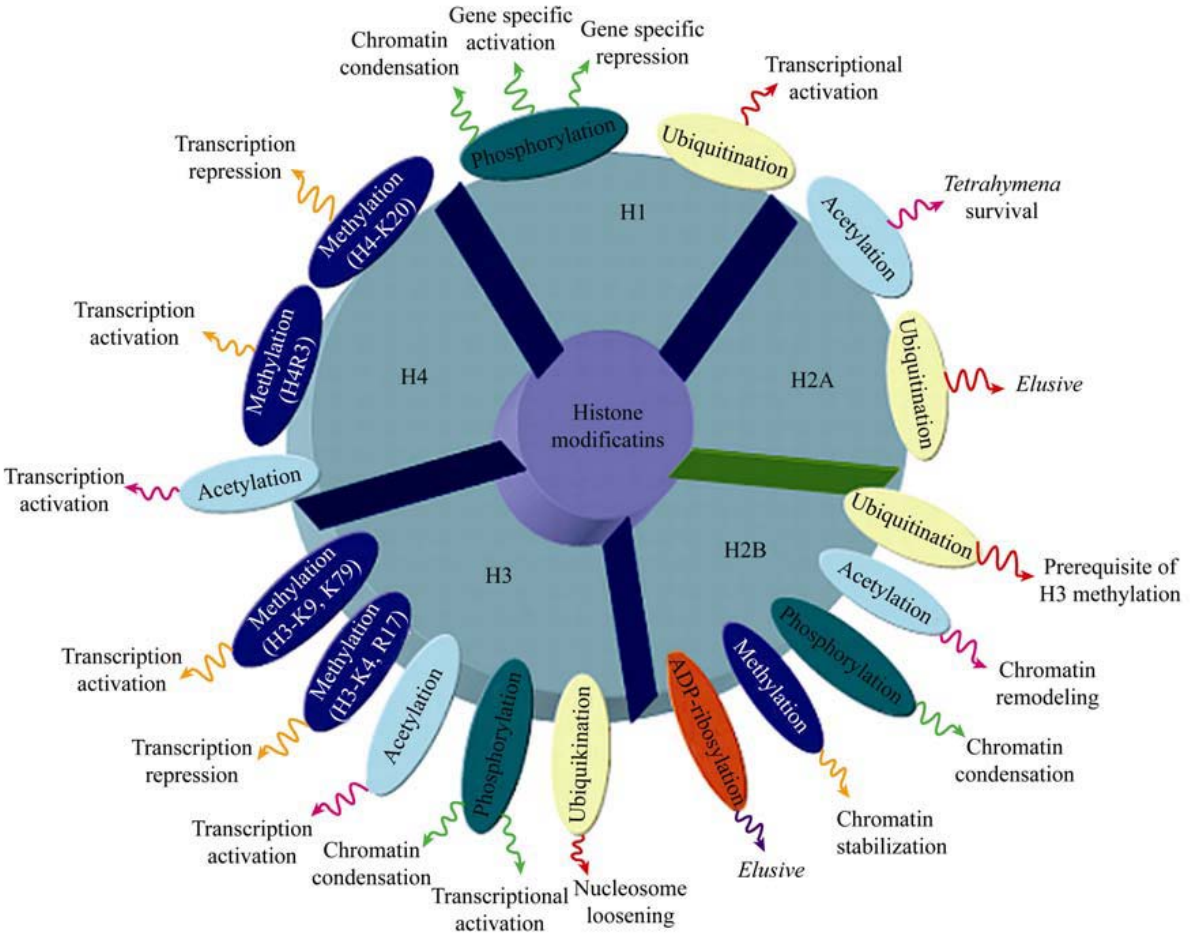


Figure 1.4 Pictorial representation of the histone modifications and their biological roles (taken from Munshi, Shafi, Aliya, & Jyothy, 2009)**.**

1.3.3 The Histone modification machinery

Multiple histone-modifying enzymes have been identified. The first histone modification enzymes, histone acetyltransferase (HAT) and histone deacetylase from *Tetrahymena sp*., were detected in 1996 (Brownell et al., 1996; Taunton, Hassig, & Schreiber, 1996). Later the enzymes responsible for adding histone modifications have been identified in different organisms. For instance, in human "Histome", a knowledgebase, listed 152 histone-modifying enzymes (Khare et al., 2012). The Histone Infobase is available online: http://www.actrec.gov.in/histome/how_to_use.php.

Histone modifications can be transmitted through cell division from parent DNA strands. During S phase, the parental nucleosomes are disrupted into two H2A-H2B dimers and an (H3-H4)2 tetramer or H3-H4 dimers at the replication fork. After replication, the strand containing the histone modification is recognized as the template for adding the histone modification on the newly incorporated histones.

1.4 RNA interference

RNA interference (RNAi) is a regulation system that controls inactivation and activation of genes. RNAi has been found to have a crucial role not only in defending cells against viruses and TEs but also in regulating gene expression, implying its profound impact on many biological processes. RNAi has been extensively studied in different models. RNA molecules involved in this pathway mainly include micro RNAs (miRNA), small interfering RNAs (siRNA) and Piwi-interacting RNAs (piRNA), reflecting their distinct ways of production in plants and animals.

The RNAi pathway is initiated by short fragments (20-30 nucleotides) of double stranded RNA cleaved by an enzyme called Dicer from long double-strand RNA (dsRNA)(Macrae et al., 2006). dsRNA can be exogenous, e.g., resulting from infection by viruses or a result of bacteria or laboratory manipulations. These short fragments of double stranded RNA are termed small RNAs (sRNA). One strand of sRNA is the guide strand which is incorporated into the RNA-induced silencing complex (RISC) containing the catalytic component of the RISC complex-known as Argonaute. The initiating dsRNA can also be endogenous, generated from pre-microRNAs. The primary transcripts of about 70 bp from RNA-coding

genes first form a characteristic stem-loop structure of pre-miRNA in the nucleus. The “stem” part is then bound and cleaved by Dicer into 22 bp mature miRNA which can incorporate into the RISC complex. Exogenous and endogenous dsRNA converge at the step of RISC incorporation (Bagasra & Prilliman, 2004). The RISC complex can cleave the mRNA which is the complimentary sequence of the guide strand sRNA or miRNA which incorporate into the complex through Argonaute. The whole process leads to the degradation of mRNA. This phenomenon is well studied and called post-transcriptional gene silencing (Hammond, Caudy, & Hannon, 2001).

miRNAs are the most well studied class of sRNAs. miRNAs were found in plant and animals and have an average length of 22 nucleotides. They exert their functions by targeting mRNA and result in transcriptional repression, degradation of transcripts, and gene silencing (Bartel, 2009). The human genome encodes over 1000 miRNAs which may regulate up to 60% of total genes (Lewis, Burge, & Bartel, 2005). miRNAs have different characteristics in plant and metazoans. In plants, miRNAs are usually perfectly or nearly perfectly complementary to their target genes whereas metazoan miRNAs are more divergent (Saumet & Lecellier, 2006).

siRNAs were first discovered as part of post-transcriptional gene silencing (PTGS) and primarily thought to be exogenous, generated by viruses, TEs and transgenes (Hamilton & Baulcombe, 1999). In 2004, another form of siRNA was found in plants known as trans-acting siRNA (ta-si RNA). These are transcribed from the genome to form polyadenylated and double-stranded precursors to regulate sets of targeted genes (Vazquez et al., 2004). More recently additional endogenous siRNAs were discovered such as natural siRNAs (nat-siRNA), repeat associated siRNAs (ra siRNA), convergent mRNA transcripts, duplexes involving pseudogene-derived antisense transcripts, mRNAs from their cognate genes, and hairpin RNAs (hpRNAs) from different sources (Golden, Gerbasi, & Sontheimer, 2008). It is clear that beside the job of defending against external foreign nucleic acids, siRNAs have an inside job of regulating endogenous genes.

In plants it is also known that dsRNAs are involved in RdDM (RNA directed DNA methylation) which is a mechanism whereby small double-stranded RNAs are processed to guide *de–novo* methylation of complementary DNA for transcriptional gene silencing (TGS). The siRNAs triggered for TGS in plants are mainly 24 nucleotides in length which

distinguishes them from the 22 nucleotide long micro RNAs (Aufsatz et al., 2002; Xizhe Zhang & Rossi, 2012). Beside miRNA and siRNA, Piwi-interacting RNAs (piRNA) have been implicated in both epigenetic and post-transcriptional gene silencing of retrotransposons and other genetic elements in animal germ line cells (Siomi, Sato, Pezic, & Aravin, 2011). The average size of piRNA is 26-31nt which is longerthan the average size of miRNAs. piRNA biogenesis mechanisms are still under discussion. However, it is certain that piRNA production does not require Dicer cleavage which is the common step in miRNA and siRNA production. A "ping-pong" model has been proposed to describe piRNA biogenesis. In this model, primary piRNA and secondary piRNA are involved in a cycle for piRNA synthesis and functioning (Brennecke et al., 2007).

Although, there is no evidence for 5-methylcytosine in small non coding RNAs, there are some other modifications. For example, 2'-O-methylation catalyzed by HEN1 has been shown to function in spliceosomal assembly of small silencing RNAs (miRNAs, siRNAs, piRNAs) and stabilization of Piwi-associated small RNAs (Kurth & Mochizuki, 2009).

1.5. Interactions between epigenetic components

Epigenetic components are not independent: they interact with each other, either positively or negatively influencing each other. Research on global gene expression and epigenetic profiles in different model organisms has revealed a well-interconnected and dynamic epigenetic network for transcription. The pattern of epigenetic marks defines the conformation of chromatin and further control the transcriptional profile. It has been shown that the crosstalk between DNA methylation and histone modifications can be mediated by biochemical interactions between histone methyltransferases and DNA methyltransferases. For example, in the mammalian embryo the methylation of H3K4 including mono, di and tri methyltion might form across the genome before *de novo* DNA methylation. Once the H3K4 tail is methylated, the recruitment of *de novo* methylation will be inhibited (Cedar & Bergman, 2009). DNMT3L, which binds the H3K4 tail and recruits methyltransferases to DNA cannot bind the H3K4 tails anymore because of the methylation of the latter. As a result, the presence of H3K4me at CpG islands prevents DNA methylation at these regions (Ooi et al., 2007).

DNA methylation of CpG islands at promoter regions negatively correlates with gene expression. In this way, DNA methylation and H3K4me are antagonistically correlated to

each other for controlling gene expression. On the other hand, DNA methylation and histone modification can coordinate together to regulate gene silencing. During mammalian embryo development, a number of these genes might become the targets for *de novo* DNA methylation and a large proportion of these genes are initially marked by Polycomb which catalyzes trimethylation of H3K27. These three epigenetic marks cooperate to achieve silencing during cell differentiation (Mohn et al., 2008).

It has also been found that the conversation among H3K9 methylation, histone deacetylation and DNA methylation leads to transcriptional silencing (Cedar & Bergman, 2009). Furthermore, observations in plants, fungi, and mammals show that methylation of H3K9 is a prerequisite for DNA methylation (Jackson, Lindroth, Cao, & Jacobsen, 2002; Lehnertz et al., 2003; Tamaru & Selker, 2001). The transcriptional repression by histone modifications is usually considered as a liable mark while DNA methylation is a relatively stable silencing mark. These two layers of information team up for gene silencing at certain regions. In *A. thaliana*, H3K9me2 positively correlates with DNA methylation in constitutive heterchromatin regions rich with hypermethylated TEs (Schoft et al., 2009). H3K9me2 is catalyzed by the histone methyltransferase SUVH4/KYP. SUVH4/KYP is also required for maintenance of non-CG methylation (Jackson et al., 2002).

DNA methylation and small RNA pathways are closely associated with each other. In plants, RdDM is a conserved de novo DNA methylation pathway for gene silencing initiated by small RNA. In A. thaliana approximately one third of methylated DNA is enriched in siRNAs (Lister et al., 2008). Ahmed et al further provided evidence that in Arabidopsis the methylated TE sequences without matching siRNA acquire DNA methylation through spreading from adjacent siRNA-targeted regions (Ahmed et al., 2011). Another example of correlation between different epigenetic components is heterchromatinization of pericentromeric satellite repeats involved in de novo DNA methylation, histone modification and small RNA machinery during early embryo development in animals. The heterochromatinization is initiated by a Dicer mediated mechanism that recognizes RNA duplexes that naturally form at satellite sequences. The key complex of the small RNA machinery, RISC, is then specifically targeted back to pericentromeric regions where it probably recruits SUV39H1 and SUV39H2 responsible for trimethylation of H3K9. These proteins are also required for the recruitment of DNMT3A and DNMT3B for DNA methylation (Fuks, 2003).

1.6. Epigenetics and the environment

It is known that genetic regulation is implicated in dealing with adaptation to a changing environment. However, genetics alone cannot explain all the rapid changes observed in response to fluctuating environments. It is increasingly being considered that epigenetics may be behind some processes that permit some organisms to cope with hostile environments. The impact of epigenetic regulation on diversity and adaptation is largely hypothetical even though the phenomena of epigenetic changes on individuals and populations in response to environmental stresses have been observed in many organisms (Bossdorf, Richards, & Pigliucci, 2008). Organisms lacking reversible without modifying DNA sequence modifications and DNA and histone modifying more likely to speciation and become reproductively isolated (Rando & Verstrepen, 2007). A “flexible adaptation” caused by epigenetic changes may be therefore appropriate for short term responses to environmental change.

A growing number of studies demonstrate the role of the epigenome in controlling gene regulation and expression leading very often to phenotypic changes, diseases such as cancer, or different behavior that cannot be explained by mutations (Dolinoy, 2008). There are accumulating evidence showing that some epigenetic modifications can be influenced by environmental cues, including diet, physical stresses such as temperature, species density or chemicals such as toxins and it can also be stochastic due to random effects.

A striking example is seen in Agouti mice in which genetically identical twins have a different size and fur color. In slim healthy brown mice, the Agouti gene is prevented from transcription by DNA methylation while in yellow obese mice prone to diabetes and cancer, the same gene is not methylated resulting in its expression (Dolinoy, 2008). One suspected trigger of these changes is bisphenol A (BPA), a ubiquitous chemical in our environment found in many plastic bottles and known to be an en endocrine disruptor. Mice exposed to food contaminated with BPA during gestation produced yellow obese progeny where DNA methylationon the agouti gene was decreased by 31% compared to normal sized brown mice fed BPA-free food (Dolinoy, Huang, & Jirtle, 2007).

Another study involving a different environmental factor was reported in the fruit fly *Drosophila melanogaster* with white eyes. Temperature treatment changes the eye color of

the fruit fly to t red, and treated individual flies can pass on the change to their offspring over several generations without further temperature treatment (Tariq, Nussbaumer, Chen, Beisel, & Paro, 2009). The DNA sequence for the gene responsible for eye color remained the same for white eyed parents and red eyed offspring and the change was attributed to a histone modification (Tariq et al., 2009). Consistent with the work described above, a more recent study in Drosophila showed that the fission yeast homolog of activation transcription factor 2 (ATF2) which usually contributes to heterochromatin formation, becomes phosphorylated leading to its release from heterochromatin upon heat shock or osmotic stress (Seong, Li, Shimizu, Nakamura, & Ishii, 2011). This new heterochromatin state which does not involve any DNA sequence change is transmitted over multiple generations (Seong et al., 2011). Crews et al. (Crews et al., 2007) showed that mate choice behavior after exposure to environmental toxin was affected by a different DNA methylation profile which is inherited to the F3 generation. Studies on the Dutch hunger winter, during the Second World War indicate that famine exposure in the peri-conceptional period led to adverse metabolic and mental phenotypes in the next generation. Interestingly, the decrease in DNA methylation at a differentially methylated region at imprinted genes involved in growth and metabolic disease was observed (Feil & Fraga, 2011). Another example is the transcriptional repression of histone deacetylation in *Arabidopsis* by gene silencing of histone deacetylase which is a homologue of the global transcriptional regulator RPD3 in yeast which caused an accumulation of various developmental abnormalities, including early senescence, suppression of apical dominance, homeotic changes, heterochronic shift toward juvenility, flower defects, and male and female sterility (Tian & Chen, 2001).

Such phenomena encompass a wide range of phyla including plants. A natural variation of asymmetrical flower development in *Linaria vulgaris* was shown to be due to an epimutation in DNA methylation of Cycloidea, a gene encoding a transcription factor (Cubas et al., 1999). Epigenetic mechanisms are not confined to few loci but act at genome wide as shown by two recent studies which used a set of useful resources, epi recombinant inbred lines propagated from one homozygous ancestor for 30 generations in *A. thaliana* where methylation was quantified by sequencing. In half of the cases, variation in methylation correlated with gene expression levels in the absence of DNA sequence change (Becker et al., 2011; Schmitz et al., 2011). In an ecological context, variation of DNA methylation was observed in a wild

population of the rare, long-lived violet *Viola cazorlensis* which is a perennial plant endemic to mountainous habitats in southeastern Spain (Herrera & Bazaga, 2010). Using a modeling approach on data collected over many years, the authors observed that epigenetic variation is significantly correlated with long-term differences in herbivory, but only weakly with herbivory-related DNA sequence variation, suggesting that, besides habitat, substrate and genetic variation, epigenetic variation may be an additional, and at least partly independent, factor influencing plant–herbivore interactions in the field (Herrera & Bazaga, 2010).

1.7 Diatoms and epigenetics

It has been proposed that the ecological success of phytoplankton is also be due to the adaptive dynamics conferred by epigenetic regulation mechanisms because point mutation-based processes may be too slow to permit adaptation to a dynamic ocean environment (Tirichine & Bowler, 2011). In marine organisms, epigenetic mechanisms are likely to be involved in and probably play even more vital roles compared to genetic regulation for better adaptation to the marine environment. Epigenetic phenomena are associated with “soft” modes of inheritance in contrast to “hard” modes of inheritance of genetic phenomena. DNA sequence based evolution spur organisms in the ocean to adapt to their environment over geological timescales, whereas epigenetic changes could confer phenotypic plasticity to individuals over much shorter timescales.

The examples discussed above and many others show a remarkable conservation among mammals, plants and invertebrates of epigenetic mechanisms regulating gene expression. This conservation seems to go beyond these species including single celled organisms such as diatoms. Diatoms are the most successful and abundant eukaryotic phytoplankton group with great diversity and wide distribution both in ocean and fresh water (Armbrust, 2009). Diatoms are believed to contribute to at least 20% of annual primary productivity which is the equivalent of all the tropical rain forests combined. It is estimated that approximately 40% of organic matter generated annually in the ocean are attributed to diatoms. Diatoms are also well known for their delicate and elegant structure of silica frustule due to their utilization of silicic acid dissolved in seawater.

The study of diatoms is a fascinating field attracting scientists from different disciplines. Nano-scientists are trying to figure out and mimic the procedures of producing the delicate

nano-scale silica frustule of diatoms (Bradbury, 2004; Kröger & Poulsen, 2008). Ancient diatom biomass was a major contributor to fossil fuels, and today's diatoms are investigated as a source for renewable, carbon-neutral fuels for the future (Larkum, Ross, Kruse, & Hankamer, 2011; Ramachandra, Mahapatra, B, & Gordon, 2009). Diatoms are also very attractive for ecologists: diatoms are environmental indicators in fresh waters and probably can compensate global warming (McQuoid & Nordberg, 2003; Rovira, Trobajo, & Ibáñez, 2012).

I am trying to explain the success of diatoms in the contemporary oceans by genome-enabled approaches. The centric diatom *Thalasiossira pseudonana* and the pennate diatom *Phaeodactylum tricornutum* have been fully sequenced which have revealed much information about genes and genome structure in diatoms (Armbrust et al., 2004; Bowler et al., 2008). Diatoms are part of the stramenopile branch of eukaryotes (**Figure 1.1**) went through secondary endosymbiotic events involving green and red algae and gained bacterial genes through horizontal gene transfer (HGT) from bacteria (**Figure 1.5**) (Bowler et al., 2008; Moustafa et al., 2009). The different origin of genes incorporated into diatom genome provide the "melting pot" genome with highly novel metabolisms never previously found in other organisms, such as the urea cycle which was previously considered only to existing in animals but *in sillico* and experimental data showed that the urea cycle exists in both *T. pseudonana* and *P. tricornutum* (Armbrust et al., 2004 Bowler et al., 2008 Allen et al., 2011; Allen, Vardi, & Bowler, 2006).

P. tricornutum is the first fully sequenced pennate diatom with a small genome (27Mb). Compared to other diatoms, it grows easily in laboratory conditions and several molecular resources, such as EST databases (http://www.diatomics.biologie.ens.fr/EST3/est3.php) are available (Maheswari et al., 2010; Maheswari, Mock, Armbrust, & Bowler, 2009). Reverse genetics in *P. tricornutum*, such as gene knockdown and overexpression have also been well established (De Riso et al., 2009; Siaut et al., 2007). *P. tricornutum* has been used as a model diatom species for decades. It is the only species in genus *Phaeodactylum*. It was first described as a single unique species by Bohlin in 1897. It was initially described as only triradiate morphotype cell with three arms, a yellow brown chloroplast, and weakly silicified frustule. Actually *P. tricornutum* has four morphotypes: fusiform, triradiate, oval and roundish (**Figure 1.6**). Both fusiform and triradiate cells are capable of forming chains and

lightly silicified. The frustule of *P. tricornutum* oval morphotype is more silicified compared to that of fusiform and triradiate morphotype. It appears to be special compared to other diatom species because it does not have an obligate requirement for silicic acid(A. D. Martino, Meichenin, Shi, Pan, & Bowler, 2007).

Until now, 11 strains of *P. tricornutum* (Pt1, Pt2, Pt3, Pt4, Pt5, Pt6, Pt7, Pt8, Pt9, Pt10 and Pt Hongkong) have been collected around the world with different morphotypes. For example, Pt1 harvested off Blackpool, UK, the most frequently used in the lab, is fusiform. However, Pt8 has dominant triradiate cells while Pt3 has dominant oval morphotype. It was shown that morphotype changes can be regulated by changing culture conditions, depending on the strain. A common trend of increased oval cell abundance was observed as a response to stress (A. De Martino et al., 2011). The transition between different morphotypes is dynamic and reversible as shown in **Figure 1.6**. The usually property of being pleiomorphic is an important characteristic of *P. tricornutum*. I speculate that probably the reversible and dynamic transition between morphotypes is related to epigenetic regulation which is more flexible compared to genetic regulation. In this respect *P. tricornutum* appear to be a good model species for epigenetic study. Furthermore, *P. tricornutum* has been studied as a model marine diatom for decades. More molecular and physiological resources in *P. tricornutum* are available. It will be very intriguing to explore epigenetic regulation in diatom model species *P. tricornutum* in the context of marine environment.

Though whole genome sequencing is essential for a better understanding of diatom ecological success, the primary sequence is only a foundation for understanding how the genetic program is read. Another layer of heritable information which is superimposed upon the DNA sequence is the epigenetic information. In the past decades, remarkable progress has been made in the mammalian and plant epigenetics field which has helped us better understand the vital roles of epigenetic regulation in chromatin organization, gene expression and development (Ahmad, Zhang, & Cao, 2010; Becker et al., 2011; Crews et al., 2007; Lister et al., 2008, 2009; Roudier et al., 2011a; Szyf, 2009; Turner, 2009).

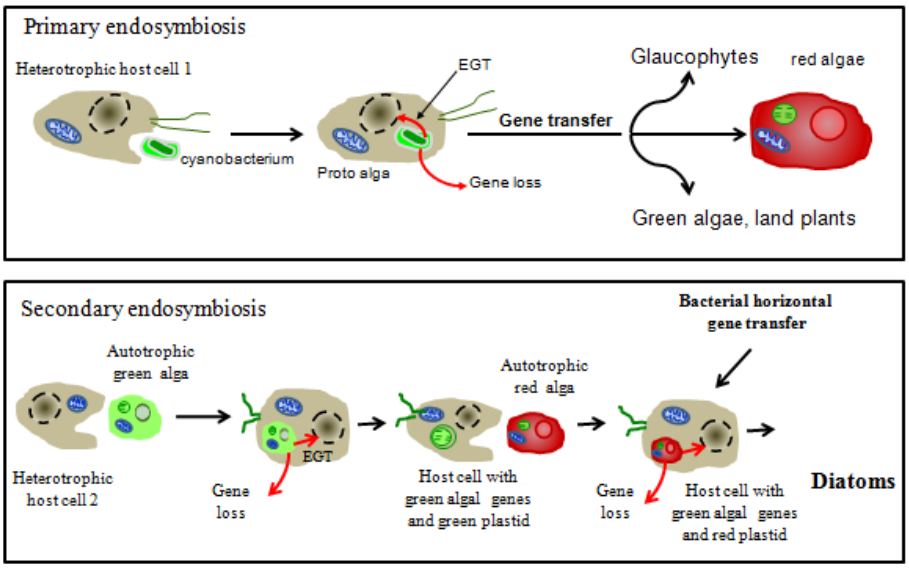


Figure 1.5 Shematic representation of the secondary endosymbiont hypothesis of diatom origin. Diatoms are chimaeras deriving from a non-photosynthetic eukaryote and combining several organisms, green, red algae and bacteria via endosymbiotic and lateral gene transfer providing them with a plethora of metabolic capabilities which may explain their ecological success. Besides genes of algal origin, Moustafa et al., (2009) have shown by comparing the genomes of *P. tricornutum* and *T. pseudonana* to hundreds of other sequenced genomes the presence of not less than 16% of green algal origin genes. EGT: endosymbiotic gene transfer; LGT: lateral gene transfer.

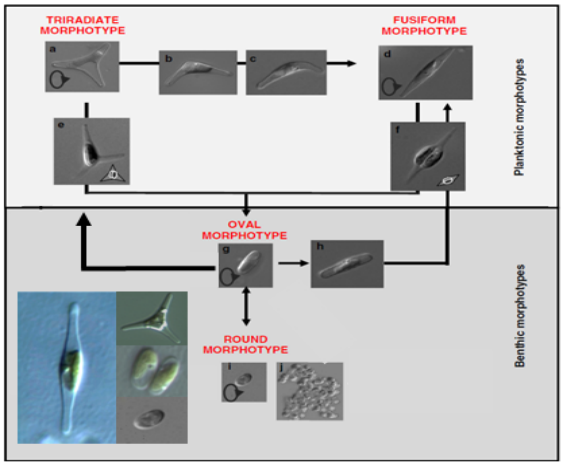


Figure 1.6 Morphotype transition in *P. tricornutum*. Fusiform, triradiate, oval, roundish morphotypes are shown in this figure. Morphotype changes can be regulated by changing culture conditions (Modified from A. De Martino et al., 2011)

1.7.1 DNA methylation and its machinery in diatoms

P. tricornutum was shown to respond to nutrient limitation and toxin exposure by decreasing DNA methylation of Long terminal repeat Retrotransposon (LTR-RTs) (Maumus et al., 2009a).The study reported that the LTR-RTs known as Blackbeard and Surcouf are methylated under normal growth conditions but become induced transcriptionally and hypomethylated under nitrate starvation and after exposure to low decadienal aldehyde, respectively without a change in genomic DNA sequence of the two retrotransposons (**Figure 1.7**) (Maumus et al., 2009a). Although not observed in their natural environment like in *Viola cazorlensis*, it is easy to imagine the impact of such factors, nutrients limitation and toxins and the resulting epigenetic influences on the development and ecological success of diatoms. The theme of "environment and epigenetics" evokes excitement because more cases of alteration of DNA methylation and histone modification triggered by environmental cues have been found. However, these studies are mostly limited to extensively studied model plants and animals. It will be very interesting to expand the "environment" to the marine environment and epigenetic research to marine organisms. Compared to multicellular organisms, the intricacy caused by different tissues and cell types is alleviated because *P. tricornutum* is unicellular and the cell cycle can be synchronized at the population level, thus the epigenetic status can also be synchronized among the cells. In a word, *P. tricornutum* can be an interesting and exciting model for epigenetic studies.

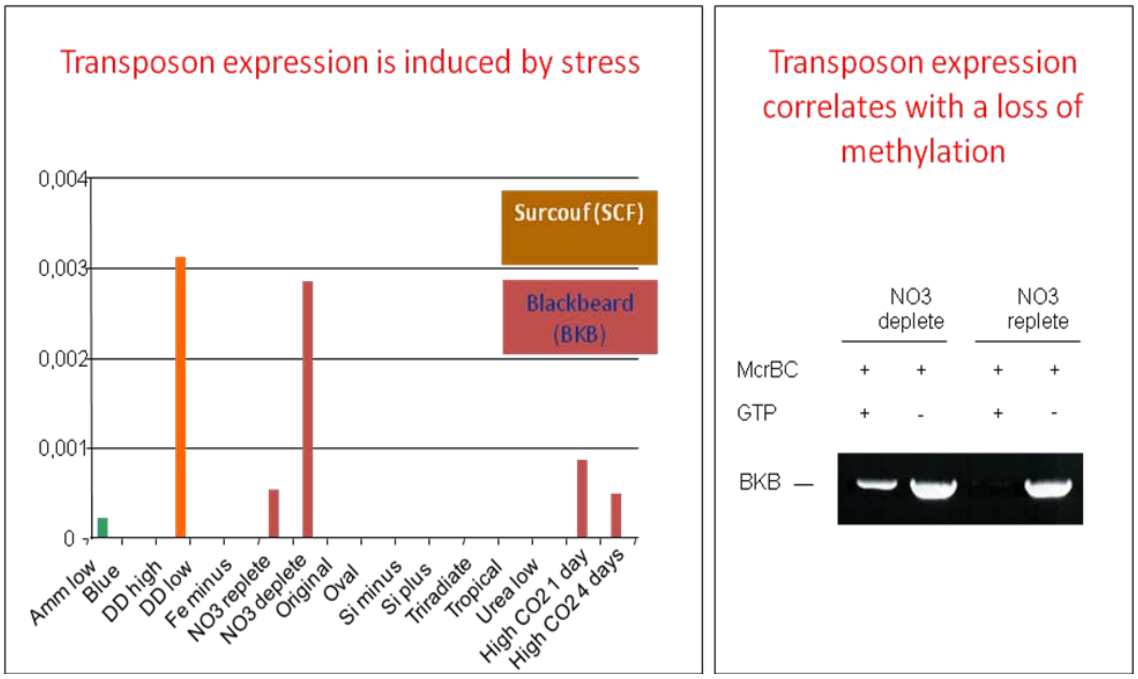


Figure 1.7 Transposon expression is induced by stress. Blackbeard became transcriptionally active and hypomethylated under nitrate starvation and without a change in genomic DNA sequence of the two retrotransposons. Surcouf became transcriptional active after exposure to low decadienal aldehyde.

Using high performance liquid chromatography, DNA methylation was reported previously in many species of algae representative of Chlorophyceae, *Charophyceae*, Prasinophyceae and Bacillariophyceae among which *P. tricornutum* (Jarvis, Dunahay, & Brown, 1992). The authors reported two ecotypes of *P. tricornutum* which have a low percentage of DNA cytosine methylation (0.11 % and 0.16% respectively),while the centric *Cyclotella cryptica* has higher cytosine methylation up to 2.23% (Jarvis et al., 1992).

The DNA methylation has been shown to occur in diatom *P. tricornutum* encodes a special set of C5-DNA methyltransferases (Maumus et al., 2011)(**Figure 1.8)**. Pt 46156 is the homologue of the DNA C5 methyltransferase Dnmt3. Pt 16674 is the putative Dnmt2 which is highly conserved in many organisms. Pt45072 combined with Pt45071 are considered as Dnmt5, a new DNA methyltransferase previously described in fungi, algae(Ponger & Li, 2005). The putative methyltransferase Pt47357 appears to have a bacterial origin (**Figure 1.8, Figure 1.9**). In bacteria, cytosine methylation is part of the restriction modification system, thus whether the bacterial Pt47357 in *P. tricornutum* is involved in a similar mechanism may be worth exploring. Interestingly, it is conserved not only in another pennate diatom *Fragilariopsis cylindrus* (Fc148014) but also in the centric diatom *T. pseudonana* (Tp 2094), from which pennate diatoms such as *P. tricornutum* diverged ~90 million years ago. This implies that a diatom common ancestor acquired DNMT from bacteria after a horizontal gene transfer prior to the centric/pennate diatom split. *P. tricornutum* does not contain a Dnmt1 homolog, but Pt44453 seems to be a Dnmt1 remnant protein which lacks the C5 methyltransferase catalytic domain but has retained two motifs characteristic of Dnmt1, the Bromo-adjacent homology (BAH) domain and a cysteine rich region (ZF_CXXX) that binds zinc ions. Putative proteins coding for plant specific DNA methyltransferase CMT3 and DRM which are responsible for non CG methylation are not detected in *P. tricornutum*. The fungal DNA methylatransferase Dim-2 also doesn't have homolog in *P. tricornutum* (Maumus et al., 2011).

In contrast of the special set of DNA methyltransferases in *P. tricornutum*, no putative DNA C5 methyltransferase could be detected in *Ectocarpus siliculosus* a brown macroalga which is also a Stramenopla. Indeed, HPLC analysis as well as McrBC-PCR experiments targeting TE loci demonstrated the complete absence of C5 DNA methylation in *E. siliculosus*. Therefore, it was speculated that C5 DNA methylation and C5 methyltransferases went through

complicated evolutionary processes leading to diverse sets of C5 methylatransferases in different lineages of life. It will be of interest to uncover the roles of the different DNMTs present in *P. tricornutum* in processes such as maintenance and *de novo* DNA methylation as well as context specificities.

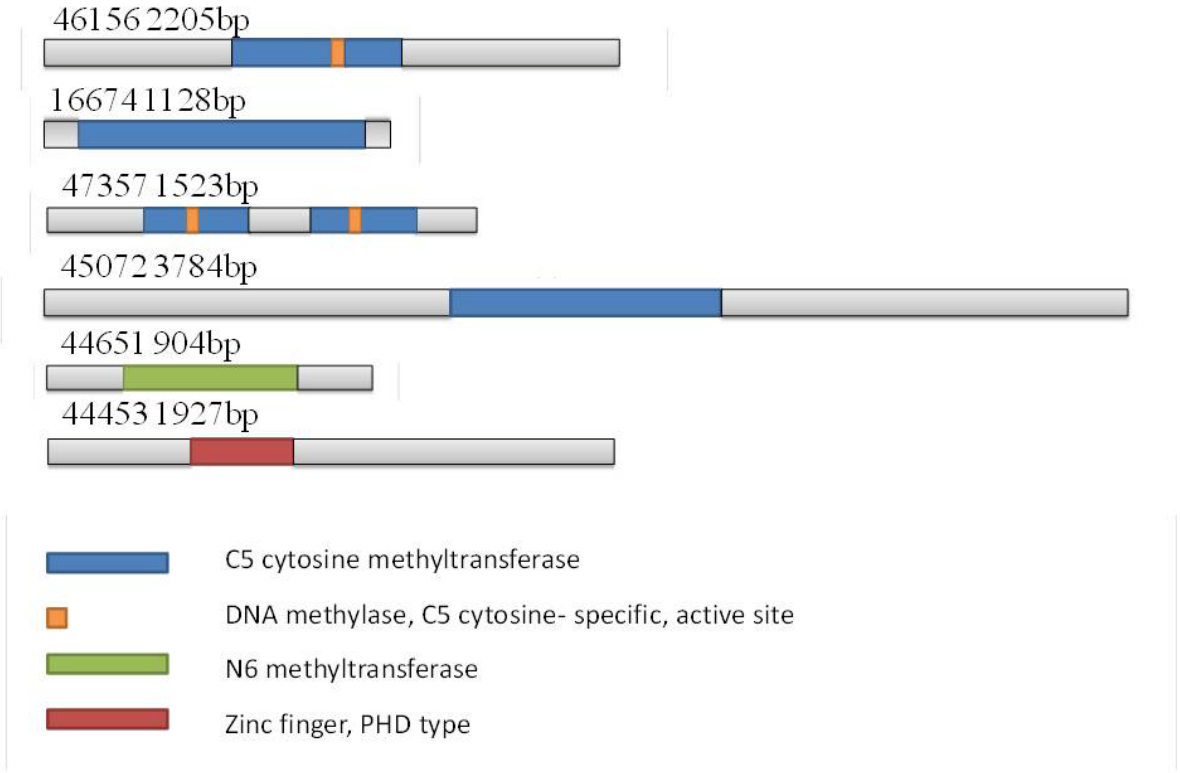


Figure 1.8 Schematic representation of putative *P. tricornutum* C5-MTases.

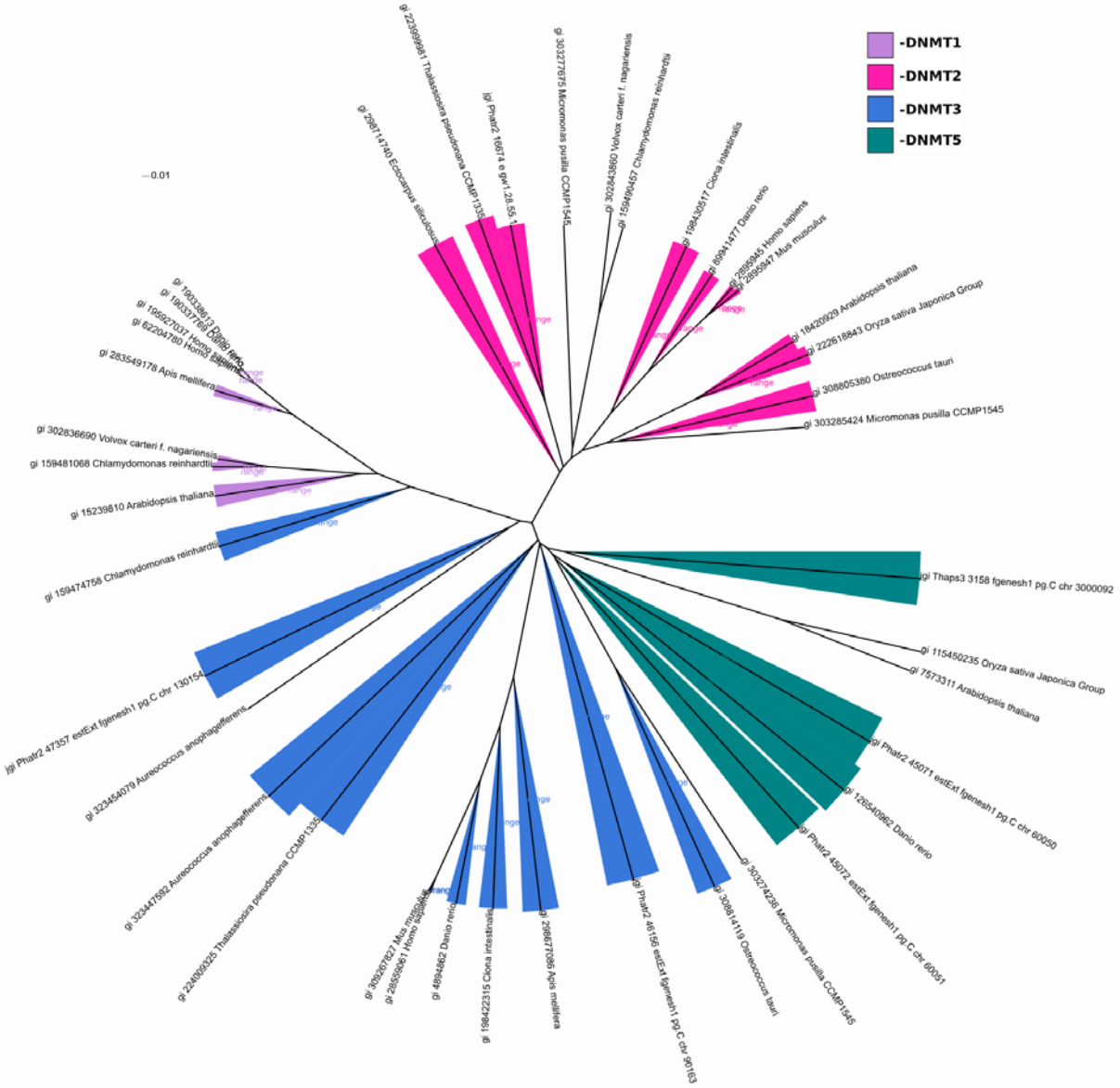


Figure 1.9 Phylogenetic tree of C5-MTases.

1.7.2 Histone modification in diatoms

Different putative genes encoding of histone methylases and demethylases have been detected in *P. tricornutum*, *T. pseudonana* and *F. cylinrus* (**Table 1.2**)(Maumus et al., 2011). The existence of these putative proteins implies that they have potential roles in diatoms. It is particularly interesting to note the presence of putative chimeric proteins related to histone modifications that have multiple-domains with antagonistic functions linked together. One interesting case is a putative Jmjc-SET fusion protein detected in *P. tricornutum* (Maumus et al., 2011). SET domains mediate lysine methylation while JmJC domains mediate demethylation of lysine. Another interesting case is a gene encoding Chromodomain-Jmjc putative fusion proteins were detected in both *P. tricornutum* and *T. pseudonana*. Chromodomains are conserved in heterochromatin protein 1(HP1) and like heterchromatin protein1 (LHP1)(Brehm, Tufteland, Aasland, & Becker, 2004; Exner et al., 2009). HP1 in *D. melanogaster* was found associated with heterchromatin formation for gene repression (James & Elgin, 1986) while LHP1 in *A. thaliana* colocalizes with H3K27me3, suggesting that LHP1 has functions similar to those of Polycomb proteins (Exner et al., 2009). In fission yeast chromodomain proteins (Chp1/Chp2/Swi6/Clr4) bind to methylated H3K9 (H3K9me) and regulate pericentric heterochromatin. H3K4 acetylation (H3K4ac) was also to have a role in the transition of dimethylated H3K9 (H3K9me2) occupancy from Chp1/Clr4 to Chp2/Swi6 (Xhemalce & Kouzarides, 2010). The functions of these chimeric proteins in diatoms are still unknown and waiting to be explore.

Previous western blotting experiments have shown the existence of H3K4me2, H3K27me3, and H3K9me2 histone modifications in *P. tricornutum* (Maumus & Bowler, 2009). To initiate histone modification study in *P. tricornutum*, I chose these three histone modifications for carrying out ChIP-seq experiments in *P. tricornutum* (see Chapter III).

Table 1.2 Gene models encoding putative enzymes responsible for histone modification which are identified in *P. tricornutum, T. psedunana* and *F. cylindrus* (adapted from Maumus et al., 2011).

Histone Modifiers	**Residues Modified**	**Homologs in *P. tricornutum***	**Homologs in *T. pseudonana***	**Homologs in *F. cylindrus***
Lysine Acetyltransferases (KATs)				
HAT1 (KAT1)	H4 (K5, K12)	54343	1397, 22580	223323
GCN5 (KAT2)	H3 (K9, K14, K18, K23, K36)	46915	15161	184437, 195196,145609, 271402, 207494, 168758,155232
Nejire (KAT3); CBP/p300 (KAT3A/B)	H3 (K14, K18, K56) H4 (K5, K8); H2A (K5) H2B (K12, K15)	45703, 45764, 54505	24331, 269496, 263785	207494, 181388, 168758, 267955, 165389, 197711, 144674, 145609, 165102, 145380
MYST1 (KAT8)	H4 (K16)	24733, 24393	37928, 36275	173464, 172955
ELP3 (KAT9)	H3	50848	9040	206186
Lysine Deacetylases (HDACs)				
RPD3 (Class I HDACS)	H2, H3, H4	51026, 49800	41025, 32098, 261393	170265, 223488,159918, 206810, 161620, 177910
HDA1 (Class II HDACS)	H2, H3, H4	45906, 50482, 35869	268655, 269060, 3235, 15819	206710, 161620, 161016, 170265, 177910, 223488
NAD+ dependant (Class III HDACS)	H4 (K16)	52135, 45850, 24866, 45909, 52718, 21543, 39523	269475, 264809, 16405, 35693, 264494, 16384, 35956	267362, 225537, 140979, 160556, 208020
Lysine				

Methyltransferases				
MLL	H3 (K4)	40183, 54436, 42693, 47328, 49473, 49476, 44935	35182, 35531, 22757	218739, 144309, 182906, 271402
ASH1/WHSC1	H3 (K4)	43275	264323	152933, 192471, 144309, 182906, 218739, 181541
SETD1	H3 (K36), H4 (K20)	not found	not found	not found
SETD2	H3 (K36)	50375	35510	192471, 144309, 152933, 182906, 259554
SETDB1	H3 (K9)	not found	not found	not found
SETMAR	H3 (K4, K36)	not found	not found	not found
SMYD	H3 (K4)	bd1647, 43708	23831, 24988	231918, 234963
TRX-related		not found	not found	not found
E(Z)	H3 (K9, K27)	32817	268872	181541, 192471, 182906, 152933, 144309, 218739, 259554
EHMT2	H3 (K9, K27)	not found	not found	not found
SET+JmjC	Unknown	bd1647	not found	not found
Lysine Demethylases (KDM)				
LSD1 (KDM1)	H3 (K4, K9)	51708, 44106, 48603	not found	138681, 264053, 240044, 261034,180908, 193836
FBXL (KDM2)	H3 (K36)	42595	not found	184520
JMJD2 (KDM4)/JARID	H3 (K9, K36)	48747	2137	206463, 237169,
JMJ-MBT	Unknown	48109	22122	237169, 206463
JMJ-CHROMO	Unknown	Pt1-40322	1863	not found

1.7.3 RNA interference in diatoms

Besidesthe numerous studies of small RNAs in plants and animals, there has been some progress reported recently in unicellular eukaryotes. Several types of sRNAs were reported in the unicellular green alga *Chlamydomonas reinhardtii* (Molnár, Schwach, Studholme, Thuenemann, & Baulcombe, 2007; T. Zhao et al., 2007). miRNAs have also been identified in the red alga *Porphyra yezoensis* and heterokont brown alga *Ectocarpus siliculosus* (Cock et al., 2010; Liang et al., 2010). It should be noted that all the miRNAs found in the organisms mentioned above show no sequence identity with known miRNAs of plants and animals in available databases.

Putative components of the RNAi machinery have been identified in multicellular organisms as well as in unicellular eukaryotes with relatively large genomes such as *Chlamydomonas reinhardtii*, but not in green and red unicellular algae with small nuclear genomes (Casas-Mollano et al., 2008). The genes encoding putative components of the RNAi machinery in diatom genomes include putative Dicer, Argonaute and RNA-dependent RNA polymerase (RdRP) (De Riso et al., 2009). Valentina De Riso et al successfully transformed a construct containing either antisense or inverted repeat sequences of the GUS reporter gene into GUS transgenic *P. tricornutum* cells which resulted in a significant down-regulation of GUS activity (De Riso et al., 2009). This work experimentally proved that siRNA generated from an endogenous dsRNA pathway does exist in *P. tricornutum.* Further molecular analyses revealed that gene silencing in *P. tricornutum* likely occurs at both transcriptional and post-transcriptional levels, although the machinery involved in the process is still obscure. Furthermore, *de novo* cytosine methylation was detected by bisulfite sequencing within antisense or inverted repeat sequence targeting regions and a discrete region up to the promoter. This shows that siRNA-mediated down-regulation of genes involves at least an RdDM process and de novo methylation cooperating together to defend against foreign DNA in *P. tricornutum*. But the mechanism of DNA methylation spreading from targeted region to flanking region was still not determined.

Huang et al identified 13 miRNAs from *P. tricornutum* cells grown under normal, silicon limited and nitrogen limited conditions(Huang, He, & Wang, 2011). None of the 13 identified miRNAs had any identifiable homologues in other organisms (Huang et al., 2011). Trina M.

Norden-Krichmar et al predicted 29 miRNA candidates over the 18–24 nucleotide size range in *T. pseudonana* (Norden-Krichmar, Allen, Gaasterland, & Hildebrand, 2011). The sequence identity analyses showed that the miRNA candidates in *T. pseudonana* also lacked strong phylogenetic conservation compared to known miRNAs. These results confirmed the existence of exogenous small RNA-miRNAs in diatoms and also imply that diatoms probably acquired a novel small RNA processing mechanism which is different from well-studied model organisms such as plants and animals.

As for siRNAs in *T. pseudonana*, a large number of sense-antisense siRNA candidates transcribed from intergenic regions were predicted, which is consistent with studies of plant small RNAs (Lu et al., 2005). However, the most interesting candidates were transcribed in the antisense direction within introns and exons, or mapped to both intergenic and protein coding regions. These characteristics suggest that the small RNAs could form double-stranded RNA with the protein coding genes, generating endogenous siRNA.

It seems that the *P. tricornutum* genome lacks methylation in the CHG context. In *P. tricornutum*, small RNAs can trigger transcriptional gene silencing (TGS) associated with DNA methylation but only in the context of CG methylation (De Riso et al., 2009). In plants three contexts of DNA methylation can be directed by small RNA. The comparison analysis of key players of RNA silencing has shown that the Argonaute-like proteins in *P. tricornutum* cluster in a clade different from those in either animals or plants. The miRNAs identified in *P. tricornutum* and another diatom *T. pseudonana* are also different from those that have been identified in animals and plants. All studies therefore indicate that diatoms probably have unique RNAi pathways.

1.8 PhD thesis description

The thesis manuscript is organized into five chapters. Chapter I provides an introduction to the others as it gives an overview of what is known in the literature about methylation, histone modifications, small RNA and their eventual interactions in a few model organisms where they have been examined, and what we know so far about diatom epigenetics.

Chapter II "methylation in the pennate model diatom *Phaeodactylum tricornutum*" is mainly about DNA methylation profiling in *P. tricornutum*. The paper "Insights into the Role of Methylation in Diatoms by Genome-Wide Profiling in *Phaeodactylum tricornutum*" has been submitted. In Chapter II, I utilized McrBC-ChIP for profiling DNA methylation distribution genome wide in *P. tricornutum*. In general, *P. tricornutum* has low DNA methylation with relative extensively methylated TEs and a few methylated genes.

In Chapter III "Genome wide mapping of H3K4me2, H3K9me2 and H3K27me3 and their correlation with DNA methylation and small RNA profiles in *Phaeodactylum tricornutum*", ChIP-seq experiments for three different histone modifications H3K4me2, H3K9me2 and H3K27me3 were conducted to explore the distribution of these histone modifications genome wide in *P. tricornutum*. RNA-seq profiling was carried out in parallel to correlate gene expression with histone modification profiles. The *in vivo* presence of these modified histones was verified by western blotting. The results show that H3K4me2 mainly marks genes while a large proportion of H3K9me2 and H3K27me3 depositions are on TEs. Thus, H3K27me3 presents an unusual pattern raising questions about its role in *P. tricornutum* and whether this particular distribution is conserved in diatoms and in single celled organisms in general. In other well studied models, such as *A. thaliana* and *D. melanogaster*, H3K27me3 mainly deposits on repressed genes not TEs. A small portion of H3K27me3 marked regions are also repressed genes which are consistent with previous studies. The Annexes of Chapter III is a submitted paper -"Chromatin immunoprecipitation coupled to detection by quantitative real time PCR to study in vivo protein DNA interactions in two model diatoms *Phaeodactylum tricornutum* and *Thalasiossira pseudonana*" which describes the chromatin immunopreccipition protocol I used for ChIP-seq experiments.

In Chapter IV " Putative epigenetic components and reverse genetic approach for dissecting epigenetic machineries in *Phaeodactylum tricornutu*m", I investigated the putative genes

encoding proteins involved in C5-DNA methylation machinery, N6 adenine methylation machinery, C5-demethylation machinery and histone lysine methyltransferases with focus on Polycomb group (PcG) proteins which play crucial roles in H3K27me3 deposition. In Chapter IV, functional analysis on three putative C-5 methyltransferases (46156, 45072 and 47357) was carried out by knocking down these genes in *P. tricornutum*. Preliminary results show that the global DNA methylation levels of 45072 and 47357knockdown lines are lower than the wild type. I investigated the functions of putative E(z) which is the key component of PcG by reverse genetic approach. The cells of E(z) 32817 *P. tricornutum* mutants tend to be roundish and oval which are different from wild type cells. It seems that E(z) in *P. tricornutum* might be involved in the morphotype fate based on current experiment results. I also explored the distribution of PcG proteins in unicellular algae including three diatom species. It is known that Polycomb group proteins are associated with H3K27me3. The novel distribution of H3K27me3 in *P. tricornutum* provoked my interest to investigate the distribution of polycomb group proteins in unicellular organisms. This is of a particular interest because the H3K27me3 mark initially considered to exist only in multicellular organisms. I showed that the PcG proteins have a distinct distribution in *P. tricornutum* compared to other well studied multicelluar organisms. The wide distribution of polycomb group proteins in unicellular algae further confirmed the fact that not only exist in unicellular organisms but also might have different functions.

Chapter V “Discussion and perspective” summarizes the thesis and gives a perspective of diatom epigenetic study and its relationship to environment.

1.9 References

Ahmad, A., Zhang, Y., & Cao, X.-F. (2010). Decoding the epigenetic language of plant development. *Molecular plant*, *3*(4), 719–28. doi:10.1093/mp/ssq026

Ahmed, I., Sarazin, A., Bowler, C., Colot, V., & Quesneville, H. (2011). Genome-wide evidence for local DNA methylation spreading from small RNA-targeted sequences in Arabidopsis. *Nucleic acids research*, (7), 1–13. doi:10.1093/nar/gkr324

Allen, A. E., Dupont, C. L., Oborník, M., Horák, A., Nunes-Nesi, A., McCrow, J. P., Zheng, H., et al. (2011). Evolution and metabolic significance of the urea cycle in photosynthetic diatoms. *Nature*, *473*(7346), 203–7. doi:10.1038/nature10074

Allen, A. E., Vardi, A., & Bowler, C. (2006). An ecological and evolutionary context for integrated nitrogen metabolism and related signaling pathways in marine diatoms. *Current opinion in plant biology*, *9*(3), 264–73. doi:10.1016/j.pbi.2006.03.013

Aravind, L., Abhiman, S., & Iyer, L. M. (2011). *Natural history of the eukaryotic chromatin protein methylation system. Progress in molecular biology and translational science* (1st ed., Vol. 101, pp. 105–76). Elsevier Inc. doi:10.1016/B978-0-12-387685-0.00004-4

Armbrust, E. V. (2009). The life of diatoms in the world's oceans. *Nature*, *459*(7244), 185–92. doi:10.1038/nature08057

Armbrust, E. V., Berges, J. a, Bowler, C., Green, B. R., Martinez, D., Putnam, N. H., Zhou, S., et al. (2004). The genome of the diatom Thalassiosira pseudonana: ecology, evolution, and metabolism. *Science (New York, N.Y.)*, *306*(5693), 79–86. doi:10.1126/science.1101156

Aufsatz, W., Mette, M. F., van der Winden, J., Matzke, A. J. M., & Matzke, M. (2002). RNA-directed DNA methylation in Arabidopsis. *Proceedings of the National Academy of Sciences of the United States of America*, *99 Suppl 4*, 16499–506. doi:10.1073/pnas.162371499

Bagasra, O., & Prilliman, K. R. (2004). RNA interference: the molecular immune system. *Journal of Molecular Histology*, *35*(6), 545–53. doi:10.1007/s10735-004-2192-8

Bannister, A. J., & Kouzarides, T. (2011). Regulation of chromatin by histone modifications. *Cell research*, *21*(3), 381–95. doi:10.1038/cr.2011.22

Bartel, D. P. (2009). MicroRNAs: target recognition and regulatory functions. *Cell*, *136*(2), 215–33. doi:10.1016/j.cell.2009.01.002

Barth, T. K., & Imhof, A. (2010). Fast signals and slow marks: the dynamics of histone modifications. *Trends in biochemical sciences*, *35*(11), 618–26. doi:10.1016/j.tibs.2010.05.006

Becker, C., Hagmann, J., Müller, J., Koenig, D., Stegle, O., Borgwardt, K., & Weigel, D. (2011). Spontaneous epigenetic variation in the Arabidopsis thaliana methylome. *Nature*. doi:10.1038/nature10555

Bernstein, B. E., Humphrey, E. L., Liu, C. L., & Schreiber, S. L. (2004). The use of chromatin immunoprecipitation assays in genome-wide analyses of histone modifications. *Methods in enzymology*, *376*(1999), 349–60. doi:10.1016/S0076-6879(03)76023-6

Bird, A. (2002). DNA methylation patterns and epigenetic memory. *Genes & development*, *16*(1), 6–21. doi:10.1101/gad.947102

Bossdorf, O., Richards, C. L., & Pigliucci, M. (2008). Epigenetics for ecologists. *Ecology letters*, *11*(2), 106–15. doi:10.1111/j.1461-0248.2007.01130.x

Bowler, C., Allen, A. E., Badger, J. H., Grimwood, J., Jabbari, K., Kuo, A., Maheswari, U., et al. (2008). The Phaeodactylum genome reveals the evolutionary history of diatom genomes. *Nature*, *456*(7219), 239–44. doi:10.1038/nature07410

Bradbury, J. (2004). Nature's nanotechnologists: unveiling the secrets of diatoms. *PLoS biology*, *2*(10), e306. doi:10.1371/journal.pbio.0020306

Brehm, A., Tufteland, K. R., Aasland, R., & Becker, P. B. (2004). The many colours of chromodomains. *BioEssays : news and reviews in molecular, cellular and developmental biology*, *26*(2), 133–40. doi:10.1002/bies.10392

Brennecke, J., Aravin, A. a, Stark, A., Dus, M., Kellis, M., Sachidanandam, R., & Hannon, G. J. (2007). Discrete small RNA-generating loci as master regulators of transposon activity in Drosophila. *Cell*, *128*(6), 1089–103. doi:10.1016/j.cell.2007.01.043

Brownell, J. E., Zhou, J., Ranalli, T., Kobayashi, R., Edmondson, D. G., Roth, S. Y., & Allis, C. D. (1996). Tetrahymena histone acetyltransferase A: a homolog to yeast Gcn5p linking histone acetylation to gene activation. *Cell*, *84*(6), 843–51. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/8601308

Casas-Mollano, J. A., Rohr, J., Kim, E.-J., Balassa, E., van Dijk, K., & Cerutti, H. (2008). Diversification of the core RNA interference machinery in Chlamydomonas reinhardtii and the role of DCL1 in transposon silencing. *Genetics*, *179*(1), 69–81. doi:10.1534/genetics.107.086546

Cedar, H., & Bergman, Y. (2009). Linking DNA methylation and histone modification: patterns and paradigms. *Nature reviews. Genetics*, *10*(5), 295–304. doi:10.1038/nrg2540

Chedin, F., Lieber, M. R., & Hsieh, C.-L. (2002). The DNA methyltransferase-like protein DNMT3L stimulates de novo methylation by Dnmt3a. *Proceedings of the National*

Academy of Sciences of the United States of America, *99*(26), 16916–21. doi:10.1073/pnas.262443999

Cock, J. M., Sterck, L., Rouzé, P., Scornet, D., Allen, A. E., Amoutzias, G., Anthouard, V., et al. (2010). The Ectocarpus genome and the independent evolution of multicellularity in brown algae. *Nature*, *465*(7298), 617–21. doi:10.1038/nature09016

Costa, F. F. (2008). Non-coding RNAs, epigenetics and complexity. *Gene*, *410*(1), 9–17. doi:10.1016/j.gene.2007.12.008

Crews, D., Gore, A. C., Hsu, T. S., Dangleben, N. L., Spinetta, M., Schallert, T., Anway, M. D., et al. (2007). Transgenerational epigenetic imprints on mate preference. *Proceedings of the National Academy of Sciences of the United States of America*, *104*(14), 5942–6. doi:10.1073/pnas.0610410104

Cubas, P., Vincent, C., Coen, E., Centre, J. I., Lane, C., & Nr, N. (1999). An epigenetic mutation responsible for natural variation in ¯ oral symmetry. *Nature*, 157–161.

De Martino, A., Bartual, A., Willis, A., Meichenin, A., Villazán, B., Maheswari, U., & Bowler, C. (2011). Physiological and molecular evidence that environmental changes elicit morphological interconversion in the model diatom Phaeodactylum tricornutum. *Protist*, *162*(3), 462–81. doi:10.1016/j.protis.2011.02.002

De Martino, A., Meichenin, A., Shi, J., Pan, K., & Bowler, C. (2007). Genetic and phenotypic characterization of Phaeodactylum tricornutum (Bacillariophyceae) accessions 1. *Journal of Phycology*, *43*(5), 992–1009. doi:10.1111/j.1529-8817.2007.00384.x

De Riso, V., Raniello, R., Maumus, F., Rogato, A., Bowler, C., & Falciatore, A. (2009). Gene silencing in the marine diatom Phaeodactylum tricornutum. *Nucleic acids research*, *37*(14), e96. doi:10.1093/nar/gkp448

Deaton, A. M., & Bird, A. (2011). CpG islands and the regulation of transcription. *Genes & development*, *25*(10), 1010–22. doi:10.1101/gad.2037511

Dolinoy, D. C. (2008). The agouti mouse model: an epigenetic biosensor for nutritional and environmental alterations on the fetal epigenome. *Nutrition reviews*, *66 Suppl 1*, S7–11. doi:10.1111/j.1753-4887.2008.00056.x

Dolinoy, D. C., Huang, D., & Jirtle, R. L. (2007). Maternal nutrient supplementation counteracts bisphenol A-induced DNA hypomethylation in early development. *Proceedings of the National Academy of Sciences of the United States of America*, *104*(32), 13056–61. doi:10.1073/pnas.0703739104

Exner, V., Aichinger, E., Shu, H., Wildhaber, T., Alfarano, P., Caflisch, A., Gruissem, W., et al. (2009). The chromodomain of LIKE HETEROCHROMATIN PROTEIN 1 is essential for

H3K27me3 binding and function during Arabidopsis development. *PloS one*, *4*(4), e5335. doi:10.1371/journal.pone.0005335

Feil, R., & Fraga, M. F. (2011). Epigenetics and the environment: emerging patterns and implications. *Nature reviews. Genetics*, *13*(2), 97–109. doi:10.1038/nrg3142

Feng, S., Cokus, S. J., Zhang, X., Chen, P.-Y., Bostick, M., Goll, M. G., Hetzel, J., et al. (2010). Conservation and divergence of methylation patterning in plants and animals. *Proceedings of the National Academy of Sciences of the United States of America*, *107*(19), 8689–94. doi:10.1073/pnas.1002720107

Feng, S., & Jacobsen, S. E. (2011). Epigenetic modifications in plants: an evolutionary perspective. *Current opinion in plant biology*, *14*(2), 179–86. doi:10.1016/j.pbi.2010.12.002

Ficz, G., Branco, M. R., Seisenberger, S., Santos, F., Krueger, F., Hore, T. a, Marques, C. J., et al. (2011). Dynamic regulation of 5-hydroxymethylcytosine in mouse ES cells and during differentiation. *Nature*, *473*(7347), 398–402. doi:10.1038/nature10008

Foret, S., Kucharski, R., Pellegrini, M., Feng, S., Jacobsen, S. E., Robinson, G. E., & Maleszka, R. (2012). DNA methylation dynamics, metabolic fluxes, gene splicing, and alternative phenotypes in honey bees. *Proceedings of the National Academy of Sciences of the United States of America*, *109*(13). doi:10.1073/pnas.1202392109

Fuks, F. (2003). The DNA methyltransferases associate with HP1 and the SUV39H1 histone methyltransferase. *Nucleic Acids Research*, *31*(9), 2305–2312. doi:10.1093/nar/gkg332

Garcia, B. a, Shabanowitz, J., & Hunt, D. F. (2007). Characterization of histones and their post-translational modifications by mass spectrometry. *Current opinion in chemical biology*, *11*(1), 66–73. doi:10.1016/j.cbpa.2006.11.022

Golden, D. E., Gerbasi, V. R., & Sontheimer, E. J. (2008). An inside job for siRNAs. *Molecular cell*, *31*(3), 309–12. doi:10.1016/j.molcel.2008.07.008

Goll, M. G., & Bestor, T. H. (2005). Eukaryotic cytosine methyltransferases. *Annual Review of Biochemistry*, *74*(1), 481–514. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/15952895

Hamilton, a J., & Baulcombe, D. C. (1999). A species of small antisense RNA in posttranscriptional gene silencing in plants. *Science (New York, N.Y.)*, *286*(5441), 950–2. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/10542148

Hammond, S. M., Caudy, a a, & Hannon, G. J. (2001). Post-transcriptional gene silencing by double-stranded RNA. *Nature reviews. Genetics*, *2*(2), 110–9. doi:10.1038/35052556

Herrera, C. M., & Bazaga, P. (2010). Epigenetic differentiation and relationship to adaptive genetic divergence in discrete populations of the violet Viola cazorlensis. *The New phytologist*, *187*(3), 867–76. doi:10.1111/j.1469-8137.2010.03298.x

Hsieh, C. (1999). In Vivo Activity of Murine De Novo In Vivo Activity of Murine De Novo Methyltransferases , Dnmt3a and Dnmt3b, *19*(12).

Huang, A., He, L., & Wang, G. (2011). Identification and characterization of microRNAs from Phaeodactylum tricornutum by high-throughput sequencing and bioinformatics analysis. *BMC genomics*, *12*(1), 337. doi:10.1186/1471-2164-12-337

Jackson, J. P., Lindroth, A. M., Cao, X., & Jacobsen, S. E. (2002). Control of CpNpG DNA methylation by the KRYPTONITE histone H3 methyltransferase. *Nature*, *416*(6880), 556–60. doi:10.1038/nature731

James, T. C., & Elgin, S. C. (1986). Identification of a nonhistone chromosomal protein associated with heterochromatin in Drosophila melanogaster and its gene. *Molecular and cellular biology*, *6*(11), 3862–72. Retrieved from http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=367149&tool=pmcentrez&rendertype=abstract

Jarvis, E. E., Dunahay, T. G., & Brown, L. M. (1992). DNA NUCLEOSIDE COMPOSITION AND METHYLATION IN SEVERAL SPECIES OF MICROALGAE1. *Journal of Phycology*, *28*(3), 356–362.

Jayani, R. S., Ramanujam, P. L., & Galande, S. (2010). *Studying histone modifications and their genomic functions by employing chromatin immunoprecipitation and immunoblotting. Methods in cell biology* (Vol. 98, pp. 35–56). Elsevier Inc. doi:10.1016/S0091-679X(10)98002-3

Jia, G., Fu, Y., Zhao, X., Dai, Q., Zheng, G., Yang, Y., Yi, C., et al. (2011). N6-Methyladenosine in nuclear RNA is a major substrate of the obesity-associated FTO. *Nature chemical biology*, *7*(december), 885–887. doi:10.1038/nchembio.687

Kerppola, T. K. (2009). Polycomb group complexes--many combinations, many functions. *Trends in cell biology*, *19*(12), 692–704. doi:10.1016/j.tcb.2009.10.001

Khare, S. P., Habib, F., Sharma, R., Gadewal, N., Gupta, S., & Galande, S. (2012). HIstome--a relational knowledgebase of human histone proteins and histone modifying enzymes. *Nucleic acids research*, *40*(Database issue), D337–42. doi:10.1093/nar/gkr1125

Kouzarides, T. (2007). Chromatin modifications and their function. *Cell*, *128*(4), 693–705. doi:10.1016/j.cell.2007.02.005

Kröger, N., & Poulsen, N. (2008). Diatoms-from cell wall biogenesis to nanotechnology. *Annual review of genetics*, *42*, 83–107. doi:10.1146/annurev.genet.41.110306.130109

Kurth, H. M., & Mochizuki, K. (2009). 2'-O-methylation stabilizes Piwi-associated small RNAs and ensures DNA elimination in Tetrahymena. *RNA (New York, N.Y.)*, *15*(4), 675–85. doi:10.1261/rna.1455509

Larkum, A. W. D., Ross, I. L., Kruse, O., & Hankamer, B. (2011). Selection, breeding and engineering of microalgae for bioenergy and biofuel production. *Trends in biotechnology*, *30*(4), 198–205. doi:10.1016/j.tibtech.2011.11.003

Law, J. a, & Jacobsen, S. E. (2010). Establishing, maintaining and modifying DNA methylation patterns in plants and animals. *Nature reviews. Genetics*, *11*(3), 204–20. doi:10.1038/nrg2719

Lehnertz, B., Ueda, Y., Derijck, A. A. H. A., Braunschweig, U., Perez-burgos, L., Kubicek, S., Chen, T., et al. (2003). Suv39h-Mediated Histone H3 Lysine 9 Methylation Directs DNA Methylation to Major Satellite Repeats at Pericentric Heterochromatin. *Current*, *13*, 1192–1200. doi:10.1016/S

Lennartsson, A., & Ekwall, K. (2009). Histone modification patterns and epigenetic codes. *Biochimica et biophysica acta*, *1790*(9), 863–8. doi:10.1016/j.bbagen.2008.12.006

Lewis, B. P., Burge, C. B., & Bartel, D. P. (2005). Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. *Cell*, *120*(1), 15–20. doi:10.1016/j.cell.2004.12.035

Liang, C., Zhang, X., Zou, J., Xu, D., Su, F., & Ye, N. (2010). Identification of miRNA from Porphyra yezoensis by high-throughput sequencing and bioinformatics analysis. *PloS one*, *5*(5), e10698. doi:10.1371/journal.pone.0010698

Lister, R., O'Malley, R. C., Tonti-Filippini, J., Gregory, B. D., Berry, C. C., Millar, a H., & Ecker, J. R. (2008). Highly integrated single-base resolution maps of the epigenome in Arabidopsis. *Cell*, *133*(3), 523–36. doi:10.1016/j.cell.2008.03.029

Lister, R., Pelizzola, M., Dowen, R. H., Hawkins, R. D., Hon, G., Tonti-Filippini, J., Nery, J. R., et al. (2009). Human DNA methylomes at base resolution show widespread epigenomic differences. *Nature*, *462*(7271), 315–22. doi:10.1038/nature08514

Lu, C., Tej, S. S., Luo, S., Haudenschild, C. D., Meyers, B. C., & Green, P. J. (2005). Elucidation of the small RNA component of the transcriptome. *Science (New York, N.Y.)*, *309*(5740), 1567–9. doi:10.1126/science.1114112

Lyko, F., Foret, S., Kucharski, R., Wolf, S., Falckenhayn, C., & Maleszka, R. (2010). The honey bee epigenomes: differential methylation of brain DNA in queens and workers. *PLoS biology*, *8*(11), e1000506. doi:10.1371/journal.pbio.1000506

Macrae, I. J., Zhou, K., Li, F., Repic, A., Brooks, A. N., Cande, W. Z., Adams, P. D., et al. (2006). Structural basis for double-stranded RNA processing by Dicer. *Science (New York, N.Y.)*, *311*(5758), 195–8. doi:10.1126/science.1121638

Maheswari, U., Jabbari, K., Petit, J.-L., Porcel, B. M., Allen, A. E., Cadoret, J.-P., De Martino, A., et al. (2010). Digital expression profiling of novel diatom transcripts provides insight into their biological functions. *Genome biology*, *11*(8), R85. doi:10.1186/gb-2010-11-8-r85

Maheswari, U., Mock, T., Armbrust, E. V., & Bowler, C. (2009). Update of the Diatom EST Database: a new tool for digital transcriptomics. *Nucleic acids research*, *37*(Database issue), D1001–5. doi:10.1093/nar/gkn905

Maumus, F., Allen, A. E., Mhiri, C., Hu, H., Jabbari, K., Vardi, A., Grandbastien, M.-A., et al. (2009). Potential impact of stress activated retrotransposons on genome evolution in a marine diatom. *BMC genomics*, *10*, 624. doi:10.1186/1471-2164-10-624

Maumus, F., & Bowler, C. (2009). " Transcriptional and Epigenetic regulation in the marine diatom Phaeodactylum tricornutum " Thesis director :*Analysis*.

Maumus, F., Rabinowicz, P., Bowler, C., Rivarola, M., Inserm, U., Nacional, I., Agropecuaria, D. T., et al. (2011). Stemming Epigenetics in Marine Stramenopiles. *Current*.

McQuoid, M. ., & Nordberg, K. (2003). The diatom Paralia sulcata as an environmental indicator species in coastal sediments. *Estuarine, Coastal and Shelf Science*, *56*(2), 339–354. doi:10.1016/S0272-7714(02)00187-7

Mohn, F., Weber, M., Rebhan, M., Roloff, T. C., Richter, J., Stadler, M. B., Bibel, M., et al. (2008). Lineage-specific polycomb targets and de novo DNA methylation define restriction and potential of neuronal progenitors. *Molecular cell*, *30*(6), 755–66. doi:10.1016/j.molcel.2008.05.007

Molnár, A., Schwach, F., Studholme, D. J., Thuenemann, E. C., & Baulcombe, D. C. (2007). miRNAs control gene expression in the single-cell alga Chlamydomonas reinhardtii. *Nature*, *447*(7148), 1126–9. doi:10.1038/nature05903

Moustafa, A., Beszteri, B., Maier, U. G., Bowler, C., Valentin, K., & Bhattacharya, D. (2009). Genomic footprints of a cryptic plastid endosymbiosis in diatoms. *Science (New York, N.Y.)*, *324*(5935), 1724–6. doi:10.1126/science.1172983

Norden-Krichmar, T. M., Allen, A. E., Gaasterland, T., & Hildebrand, M. (2011). Characterization of the Small RNA Transcriptome of the Diatom, Thalassiosira pseudonana. (I. Friedberg, Ed.)*PLoS ONE*, *6*(8), e22870. doi:10.1371/journal.pone.0022870

Okano, M., Bell, D. W., Haber, D. a, & Li, E. (1999). DNA methyltransferases Dnmt3a and Dnmt3b are essential for de novo methylation and mammalian development. *Cell*, *99*(3), 247–57. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/10555141

Ooi, S. K. T., & Bestor, T. H. (2008). The colorful history of active DNA demethylation. *Cell*, *133*(7), 1145–8. doi:10.1016/j.cell.2008.06.009

Ooi, S. K. T., Qiu, C., Bernstein, E., Li, K., Jia, D., Yang, Z., Erdjument-Bromage, H., et al. (2007). DNMT3L connects unmethylated lysine 4 of histone H3 to de novo methylation of DNA. *Nature*, *448*(7154), 714–7. doi:10.1038/nature05987

Pastor, W. a, Pape, U. J., Huang, Y., Henderson, H. R., Lister, R., Ko, M., McLoughlin, E. M., et al. (2011). Genome-wide mapping of 5-hydroxymethylcytosine in embryonic stem cells. *Nature*, *473*(7347), 394–7. doi:10.1038/nature10102

Ponger, L., & Li, W.-H. (2005). Evolutionary diversification of DNA methyltransferases in eukaryotic genomes. *Molecular biology and evolution*, *22*(4), 1119–28. doi:10.1093/molbev/msi098

Ramachandra, T. V., Mahapatra, D. M., B, K., & Gordon, R. (2009). Milking Diatoms for Sustainable Energy: Biochemical Engineering versus Gasoline-Secreting Diatom Solar Panels. *Industrial & Engineering Chemistry Research*, *48*(19), 8769–8788. doi:10.1021/ie900044j

Rando, O. J., & Verstrepen, K. J. (2007). Timescales of genetic and epigenetic inheritance. *Cell*, *128*(4), 655–68. doi:10.1016/j.cell.2007.01.023

Richmond, R. K., Sargent, D. F., Richmond, T. J., Luger, K., & Ma, A. W. (1997). Crystal structure of the nucleosome ˚ resolution core particle at 2 . 8 A. *Nature*, *7*.

Ris, H., & Kubai, D. F. (1970). Chromosome structure. *Annual review of genetics*, *4*(130), 263–94. doi:10.1146/annurev.ge.04.120170.001403

Roudier, F., Ahmed, I., Bérard, C., Sarazin, A., Mary-Huard, T., Cortijo, S., Bouyer, D., et al. (2011). Integrative epigenomic mapping defines four main chromatin states in Arabidopsis. *The EMBO journal*, *30*(10), 1928–38. doi:10.1038/emboj.2011.103

Rovira, L., Trobajo, R., & Ibáñez, C. (2012). The use of diatom assemblages as ecological indicators in highly stratified estuaries and evaluation of existing diatom indices. *Marine pollution bulletin*, *64*(3), 500–11. doi:10.1016/j.marpolbul.2012.01.005

Saumet, A., & Lecellier, C.-H. (2006). Anti-viral RNA silencing: do we look like plants? *Retrovirology*, *3*, 3. doi:10.1186/1742-4690-3-3

Schaefer, M., Pollex, T., Hanna, K., & Lyko, F. (2009). RNA cytosine methylation analysis by bisulfite sequencing. *Nucleic Acids Research*, *37*(2), e12. Retrieved from http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2632927&tool=pmcentrez&rendertype=abstract

Schaefer, M., Pollex, T., Hanna, K., Tuorto, F., Meusburger, M., Helm, M., & Lyko, F. (2010). RNA methylation by Dnmt2 protects transfer RNAs against stress-induced cleavage. *Genes & development*, *24*(15), 1590–5. doi:10.1101/gad.586710

Schmitz, R. J., Schultz, M. D., Lewsey, M. G., O'Malley, R. C., Urich, M. a, Libiger, O., Schork, N. J., et al. (2011). Transgenerational epigenetic instability is a source of novel methylation variants. *Science (New York, N.Y.)*, *334*(6054), 369–73. doi:10.1126/science.1212959

Schoft, V. K., Chumak, N., Mosiolek, M., Slusarz, L., Komnenovic, V., Brownfield, L., Twell, D., et al. (2009). Induction of RNA-directed DNA methylation upon decondensation of constitutive heterochromatin. *EMBO reports*, *10*(9), 1015–21. doi:10.1038/embor.2009.152

Seong, K.-H., Li, D., Shimizu, H., Nakamura, R., & Ishii, S. (2011). Inheritance of stress-induced, ATF-2-dependent epigenetic change. *Cell*, *145*(7), 1049–61. doi:10.1016/j.cell.2011.05.029

Sharp, A. J., Stathaki, E., Migliavacca, E., Brahmachary, M., Montgomery, S. B., Dupre, Y., & Antonarakis, S. E. (2011). DNA methylation profiles of human active and inactive X chromosomes. *Genome research*, *21*(10), 1592–600. doi:10.1101/gr.112680.110

Siaut, M., Heijde, M., Mangogna, M., Montsant, A., Coesel, S., Allen, A., Manfredonia, A., et al. (2007). Molecular toolbox for studying diatom biology in Phaeodactylum tricornutum. *Gene*, *406*(1-2), 23–35. doi:10.1016/j.gene.2007.05.022

Sidoli, S., Cheng, L., & Jensen, O. N. (2012). Proteomics in chromatin biology and epigenetics: Elucidation of post-translational modifications of histone proteins by mass spectrometry. *Journal of proteomics*, *75*(12), 3419–33. doi:10.1016/j.jprot.2011.12.029

Siomi, M. C., Sato, K., Pezic, D., & Aravin, A. a. (2011). PIWI-interacting small RNAs: the vanguard of genome defence. *Nature reviews. Molecular cell biology*, *12*(4), 246–58. doi:10.1038/nrm3089

Suzuki, M. M., & Bird, A. (2008). DNA methylation landscapes: provocative insights from epigenomics. *Nature Reviews Genetics*, *9*(6), 465–476. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/18463664

Szyf, M. (2009). The early life environment and the epigenome. *Biochimica et biophysica acta*, *1790*(9), 878–85. doi:10.1016/j.bbagen.2009.01.009

Tamaru, H., & Selker, E. U. (2001). A histone H3 methyltransferase controls DNA methylation in Neurospora crassa. *Nature*, *414*(6861), 277–83. doi:10.1038/35104508

Tariq, M., Nussbaumer, U., Chen, Y., Beisel, C., & Paro, R. (2009). Trithorax requires Hsp90 for maintenance of active chromatin at sites of gene expression. *Proceedings of the National Academy of Sciences of the United States of America*, *106*(4), 1157–62. doi:10.1073/pnas.0809669106

Taunton, J., Hassig, C. a, & Schreiber, S. L. (1996). A mammalian histone deacetylase related to the yeast transcriptional regulator Rpd3p. *Science (New York, N.Y.)*, *272*(5260), 408–11. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/8602529

Tian, L., & Chen, Z. J. (2001). Blocking histone deacetylation in Arabidopsis induces pleiotropic effects on plant gene regulation and development. *Proceedings of the National Academy of Sciences of the United States of America*, *98*(1), 200–5. doi:10.1073/pnas.011347998

Tirichine, L., & Bowler, C. (2011). Decoding algal genomes: tracing back the history of photosynthetic life on Earth. *The Plant journal: for cell and molecular biology* , *66*(1), 45–57. doi:10.1111/j.1365-313X.2011.04540.x

Turner, B. M. (2009). Epigenetic responses to environmental change and their evolutionary implications. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, *364*(1534), 3403–18. doi:10.1098/rstb.2009.0125

Tweedie, S., Ng, H. H., Barlow, a L., Turner, B. M., Hendrich, B., & Bird, a. (1999). Vestiges of a DNA methylation system in Drosophila melanogaster? *Nature genetics*, *23*(4), 389–90. doi:10.1038/70490

Vazquez, F., Vaucheret, H., Rajagopalan, R., Lepers, C., Gasciolli, V., Mallory, A. C., Hilbert, J.-L., et al. (2004). Endogenous trans-acting siRNAs regulate the accumulation of Arabidopsis mRNAs. *Molecular cell*, *16*(1), 69–79. doi:10.1016/j.molcel.2004.09.028

Wang, Z., Schones, D. E., & Zhao, K. (2009). Characterization of human epigenomes. *Current opinion in genetics & development*, *19*(2), 127–34. doi:10.1016/j.gde.2009.02.001

Williams, K., Christensen, J., & Helin, K. (2011). DNA methylation: TET proteins—guardians of CpG islands? *EMBO reports*, *13*(1), 28–35. doi:10.1038/embor.2011.233

Xhemalce, B., & Kouzarides, T. (2010). A chromodomain switch mediated by histone H3 Lys 4 acetylation regulates heterochromatin assembly. *Genes & development*, *24*(7), 647–52. doi:10.1101/gad.1881710

Xiang, H., Zhu, J., Chen, Q., Dai, F., Li, X., Li, M., Zhang, H., et al. (2010). Single base-resolution methylome of the silkworm reveals a sparse epigenomic map. *Nature Biotechnology*, *28*(5), 516–520. Retrieved from http://discovery.ucl.ac.uk/155255/

Zemach, A., McDaniel, I. E., Silva, P., & Zilberman, D. (2010). Genome-wide evolutionary analysis of eukaryotic DNA methylation. *Science (New York, N.Y.)*, *328*(5980), 916–9. doi:10.1126/science.1186366

Zhang, X., & Rossi, J. J. (2012). Phylogenetic Comparison of Small RNA-triggered Transcriptional Gene Silencing *, *286*(34), 29443–29448. doi:10.1074/jbc.R111.276378

Zhao, T., Li, G., Mi, S., Li, S., Hannon, G. J., Wang, X.-J., & Qi, Y. (2007). A complex system of small RNAs in the unicellular green alga Chlamydomonas reinhardtii. *Genes & development*, *21*(10), 1190–203. doi:10.1101/gad.1543507

Chapter II

Insights into the Role of Methylation in Diatoms by Genome-Wide Profiling in *Phaeodactylum tricornutum*

Table of contents

Chapter II **53**

Insights into the Role of Methylation in Diatoms by Genome-Wide Profiling in *Phaeodactylum tricornutum* **53**

2.1 Abstract **57**

2.2 Introduction **58**

2.3 Results **59**

2.3.1 Whole genome methylation landscape 59

2.3.2 HMRs in TEs and other repeat loci 60

2.3.3 Gene methylation profiles 65

2.3.4 Genomic distribution of methylated genes 66

2.3.5 Methylation, gene expression, and gene product function 70

2.3.6 Methylation and non-autonomous Class II TEs 73

2.4 Discussion **74**

2.5 Materials and Methods **77**

2.5.1 Culture conditions 77

2.5.2 DNA preparation, microarray hybridization, and validation 78

2.5.3 RNA-seq preparation 78

2.5.4 Identification of methylated regions, distribution and expression analysis 79

2.7 References **80**

2.8 Supporting information **84**

2.8.1 Supporting Supplementry Figures 84

2.8.2 Detailed Methods 99

2.8.3 Supplementary References 102

Insights into the Role of Methylation in Diatoms by Genome-Wide Profiling in *Phaeodactylum tricornutum*

Xin Lin[1*], Florian Maumus[1*$], Alaguraj Veluchamy[1*], Edda Rayko[1], Ikhlak Ahmed[1], Stéphane Le Crom[2], Gregory Farant[1], Jean-Yves Sgro[3], Sue A. Olson[3], Sandra Splinter Bondurant[3], Michael R. Sussman[3] , Chris Bowler[1#] and Leïla Tirichine[1#]

[1]Environmental and Evolutionary Genomics Section, Institut de Biologie de l'École Normale Supérieure (IBENS), CNRS UMR 8197 INSERM U1021, 46 rue d'Ulm 75005 Paris, France

[2] Plateforme Génomique, Institut de Biologie de l'École Normale Supérieure (IBENS), CNRS UMR 8197 INSERM U1021, 46 rue d'Ulm 75005 Paris, France

[3] Roche NimbleGen, Inc. Production Bioinformatics 500 S. Rosa Road Madison WI 53719, USA

[$] Current address: Unité de Recherche en Génomique-Info, UR 1164, INRA Centre de Versailles-Grignon, route de Saint-Cyr 78026 Versailles Cedex, France

[*] These authors contributed equally to this work

[#] Co-corresponding authors: cbowler@biologie.ens.fr and tirichin@biologie.ens.fr

2.1 Abstract

DNA cytosine methylation is a widely conserved epigenetic mark in eukaryotes that appears to play critical roles in the regulation of genome structure and transcription. Although the evolution and function of DNA methylation have been explored in a range of eukaryotes, genome-wide methylation maps have so far only been established from the supergroups Archaeplastida and Unikont. Here we report the first whole genome methylome from a stramenopile, the marine model diatom *Phaeodactylum tricornutum*. Around 6% of the genome was methylated in a mosaic landscape. We found extensive methylation in transposable elements (TEs), especially in recently amplified Copia-like elements. We also detected methylation in over 320 genes occurring in three different genomic contexts: in the proximity of TEs, in clusters of methylated genes, and in single genes, suggesting that gene methylation is mediated by at least two distinct mechanisms, including spreading from repeats. Extensive gene methylation correlated strongly with transcriptional silencing and differential expression under specific conditions. By contrast we found that genes with partial methylation tend to be constitutively expressed. These patterns contrast with those found previously in other eukaryotes, where the highest levels of methylation are typically observed in moderately expressed genes. We further observed that methylated genes are enriched in products encoding basal metabolic functions, although highly methylated gene products appear to have functions in signal transduction. Finally, it was found that genes likely acquired by horizontal gene transfer from bacteria were preferentially inserted within TE-rich regions, suggesting a mechanism whereby the expression of foreign genes can be buffered following their insertion in the genome. By going beyond plants, animals, and fungi, this stramenopile methylome adds significantly to our understanding of the evolution of DNA methylation in eukaryotes.

2.2 Introduction

DNA cytosine methylation (m5C) is a conserved epigenetic mark in eukaryotes, involved in several important biological processes such as silencing of transposable elements (TEs) and other repeat loci (1), X chromosome inactivation in female mammals (2), parent-of-origin genomic imprinting (3), and the regulation of gene expression (4). Recently, whole genome methylomes have been reported from a range of plants, fungi, and animals. These have shown that in addition to the methylation found in TEs and other repeat sequences, the presence of m5C in the bodies of genes also appears to be common in many eukaryotic genomes (5-10). In most organisms, the presence of m5C in repeat loci represents the primary mechanism of transposon suppression, whereas the function of intragenic methylation remains elusive.

In addition to revealing common aspects of eukaryotic methylation systems, these studies have also shed light on the highly variable evolution of m5C functions, patterns, and landscapes across eukaryotic groups and lineages. For example, transcriptionally silent repeat loci have been observed to be hypomethylated in invertebrates (8, 9, 11), suggesting that m5C may not be involved in TE suppression in these organisms. Conversely, genes from the early diverging vascular plant *Selaginella moellendorffii* and the moss *Physcomitrella patens* contain only trace levels of m5C compared to those found in angiosperms (8). In fungi, m5C is concentrated in repeat loci whereas active genes are not methylated (7-9). Furthermore, several model eukaryotes are devoid of DNA methylation altogether, including the yeast *Saccharomyces cerevisiae*, the nematode *Caenorhabditis elegans*, the fruitfly *Drosophila melanogaster* (expect in the early stages of embryo), and the brown alga *Ectocarpus siliculosus.*

Whole genome m5C patterns are also variable. Mammalian genomes are characterized by 'global methylation', in which m5C is found throughout the genome with the exception of CpG islands, whereas small angiosperm genomes (e.g., *Arabidopsis thaliana* and rice) and invertebrate genomes display 'mosaic methylation', in which domains of densely methylated DNA are interspersed with domains that are free of methylation (9, 12). From the methylomes examined thus far it is therefore unclear which are the ancestral underlying mechanisms at work and those that have been co-opted to distinct biological roles in different eukaryotic

groups. To address such evolutionary issues a more thorough exploration of the distribution of cytosine methylation throughout the genomes of a wider range of eukaryotes is required.

To date, all the whole genome methylomes that have been reported are from two major eukaryotic groups: Unikont and Archaeplastida. Stramenopiles, on the other hand, represent a major lineage of eukaryotes that appeared following a secondary endosymbiosis event involving a heterotrophic exosymbiont host and algal endosymbionts (13, 14). Among these, diatoms constitute a highly successful and diversified group, with possibly over 10,000 extant species. The contribution of diatoms to marine primary productivity has been estimated to be around 40% and they play a key role in the biological carbon pump as well as a major resource at the base of the food chain (14). *Phaeodactylum tricornutum* has become an attractive model diatom because of the availability of genetic tools and a fully sequenced genome (15, 16). It contains a range of genes with characterized evolutionary histories (13, 16), and in addition to genes of exosymbiont and algal endosymbiont origins, comparative analyses suggest that a significant number (> 500) of genes are most closely related to genes of bacterial origin (16). The genome also contains a diverse set of DNA methyltransferases (DNMTs) (17). *P. tricornutum* can therefore be used to probe the evolutionary history of DNA methylation, and to ask whether genes of different origins have maintained distinctive epigenetic marks. In this report we combine McrBC digestion with whole genome tiling array hybridization and RNA sequencing to address the genome-wide distribution of methylation and its potential role in genome regulation and repression of transcription in *P. tricornutum*.

2.3 Results

2.3.1 Whole genome methylation landscape

Methylated and unmethylated DNA from *P. tricornutum* were fractionated following a protocol based on the exclusion of methylated DNA by digestion with the methyl-sensitive endonuclease McrBC (18). After whole genome amplification, the samples were hybridized to a high definition tiling array of the *P. tricornutum* genome (see Materials and Methods). We found a total of 98,080 probes out of approx. 2.2 million on the array with significant enrichment probability which we further clustered into ‘highly methylated regions' (HMRs) that we arbitrarily defined as loci with at least three overlapping enriched probes. We

purposely chose this conservative cutoff in order to reduce falsely identified regions and to focus on the most significant signals in our analysis. Only these HMRs were used for further analysis. We validated our methylation mapping approach by bisulfite sequencing of 28 randomly chosen loci including genes and TEs that are distributed at different locations in the genome (**Table S2.1**). From these analyses, 5mC was found exclusively in the CpG context.

We detected 3,887 HMRs that together cover 1,412,473 bp (~5.16 %) of the 27.4 Mb *P. tricornutum* nuclear genome (**Table S2.2**). The length of HMRs ranged from 60 to 5700 nucleotides, the majority being shorter than 500 bp (**Fig. S2.1**). We used HMRs to construct a methylation map for each *P. tricornutum* genomic scaffold (**Fig**. **S2.2**). As expected, we observed extensive DNA methylation in repeat-rich regions (39% of such sequences), including in sub-telomeric regions, although no obvious centromeric regions could be detected, neither at DNA sequence level nor in terms of DNA methylation enrichment (**Fig**. **S2.2**). We also found a significant number of HMRs in repeat-free regions. A total of 587 HMRs mapped to intergenic regions, whereas 505 HMRs mapped within predicted genes (including 500 bp upstream and downstream of predicted coding sequences, CDS). In addition, 604 HMRs mapped to predicted genes that overlap with TE annotations. For further analysis, such genes (n= 766) were omitted from the regular gene set and considered as a distinct annotation class in order to circumvent a bias due to erroneous gene predictions at TE loci and to focus on the most reliable gene predictions.

2.3.2 HMRs in TEs and other repeat loci

We first characterized HMRs mapping within known autonomous TEs. The *P. tricornutum* TE complement consists principally of LTR retrotransposons (CoDis) and a few copies of DNA transposons including PiggyBac and MuDR-like elements (19). We observed heterogeneous distribution of DNA methylation across the different groups of TEs (**Fig. 2.1**). For example, while most LTR-RT annotations contain HMRs, a significant fraction of MuDR-like annotations do not (**Fig. 2.1A, B**). Furthermore, we observed that TE annotations corresponding to LTR-RT elements are extensively methylated, while those corresponding to DNA transposons are methylated to a lesser extent (**Fig. 2.1B**). In parallel, we noticed that the coverage and presence of HMRs increase linearly with the length of TE annotations (**Fig.2.2C,D**). These observations are consistent with a tighter control of young and

potentially active TE copies, especially against CoDis which have recently amplified in this genome (19), and a relaxed control of older and inactive insertions.

We next addressed the distribution of HMRs in the 766 predicted genes that overlap with TE annotations. We sorted these into two categories: genes with partial TE coverage that may correspond to genes with TE insertions, and genes with complete TE coverage which likely correspond to *bona fide* TE loci with predicted 'gene' models. As for TEs, we observed that most 'genes' with complete TE coverage contain HMRs (**Fig. S2.3**). By contrast, most of the genes with internal TE insertions do not, suggesting that they may correspond to old insertions or that TE methylation may be suppressed in the case of intragenic insertions (especially within introns; see below).

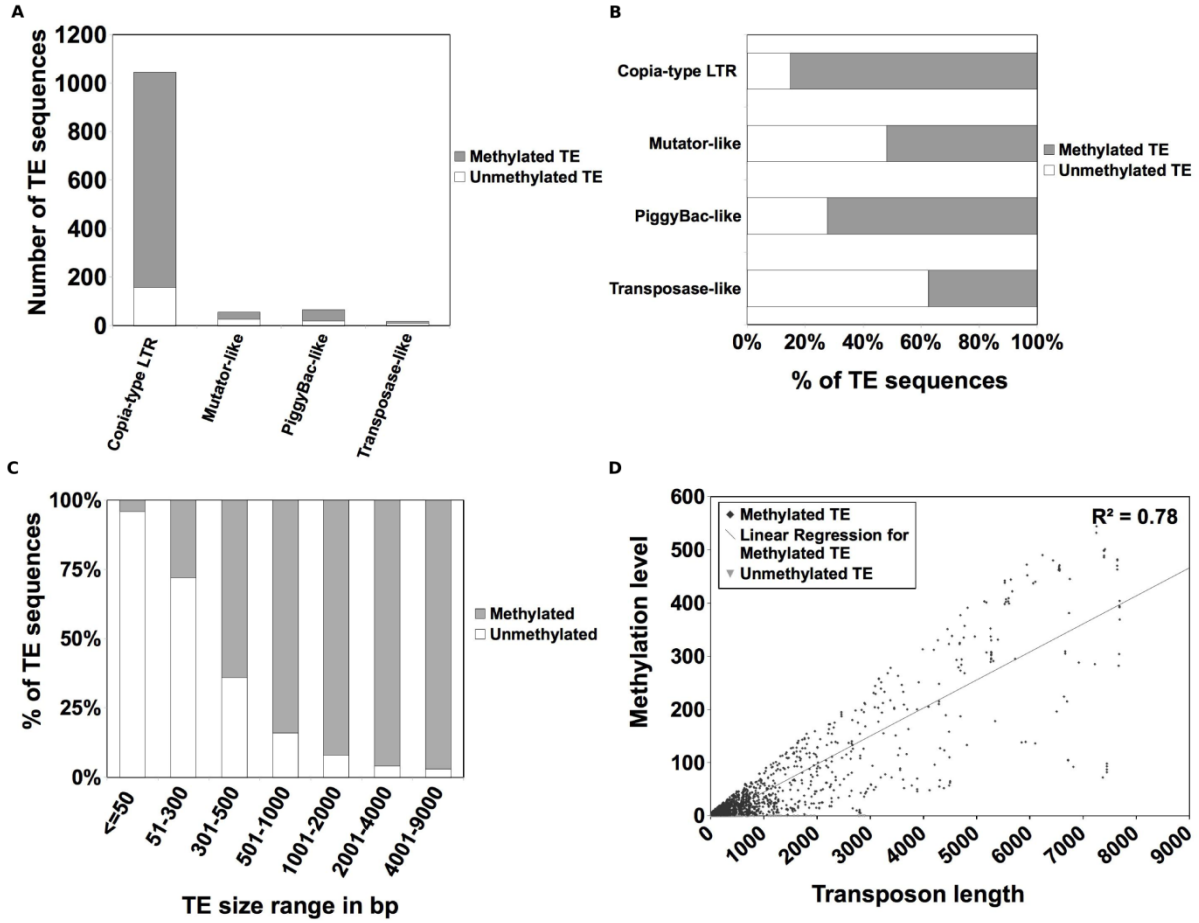


Figure 2.1 DNA methylation in transposable elements. (A-B) Number of methylated sequences and methylation coverage for different types of TEs (only for TE annotations above 300 bp). (C) Proportion of methylated sequences across size ranges of TEs and other repeats. (D) Plot of the methylation level in repeat annotations of increasing size with the correlation value. A total of 1368 TEs and repeat sequences were found to be methylated (39%).

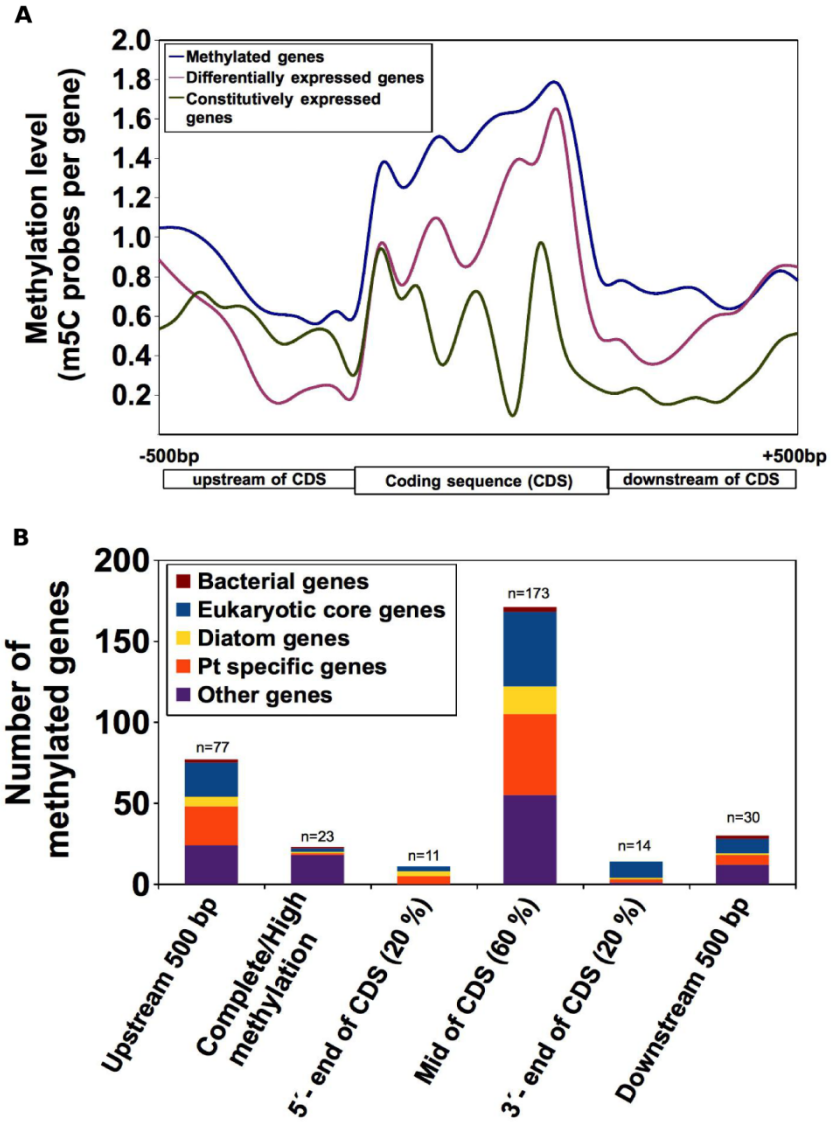


Figure 2.2 Methylation profiles on genes. (A) Methylation pattern on genes. DNA methylation pattern is shown along gene bodies and 500 bp gene-flanking regions. A moving window of 50 bp along the gene was used to calculate the average number of m5C probes (y-axis). Gene methylation profiles with respect to their expression are also shown. Constitutively expressed genes display low methylation levels along their CDS, whereas differentially expressed genes have an overall higher and increasing methylation level from 5' to 3' end of CDS. (B) Gene methylation distribution among different categories of genes. *P. tricornutum* genes have been previously grouped into different classes, including those that are present only in this diatom (Pt specific), those present in both sequenced diatoms but in no other eukaryote (diatom specific), those present in all eukaryotes examined (Eukaryotic core

genes), and those proposed to be acquired from bacteria by horizontal gene transfer (bacterial genes) (for further information about gene categories, see (20)). By examining the content of methylated genes in these different categories, we found an equal distribution among Pt specific, Eukaryotic core genes, and the rest of all genes. However, only a few methylated genes are of bacterial origin. Methylated genes are also grouped by HMR localisation. A preference for gene-body methylation, across all four classes of genes, is apparent. In general, the proportion of methylated genes in each class corresponds to the gene content of that class in the genome.

2.3.3 Gene methylation profiles

The 505 HMRs mapping within genes were distributed in the bodies and flanking 500 bp sequences of 326 genes. As found in most eukaryotes examined to date, methylation in the body of *P. tricornutum* genes occurs almost exclusively in exons (**Table S2.3**). Interestingly, we found that although most *P. tricornutum* genes contain 1-2 exons, methylated genes tend to contain more exons than unmethylated genes (**Fig**. **S2.4**). Additionally, although highly methylated genes are typically single exon genes, the few *P. tricornutum* genes with five or more exons show higher methylation levels than those with 3-4 exons (**Fig**. **S2.4B**). Furthermore, the size distribution of methylated genes is skewed towards longer genes as compared to unmethylated genes, i.e., the frequency of methylated genes longer than 2 kb is higher than that of unmethylated genes (Welsch two sample t-test p-value = 3.612e-06) (**Fig**. **S2.4C**). Overall then, methylated genes in *P. tricornutum* tend to be longer and to contain more exons than unmethylated genes.

Ends analysis show that, on average, methylation levels in genes increase from 5' to 3' within the CDS, with sharp reductions at the ends (**Fig**. **2.2A**). Such a pattern is most similar to that observed in the bodies of Arabidopsis, mammalian and fish genes, and is in contrast to what has been described in the genomes of most invertebrates where methylation is found predominantly within the first half of the CDS (6-8).

We further distinguished several patterns of DNA methylation in genes: extensive methylation from upstream to downstream of the CDS, as well as partial methylation, which we sub-divided into categories following the position of the methylation peak: upstream 500 bp, 5' end of CDS (relative first 20% of CDS), middle of CDS, 3' end of the CDS (relative last 20% of CDS), and downstream 500 bp. We found that intragenic methylation occurs mainly in the mid-CDS region while relatively few methylation profiles peak in the 5' or 3' ends of CDS (**Table S2.3**, **Fig**. **2.2B**). We also noticed a substantial number of genes with the highest levels of methylation in their promoters. Besides these, we detected 25 genes with extensive methylation throughout.

Methylation appears to be distributed evenly among genes belonging to different orthology groups, previously defined by (16, 20) as being present either in all eukaryotes, as being *P.*

tricornutum-specific or diatom-specific, or predicted to have been acquired from bacteria by horizontal gene transfer (**Fig**. **2.2B**, **Fig**. **S2.5**). However, the most strongly methylated genes appear to be depleted in *P. tricornutum*-specific genes and eukaryotic core genes, and rather to be enriched in other genes with unclear phylogenetic affiliations (**Fig**. **2.2B**).

2.3.4 Genomic distribution of methylated genes

In order to analyze the chromosomal distribution of body-methylated genes, they were mapped onto the *P. tricornutum* scaffolds and their positions were compared to TE annotations. We observed that body-methylated genes are found in different genomic contexts. First, we found that many methylated genes are located in the vicinity of TEs (e.g., **Fig**. **2.3A**). A more detailed analysis indeed revealed that methylated genes tend to map more closely to TEs than non-methylated genes (**Fig**. **2.4**). Considering that TEs are extensively methylated in the genome, we postulate that DNA methylation in such genes may result by spreading from TEs, as reported in *A. thaliana* (21). This suggests that TE insertion followed by DNA methylation may impact the epigenetic status and expression levels of flanking genes (see below). However not all TE-flanking genes are methylated (**Fig**. **2.3B,C**), suggesting that spreading, or its avoidance, is a selective process. In repeat-free regions, we found that methylated genes are isolated (**Fig**. **2.3D**) or juxtaposed to one another in clusters comprising 2-3 genes (e.g., **Fig**. **2.3E**).

We next analyzed the distribution of TEs in the vicinity of genes previously defined as being present in all eukaryotes, or predicted to have been acquired from bacteria by horizontal gene transfer (16, 20). Interestingly, we noticed an increased presence of TEs upstream of putative bacterial genes (**Fig**. **2.4**), represented by gene IDs 46924 and 47160 in **Fig**s. **2.3B** and **C**, respectively. In spite of this, the proportion of methylated bacterial genes in the genome was less than in other gene categories (**Fig**. **2.3B** and **Fig**. **S2.5**).

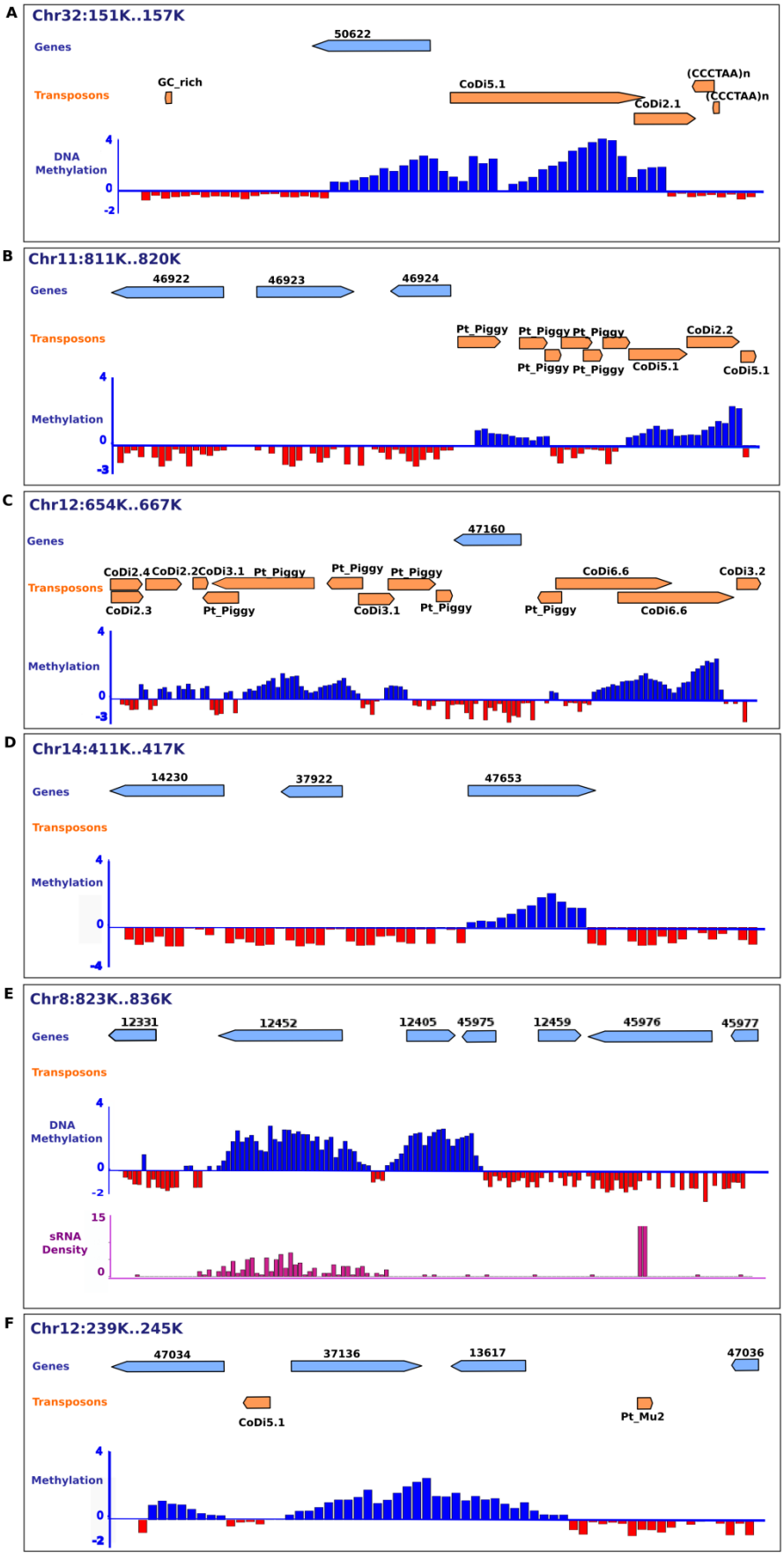

A
Chr32:151K..157K
Genes
50622
Transposons
GC_rich
CoDi5.1
(CCCTAA)n
CoDi2.1
(CCCTAA)n
DNA Methylation
4
0
-2
B
Chr11:811K..820K
Genes
46922
46923
46924
Transposons
Pt_Piggy
Pt_Piggy
Pt_Piggy
CoDi2.2
Pt_Piggy
Pt_Piggy
CoDi5.1
CoDi5.1
Methylation
4
0
-3
C
Chr12:654K..667K
Genes
47160
Transposons
CoDi2.4
CoDi2.2
CoDi3.1
Pt_Piggy
Pt_Piggy
Pt_Piggy
CoDi6.6
CoDi3.2
CoDi2.3
Pt_Piggy
CoDi3.1
Pt_Piggy
Pt_Piggy
CoDi6.6
Methylation
4
0
-3
D
Chr14:411K..417K
Genes
14230
37922
47653
Transposons
Methylation
4
0
-4
E
Chr8:823K..836K
Genes
12331
12452
12405
45975
12459
45976
45977
Transposons
DNA Methylation
4
0
-2
sRNA Density
15
0
F
Chr12:239K..245K
Genes
47034
37136
13617
47036
Transposons
CoDi5.1
Pt_Mu2
Methylation
4
0
-2

Figure 2.3 Methylation patterns of selected genes. Three tracks (genes, transposons, DNA methylation) along with chromosome position are shown for each example. (A) Region on chromosome 32 containing a methylated gene bordering a cluster of methylated TEs. (B) Region on chromosome 11 containing methylated TEs with a cluster of nonmethylated genes. Gene 46924 is derived from bacteria. (C) Region on chromosome 12 containing a nonmethylated bacterial gene (Id 47160) surrounded by methylated TEs. (D) Example of highly methylated gene. Under normal conditions, gene 47653 (annotated as a neurotransmitter transporter) is methylated with no expression in Pt1 accession but expressed highly in the Pt9 tropical accession. (E) Example of highly methylated gene cluster. Region on chromosome 8, isolated from methylated TEs, containing a cluster of methylated genes. Note that gene 12452, encoding a P-type ATPase, is also targeted by small RNAs (data from (36)). (F) Example of highly methylated gene displaying strong differential expression. Under normal conditions, gene 13617 (encoding a serine/threonine protein kinase) is methylated with no expression but expressed specifically under silicate deplete conditions (20). Heights of the peak represent the normalized log ratio (score) of the m5C probes. Genes and TE annotations are indicated.

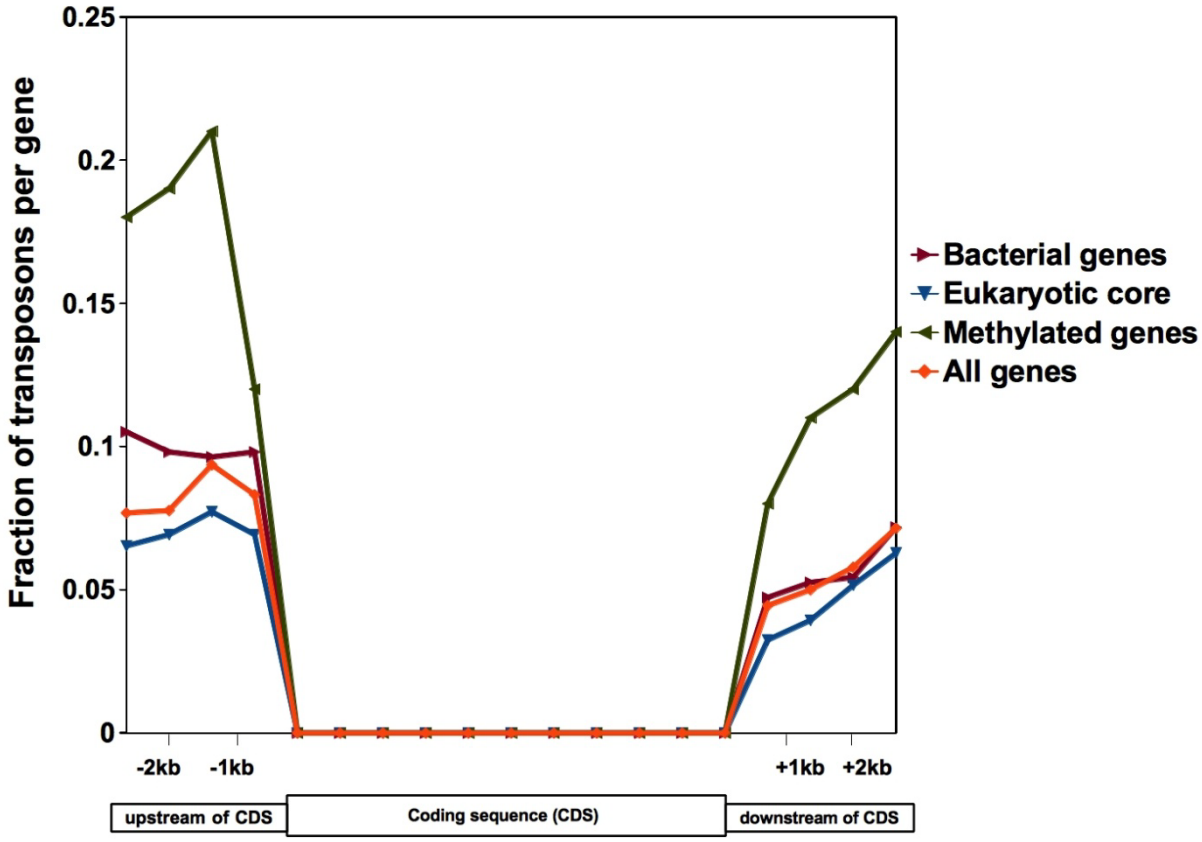


Figure 2.4 Distribution of transposable elements around genes. The plot shows TE distribution within 2 kb upstream and downstream of all genes (n=10,408), bacterial genes (n=571), methylated genes (n=326), and eukaryotic genes (n=2775). Clusters of TEs within the 2kb region upstream of bacterial genes were more common than the other classes.

2.3.5 Methylation, gene expression, and gene product function

To assess whether gene methylation impacts transcriptional regulation in *P. tricornutum*, we compared the expression levels of methylated versus unmethylated genes. Expression levels were quantified using RNA-seq data obtained from *P. tricornutum* cells grown under normal conditions and normalized to CDS size and library size (fragments per kilobase of exon per million fragments mapped, FPKM). We analyzed separately the genes with different HMR peak locations and genes with extensive HMR coverage. Interestingly, while genes with partial intragenic methylation displayed expression levels similar to unmethylated genes, those with extensive HMR coverage show on average a markedly lower FPKM (**Fig**. **2.5A**). In analogy with *Ascobulus immersus* (22), such a negative correlation between the extent of methylation and gene expression levels suggests a suppressive role for extensive gene methylation. We also observed that methylation in the promoter (upstream 500 bp) of genes does not appear to impact transcription levels.

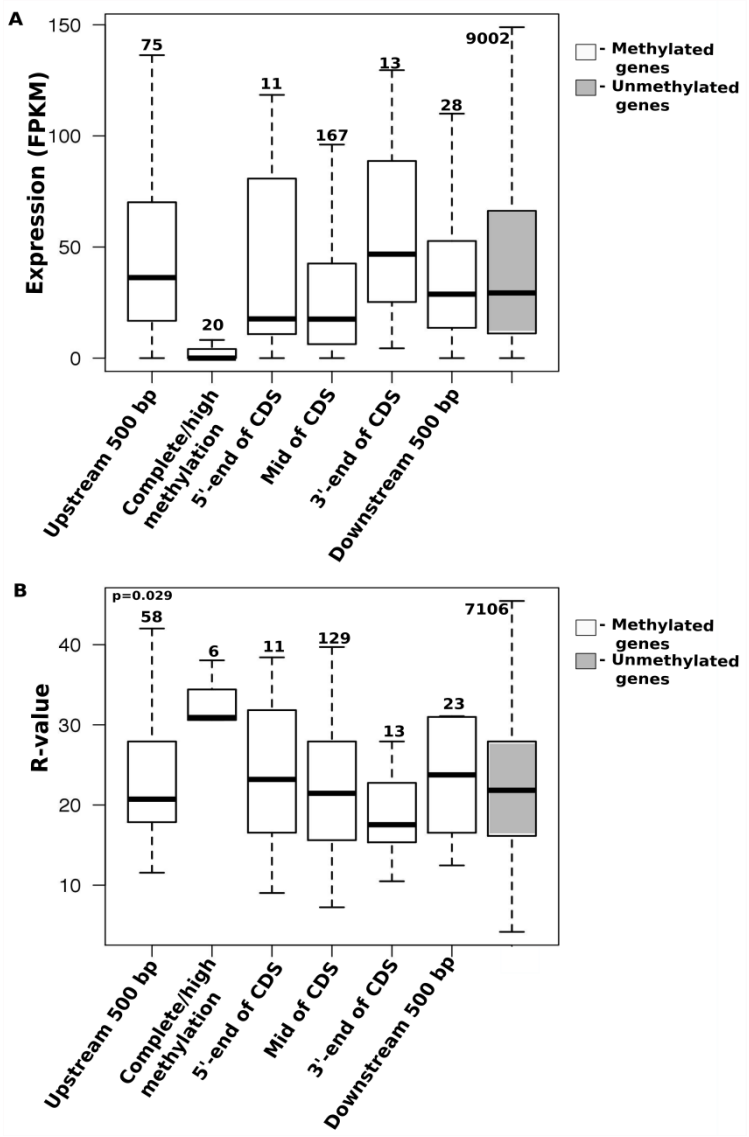


Figure 2.5 Expression profiles of methylated genes. (A) Expression levels of methylated genes. Gene expression was quantified in standard growth conditions using RNA-seq data. About 85 % of genes are expressed and quantified as fragments per kilobase of exons per million reads mapped (FPKM). (B) Differential expression profiles of methylated genes. Boxplots show the ranges of R-values for each category of methylated genes. Genes with R-values below 12 are considered to be constitutively expressed (20). Genes with dense HMRs are revealed as being differentially expressed (p-value of 0.029). Numbers above each column show the number of genes in each category.

We had previously estimated the degree of differential expression of *P. tricornutum* genes across 16 cDNA libraries by calculating the statistical significance of differential mRNA levels in specific conditions compared to random distribution (log likelihood ratio, R-value) (20, 23). Considering these criteria, constitutively expressed genes have low R-values while genes that are significantly over-represented in specific growth conditions have high R-values. We examined whether we could detect a correlation between gene methylation and R-value. Interestingly, we found that genes with extensive HMR coverage have significantly increased R-values compared to genes with partial methylation and unmethylated genes (**Fig**. **2.5B**). These observations suggest that the transcription of genes with extensive methylation is under tight control, i.e., that this gene population tends to be silenced and/or expressed only under specific conditions. For instance, gene model 13617, which encodes a serine/threonine protein kinase belonging to the eukaryotic core genes, shows null FPKM under normal conditions but is specifically represented in the cDNA library prepared from *P. tricornutum* cells grown under silicate deplete conditions (**Fig**. **2.3F**). In contrast, genes showing partial methylation are more likely to be expressed constitutively. More specifically, body methylated genes appear to be expressed at relatively low to moderate levels (**Fig**. **S2.6**). We also addressed whether we could detect a link between expression profiles and orthology groups but we found no significant correlation other than that methylated eukaryotic core genes were more likely to have lower R values than the other categories (**Table S2.4**). Additionally, methylated *P. tricornutum*-specific genes seem to be more tightly regulated as shown by their higher R values, suggesting their potential importance in regulating gene expression under specific stress conditions (**Table S2.4**).

To assess the function of the 326 methylated gene products, we performed a gene ontology (GO) analysis. Overall we found that the methylated gene set is enriched in GO categories such as ‘transferase activity,’ ‘transporter activity,’ ‘carbohydrate binding,’ and ‘nutrient reservoir activity’ (**Fig**. **S2.7A**). However, when comparing GO enrichment between sets of genes with different methylation profiles, we observed that, in contrast, the subset of genes with extensive methylation is enriched in ‘protein kinase’ and ‘signal transducer activity’ categories which are evocative of regulatory and signaling functions (**Fig**. **S2.7B**). Interestingly, when looking in more detail into metabolic pathways, we found that methylated genes are especially represented among genes predicted to be involved in pentose phosphate

metabolism (**Fig**. **S2.8A**). This pathway was reported previously to have unusual features in diatoms and to likely not be subject to regulation by thioredoxin, as is the case in other photosynthetic organisms (24). DNA methylation may therefore play an important role in the regulation of glucose turnover that produces NADPH and pentoses as essential backbones of nucleotides. When focusing on regulatory pathways, we found that methylated genes represent a significant number of *P. tricornutum* genes predicted to be involved in DNA mismatch repair, suggesting a role for DNA methylation in the maintenance of DNA integrity (**Fig**. **S2.8B**).

2.3.6 Methylation and non-autonomous Class II TEs

In a previous study we screened for autonomous TEs in the *P. tricornutum* genome using similarity-based approaches and searching for sequence structural characteristics specific of TEs. Here, in order to improve the quality of HMR mapping, we used the *de novo* repeat identification program Recon (25) and the tandem repeat finder program TRF (26) in an attempt to detect and annotate potentially unclassified or simple repeats in the genome. Most of the newly identified repeat loci lack HMRs (**Fig**. **S2.3B**). More specifically, although we were not able to classify most of the unknown repeats detected by Recon, we identified two families of non-autonomous Class II TEs with captured exons and analyzed their m5C patterns individually.

A first family, called R33, consists of six copies whose extremities are highly similar to the terminal inverted repeats (TIRs) of an inactivated MuDR-like element, and whose internal sequence contains an exon from the single-copy gene encoding 2-oxoglutarate dehydrogenase component E1 (**Fig**. **S2.9A**). We found that none of the R33 copies nor the original gene overlap with HMRs, suggesting that R33 repeats are not targeted by the DNA methylation machinery, which is consistent with the inactive state of the cognate autonomous element.

The second non-autonomous Class II family, called R59, comprises four copies and appears to be linked to PiggyBac elements (**Fig**. **S2.9B**). R59 has captured a fragment of exon from a single copy gene encoding Heat Shock Protein 70 (*HSP70*). Interestingly, although PiggyBac elements in general were observed to be only moderately methylated, R59 displays much higher methylation levels. Even more unexpectedly, the original *HSP70* gene (gene model

41417) also appears to be methylated, with HMR coverage extending out of the region of similarity with the captured region found in R59 (**Fig**. **S2.9C**). This suggests that the presence of R59 copies in the genome may affect the epigenetic regulation of the *HSP70* gene.

2.4 Discussion

In the work described herein, we have obtained the first genome-wide DNA methylation map of the nuclear genome of a stramenopile, namely the marine diatom *P. tricornutum*. Overall, DNA methylation in *P. tricornutum* is low, as previously reported using reversed-phase high performance liquid chromatography (27). It shows ~5% of global methylation with only 3.3% of genes methylated. This is lower than what is seen in mammals and in plants such as Arabidopsis and rice, in which over 30% of genes are methylated (5-8), but is similar to the marine tunicate *Ciona intestinalis* and the early-diverging land plant *Selaginella moellendorffii* (16). Consistent with previous studies (19), we also found a significant enrichment of DNA methylation in TEs. Furthermore, we detected scarce DNA methylation in the intergenic space. The DNA methylation landscape of *P. tricornutum* is therefore reminiscent of the 'mosaic' landscapes observed in angiosperms with small genomes and invertebrates (12), with islands of HMRs surrounded by methylation-free regions.

This first methylome from a stramenopile confirms the evolutionary conservation of gene-body methylation among eukaryotes (7, 8). Gene-body methylation was found to occur in various (epi)genomic contexts: in close proximity to TEs, in clusters of methylated genes, and in single genes. In the case of methylated genes that are flanked by repeats, we assume that methylation occurs through spreading from repeats. This indicates that, as seen in *A. thaliana* (21), insertion of TEs can trigger the formation of heterochromatin around and within flanking genes. By contrast, methylated genes in repeat-free regions are likely to be methylated following a distinct and more specific mechanism. Methylated genes organized in clusters are evocative of coordinated transcriptional regulation, and examples were indeed found of methylated gene clusters whose genes displayed similar expression profiles (**Fig**. **2.3F**). In all contexts, gene-body methylation was found almost exclusively in exons, which is the case for most organisms investigated so far (7-9).

The functional annotation of body-methylated genes revealed that many encoded important metabolic activities, such as transferases, transporters, carbohydrate-binding proteins, and other components involved in nutrient reservoir maintenance (**Fig**. **S2.7A**). In contrast with such apparently housekeeping functions, genes that are extensively methylated tend to encode signaling components (**Fig**. **S2.7B**). Furthermore, such genes tend to be silent under most conditions and differentially expressed only under specific conditions (**Fig**. **2.5B**). It will be of interest to examine whether methylation plays a role in the suppression of transcription of these genes, and whether it can be reversible as a function of gene expression. We have observed previously that induction of the *Blackbeard* LTR-RT element in response to nitrate limitation is accompanied by loss of methylation (19). It is therefore possible that in diatoms the perception of changing environments can trigger the hypomethylation of specific genes and release their transcriptional suppression.

A significant proportion (ca. 30%) of the approx. 10,000 genes in *P. tricornutum* has been assigned a putative evolutionary origin, either from the ancestral exosymbiont, from one of the two algal endosymbionts, or by horizontal gene transfer from bacteria (13, 16). A further 25% are either specific to *P. tricornutum* or are specific to diatoms (16, 20). Such information provides an opportunity to examine whether distinct gene methylation patterns have been conserved during diatom evolution since they were acquired. We were unable to detect any such signature, although we did note that all such categories were under-represented in the gene set that was extensively methylated (**Fig**. **2.2B**). We further observed that bacterial genes tended not to be methylated, when compared with other gene classes (**Fig**. **S2.5**). Notwithstanding, bacterial genes are often associated with TE-rich regions (**Fig**. **2.4**). This may suggest that horizontal gene transfer of environmental DNA may be facilitated by transposable elements in a mechanism such as retroposition (28) (29), or perhaps more likely that the insertion of such extraneous genes in repeat-rich regions may provide a probationary period in which their expression is attenuated and only released from repression gradually in case their effects may be lethal.

Eukaryotes have evolved and/or retained different DNA methyltransferase complements. Metazoans commonly encode DNMT1 and DNMT3 proteins, while higher plants additionally have plant-specific chromomethylase (CMT), and fungi have DNMT1, Dim-2, DNMT4, and

DNMT5 (30, 31). The *P. tricornutum* genome encodes a peculiar set of DNMTs as compared to other eukaryotes (17). DNMT1 appears to be absent, and in addition to DNMT3, diatom genomes also encode a DNMT5 protein as well as a bacterial-like DNMT. Because DNMT5 is also found in other algae and fungi, we postulate that it was present in a common ancestor. Furthermore, structural, functional, and phylogenetic data suggest that CMT, Dim-2 and DNMT1 are monophyletic (30, 31). Therefore, we propose that the common ancestor of plants, unikonts and stramenopiles possessed DNMT1 (subsequently lost in diatoms), DNMT3, and probably also DNMT5 (lost in metazoans and higher plants). This evolutionarily important loss is supported by absence of DNA methyltransferases in brown algal species *E.siliculosus*.In bacteria, cytosine methylation acts in the restriction-modification system. Thus, the function of a bacterial-like DNMT in *P. tricornutum* is unclear. Interestingly, it is conserved in the centric diatom *Thalassiosira pseudonana,* from which pennate diatoms such as *P. tricornutum* diverged ~90 million years ago. This implies that a diatom common ancestor acquired DNMT from bacteria after a horizontal gene transfer prior to the centric/pennate diatom split (14). Conservation of this gene in diatoms over this length of time suggests that it is functional. It will therefore be of interest to uncover the roles of the different DNMTs present in *P. tricornutum* in processes such as maintenance and *de novo* DNA methylation as well as context specificities. Until now, experimental validation by bisulfite sequencing indicates a clear CpG context preference in diatoms.

P. tricornutum possesses an active small RNA-mediated silencing machinery (32). This suggests that dsRNAs are efficiently processed into small RNAs in *P. tricornutum* and that they are capable of guiding DNA methylation in an RNA-dependent DNA methylation (RdDM) fashion (33-35). Furthermore, Huang and collaborators recently reported the presence of small RNAs in *P. tricornutum* (36). Interestingly, more than half of the highly methylated genes that are differentially expressed are targeted by small RNAs (e.g., **Fig. 2.3E**), suggesting that RdDM may play a role in the regulation of transcription of a subset of genes in diatoms, as recently inferred from studies of the atypical DNA methylation on some genes in *A. thaliana* (37). Furthermore, we observed that the captured exon found in the R59 repeat is inserted in reverse orientation with respect to the PiggyBac backbone (**Fig. S2.9**). A scenario explaining the methylation found in the R59 repeat and the HSP70 gene could be that R59 is a source of transcripts with complementarity to HSP70 transcripts. The formation of

double-stranded RNA duplexes may trigger their processing into small RNA that would target both R59 and HSP70 loci and methylate them through RdDM. The cognate *HSP70* gene was indeed found to be targeted by small RNAs. RdDM may therefore represent an important mechanism of genome regulation in diatoms.

In conclusion, the present work brings substantial information about the *P. tricornutum* methylome that enables analysis of methylation patterns and landscapes beyond animals, plants, and fungi. *P. tricornutum* is of significant interest for such studies because it can be readily manipulated by reverse genetics. Unlike the other unicellular model organisms *Saccharomyces cerevisiae*, *Schizosaccharomyces pombe* and *Chlamydomonas reinhardtii*, it has a small compact genome that displays all the key features of more complex genomes, such as DNA methylation, RNAi, and histone modifications. Furthermore, *P. tricornutum* has the peculiarity to be pleiomorphic as it can be found in the form of four different morphotypes: fusiform, oval, round, and triradiate (38). Significantly, morphotype transition occurs in response to specific environmental conditions. Therefore, *P. tricornutum* also constitutes an excellent model to study the basis of epigenomic reprogramming events that lead to morphological variations in response to external stimuli, e.g., to assess the influence on adaptive evolutionary processes of the increased susceptibility of methylated genes to mutation. We therefore hope that our work on DNA methylation and its role in gene regulation in the diatom *P. tricornutum* will be the foundation for future work, and an exciting opportunity for comparative epigenomics and the elucidation of the dynamics of genome evolution in relation to the epigenetic regulation of gene expression.

2.5 Materials and Methods

2.5.1 Culture conditions

Cultures of *P. tricornutum* Bohlin clone Pt1 8.6 (CCMP2561) were grown in f/2 medium made with 0.2-µm-filtered and autoclaved local seawater supplemented with f/2 vitamins and inorganic filter sterilized nutrients. Cultures were incubated at 19°C under cool white fluorescent lights at approximately 75 µmol.m-2.s-1 in 12h light: 12h dark conditions and maintained in exponential phase in semicontinuous batch cultures.

2.5.2 DNA preparation, microarray hybridization, and validation

To optimize the reproducibility and efficiency of methylated DNA exclusion, we modified the original protocol in a method called 'Window McrBC Restriction' (WMR; see Supplementary Methods). Genomic DNA from three *P. tricornutum* cultures (biological replicates) was sonicated, size fractionated, and incubated with McrBC enzyme (New England Biolabs). In negative controls, GTP, which is the co-factor required for McrBC activity, was replaced by water. Prior to hybridization, the DNA was further size selected as 500-700 nt fragments. Microarray hybridization was performed according to Lippman et al. (18), following NimbleGen's "NimbleChip Arrays User's Guide: DNA Methylation Analysis v2.0" (Roche NimbleGen, Germany). NimbleGen 2.1M *Phaedoctylum tricornutum* tiling arrays were designed based on the JGI Phatr2 genome. A total of 2.1 million probes are represented on this array, which represents the entire + strand of the nuclear genome at 12 nt overlapping intervals. The average probe length was 56 nt. Both chloroplastic and mitochondrial genomes, representing respectively 117 kb and 44 kb, were excluded from the array. NimbleGen provided design and probe annotation.

We validated our methylation mapping approach by bisulfite sequencing of 28 randomly chosen loci including genes and TEs that are distributed at different locations in the genome (**Table S2.1**). We included in the validation procedure one, two and three sparsely enriched probes to confirm that they had not been falsely discarded as a result of our strict filtering process. Altogether, our results enabled the validation of the mapping approach used for the identification of methylated regions.

2.5.3 RNA-seq preparation

P. tricornutum clone Pt1 8.6 cells were harvested at exponential phase and total RNA was used for first strand cDNA synthesis followed by double strand cDNA using Mint Universal Kit from Evrogen (SK002). cDNA was used for non-directional cloning and cDNA library construction for Illumina sequencing by Beckman Coulter Genomics. Sequencing was performed with a read length of 75 bp and sequencing coverage of 1.5 Gb.

2.5.4 Identification of methylated regions, distribution and expression analysis

Statistically significant probe bound regions (ChIP-enriched genomic regions) were detected using the RINGO (39) in R Bioconductor package. A cut-off P-value of 0.02 was set for normalization. Boundaries for methylated regions were defined as those with a minimum of three enriched overlapping enriched probes, using a moving window of 50 bp. We found at least 98,080 enriched probes, which amounts to 4.5 % of probes covered. Normalization on the three biological replicates yielded robust consistency which was statistically validated using a Student t-test and showed a Pearson R-value between 0.92 – 0.93 (**Fig**. **S2.10**, **S2.11**). Expression correlations with methylation were done using R-values derived from the EST sequences (20) and cDNA sequence data. Data processing, analysis, and plotting were done using Python, R/Bioconductor, and CIRCOS (40) (see Supplementary Methods). A genome browser based on Gbrowse is available to explore this methylome data (http://ptepi.biologie.ens.fr/cgi-bin/gbrowse/Pt_Epigenome).

2.7 References

1. Kato M, Miura A, Bender J, Jacobsen SE, & Kakutani T (2003) Role of CG and non-CG methylation in immobilization of transposons in Arabidopsis. *Curr Biol* 13(5):421-426.

2. Bird AP (1986) CpG-rich islands and the function of DNA methylation. *Nature* 321(6067):209-213.

3. Feil R & Berger F (2007) Convergent evolution of genomic imprinting in plants and mammals. *Trends Genet* 23(4):192-199.

4. Zilberman D, Gehring M, Tran RK, Ballinger T, & Henikoff S (2007) Genome-wide analysis of Arabidopsis thaliana DNA methylation uncovers an interdependence between methylation and transcription. *Nat Genet* 39(1):61-69.

5. Lister R*, et al.* (2009) Human DNA methylomes at base resolution show widespread epigenomic differences. *Nature* 462(7271):315-322.

6. Cokus SJ*, et al.* (2008) Shotgun bisulphite sequencing of the Arabidopsis genome reveals DNA methylation patterning. *Nature* 452(7184):215-219.

7. Feng S*, et al.* (2010) Conservation and divergence of methylation patterning in plants and animals. *Proc Natl Acad Sci U S A* 107(19):8689-8694.

8. Zemach A, McDaniel IE, Silva P, & Zilberman D (2010) Genome-wide evolutionary analysis of eukaryotic DNA methylation. *Science* 328(5980):916-919.

9. Xiang H*, et al.* (2010) Single base-resolution methylome of the silkworm reveals a sparse epigenomic map. *Nat Biotechnol* 28(5):516-520.

10. Lyko F*, et al.* (2010) The honey bee epigenomes: differential methylation of brain DNA in queens and workers. *PLoS Biol* 8(11):e1000506.

11. Su Z, Han L, & Zhao Z (2010) Conservation and divergence of DNA methylation in eukaryotes: new insights from single base-resolution DNA methylomes. *Epigenetics* 6(2):134-140.

12. Suzuki MM & Bird A (2008) DNA methylation landscapes: provocative insights from epigenomics. *Nat Rev Genet* 9(6):465-476.

13. Moustafa A*, et al.* (2009) Genomic footprints of a cryptic plastid endosymbiosis in diatoms. *Science* 324(5935):1724-1726.

14. Bowler C, Vardi A, & Allen AE (2010) Oceanographic and biogeochemical insights from diatom genomes. *Ann Rev Mar Sci* 2:333-365.

15. Bowler C, De Martino A, & Falciatore A (2010) Diatom cell division in an environmental context. *Curr Opin Plant Biol* 13(6):623-630.

16. Bowler C*, et al.* (2008) The Phaeodactylum genome reveals the evolutionary history of diatom genomes. *Nature* 456(7219):239-244.

17. Maumus F, Rabinowicz P, Bowler C, & Rivarola M (2011) Stemming Epigenetics in Marine Stramenopiles. *Current Genomics* 12(5):357-370.

18. Lippman Z, Gendrel AV, Colot V, & Martienssen R (2005) Profiling DNA methylation patterns using genomic tiling microarrays. *Nat Methods* 2(3):219-224.

19. Maumus F*, et al.* (2009) Potential impact of stress activated retrotransposons on genome evolution in a marine diatom. *BMC Genomics* 10:624.

20. Maheswari U*, et al.* (2010) Digital expression profiling of novel diatom transcripts provides insight into their biological functions. *Genome Biol* 11(8):R85.

21. Ahmed I, Sarazin A, Bowler C, Colot V, & Quesneville H (2011) Genome-wide evidence for local DNA methylation spreading from small RNA-targeted sequences in Arabidopsis. *Nucleic Acids Res* 39(16):6919-6931.

22. Barry C, Faugeron G, & Rossignol JL (1993) Methylation induced premeiotically in Ascobolus: coextension with DNA repeat lengths and effect on transcript elongation. *Proc Natl Acad Sci U S A* 90(10):4557-4561.

23. Maheswari U, Mock T, Armbrust EV, & Bowler C (2009) Update of the Diatom EST Database: a new tool for digital transcriptomics. *Nucleic Acids Res* 37(Database issue):D1001-1005.

24. Kroth PG*, et al.* (2008) A model for carbohydrate metabolism in the diatom Phaeodactylum tricornutum deduced from comparative whole genome analysis. *PLoS One* 3(1):e1426.

25. Bao Z & Eddy SR (2002) Automated de novo identification of repeat sequence families in sequenced genomes. *Genome Res* 12(8):1269-1276.

26. Benson G (1999) Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res* 27(2):573-580.

27. Jarvis EE, Dunahay TG, & Brown LM (1992) DNA NUCLEOSIDE COMPOSITION AND METHYLATION IN SEVERAL SPECIES OF MICROALGAE. *J. Phycol.* 28(3):356-362.

28. Paul JH, Jeffrey WH, & DeFlaun MF (1987) Dynamics of extracellular DNA in the marine environment. *Appl Environ Microbiol* 53(1):170-179.

29. Wang W*, et al.* (2006) High rate of chimeric gene origination by retroposition in plant genomes. *Plant Cell* 18(8):1791-1802.

30. Ponger L & Li WH (2005) Evolutionary diversification of DNA methyltransferases in eukaryotic genomes. *Mol Biol Evol* 22(4):1119-1128.

31. Goll MG*, et al.* (2006) Methylation of tRNAAsp by the DNA methyltransferase homolog Dnmt2. *Science* 311(5759):395-398.

32. De Riso V*, et al.* (2009) Gene silencing in the marine diatom Phaeodactylum tricornutum. *Nucleic Acids Res* 37(14):e96.

33. Wassenegger M, Heimes S, Riedel L, & Sanger HL (1994) RNA-directed de novo methylation of genomic sequences in plants. *Cell* 76(3):567-576.

34. Teixeira FK*, et al.* (2009) A role for RNAi in the selective correction of DNA methylation defects. *Science* 323(5921):1600-1604.

35. Mette MF, Aufsatz W, van der Winden J, Matzke MA, & Matzke AJ (2000) Transcriptional silencing and promoter methylation triggered by double-stranded RNA. *EMBO J* 19(19):5194-5201.

36. Huang A, He L, & Wang G (2011) Identification and characterization of microRNAs from Phaeodactylum tricornutum by high-throughput sequencing and bioinformatics analysis. *BMC Genomics* 12:337.

37. You W*, et al.* (2012) Atypical DNA methylation of genes encoding cysteine-rich peptides in Arabidopsis thaliana. *BMC Plant Biol* 12(1):51.

38. De Martino A*, et al.* (2011) Physiological and Molecular Evidence that Environmental Changes Elicit Morphological Interconversion in the Model Diatom Phaeodactylum tricornutum. *Protist* 162(3):462-481.

39. Toedling J*, et al.* (2007) Ringo--an R/Bioconductor package for analyzing ChIP-chip readouts. *BMC Bioinformatics* 8:221.

40. Krzywinski M*, et al.* (2009) Circos: an information aesthetic for comparative genomics. *Genome Res* 19(9):1639-1645.

2.8 Supporting information

2.8.1 Supporting Supplementry Figures

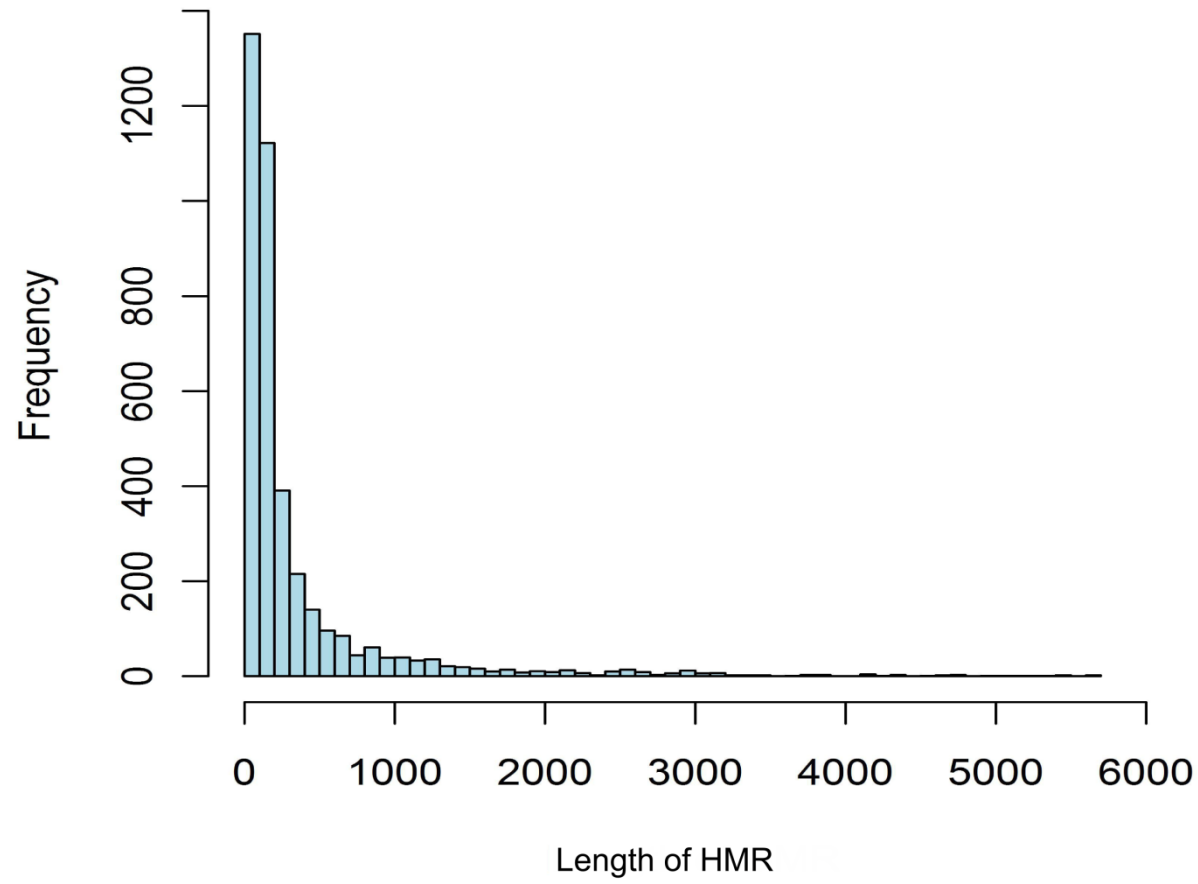


Figure S2.1 Distribution of peak lengths (in bp) of densely methylated regions (HMR). The majority of HMRs are less than 500 bp.

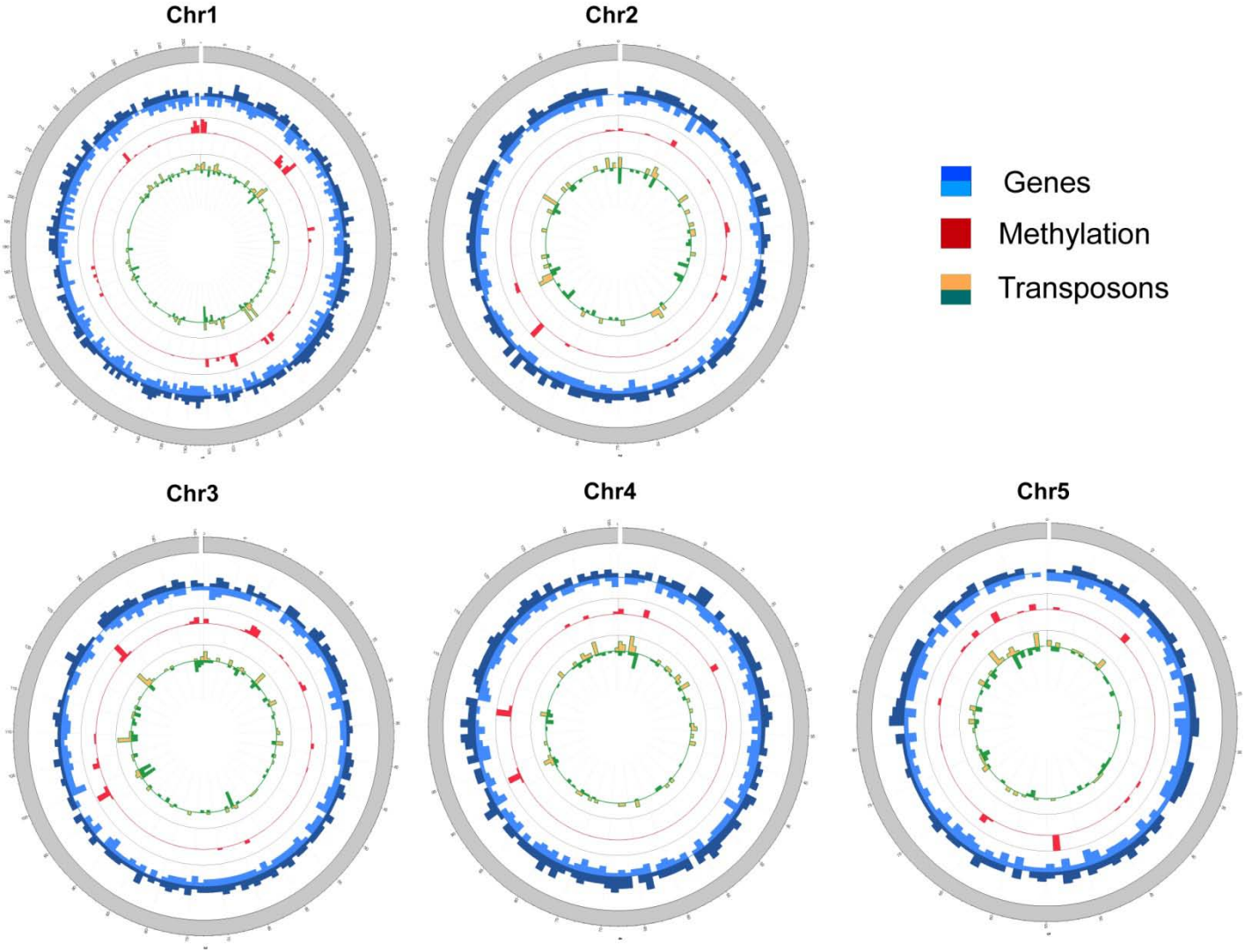

Chr1
Chr2
Chr3
Chr4
Chr5
Genes
Methylation
Transposons

Figure S2.2 Chromosome-wide patterns of methylation and other genomic components. Each of the 33 genomic scaffolds of *P. tricornutum* are shown here in a circular view. Methylation is shown as HMRs with a moving window of 10 kb. Normalized methylation (in red) is shown along with TEs (green) and genes (blue). Strand specific information is shown above and below the line with different colors.

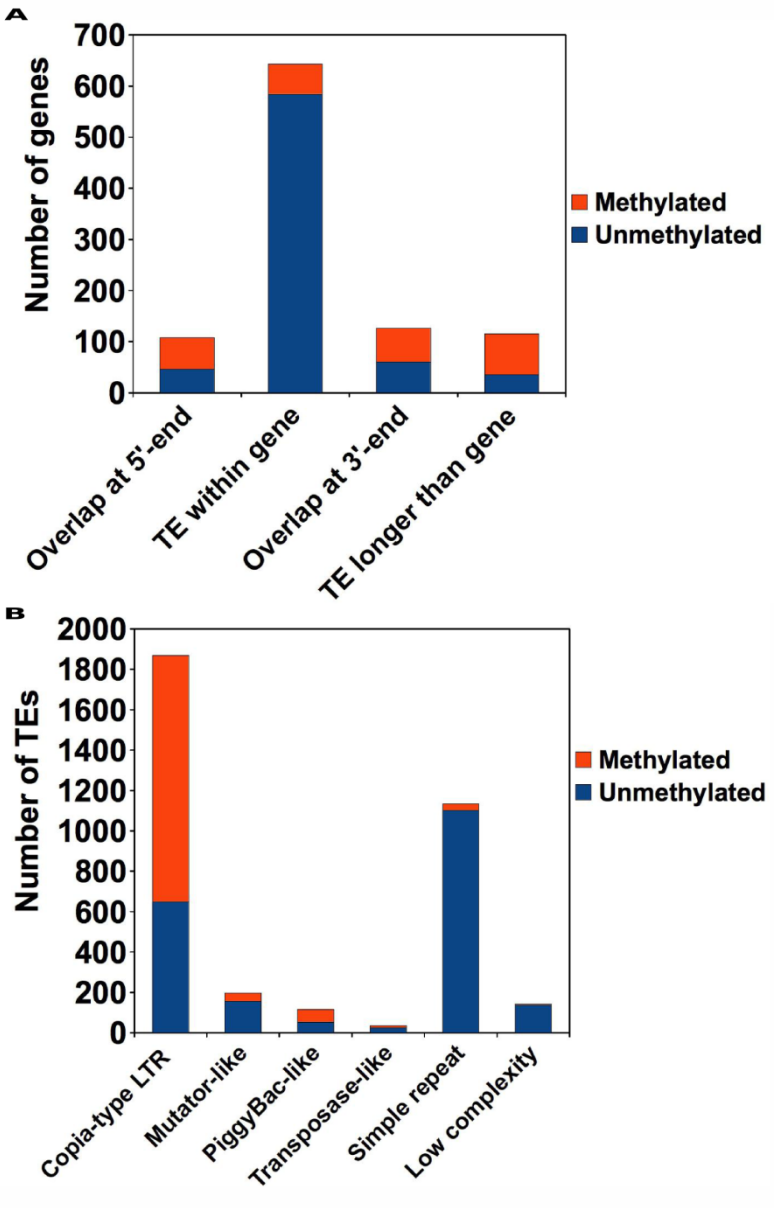


Figure S2.3 TEs within different contexts display different methylation patterns. (A) Methylation within genes that contain TE-derived sequences. A total of 767 genes were TE-overlapped and most of them are inserted as simple repeats within genes. More than half of the genes whose 5' or 3' ends overlap with TE annotations are methylated, suggesting that some may correspond to erroneous gene predictions (e.g., chimeric TE-gene predictions) or to pseudogenes. **(**B) Methylation of different classes of TEs (including elements smaller than 300 bp). The most abundant Copia-type class of retrotransposable elements (5.8 % of genome) are methylated in large numbers. Most simple repeats were devoid of methylation (see also Fig. 1C,D).

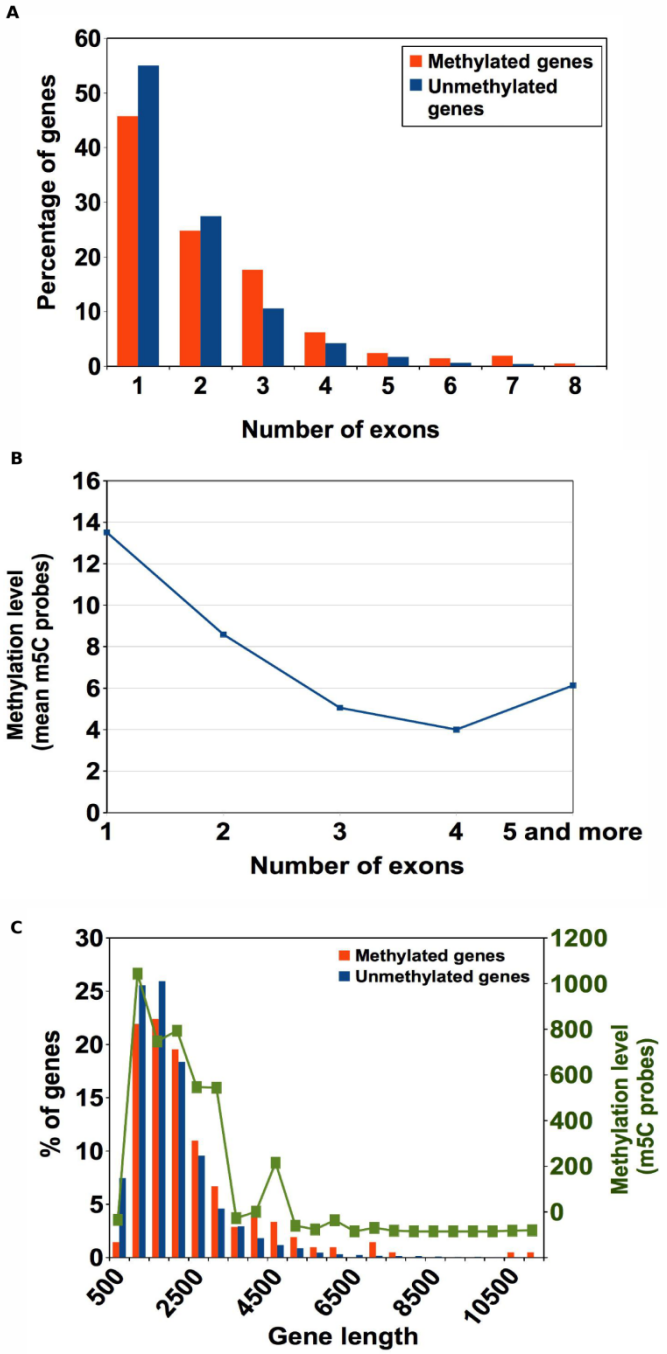


Figure S2.4 Analysis of methylation over exons. (A) Distribution of the methylated and unmethylated gene pools across classes of genes with increasing numbers of exons. (B) Methylation level across classes of genes with increasing numbers of exons. (C) Analysis of gene length and methylation. Distribution of the methylated and unmethylated gene pools across classes of genes with increasing size is shown.

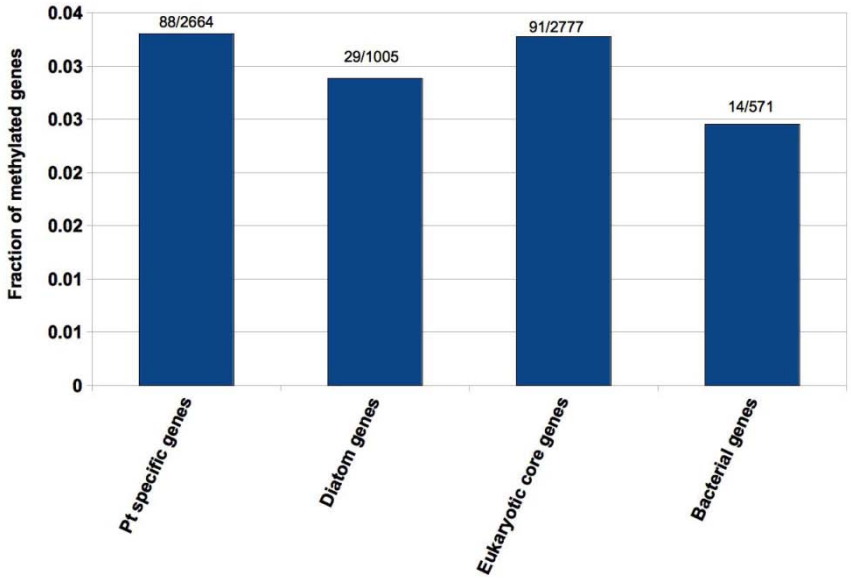


Figure S2.5 Proportion of methylated genes in four evolutionarily important gene classes. Gene content in *P. tricornutum* can be organized into 4 groups (1): Eukaryotic core genes, which are present in all eukaryotes examined, Diatom-specific genes, which are present in both the sequenced diatoms and not elsewhere, Bacterial genes, which are believed to have been acquired by horizontal gene transfer from bacteria, and unique *P. tricornutum* specific genes which are present in this diatom alone.

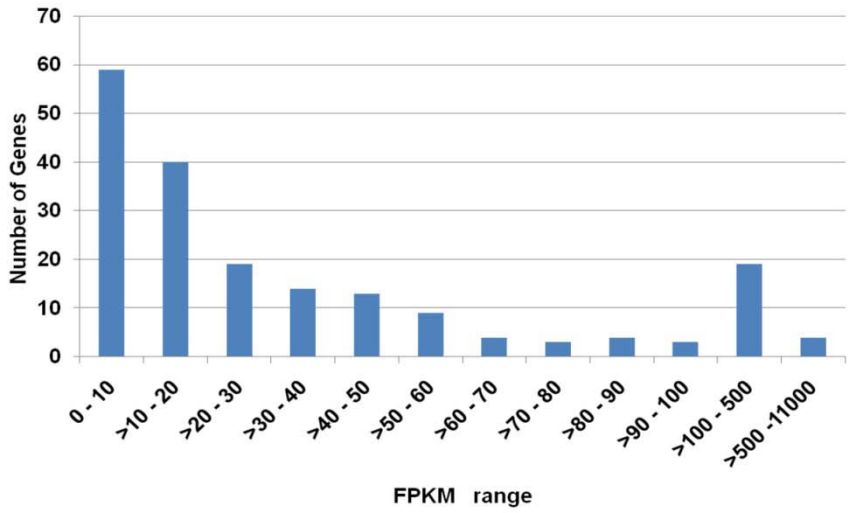


Figure S2.6 Expression analysis of body methylated genes. Fragments Per Kilobase of exon per Million fragments mapped (FPKM) were obtained from RNA-seq fragment counts.

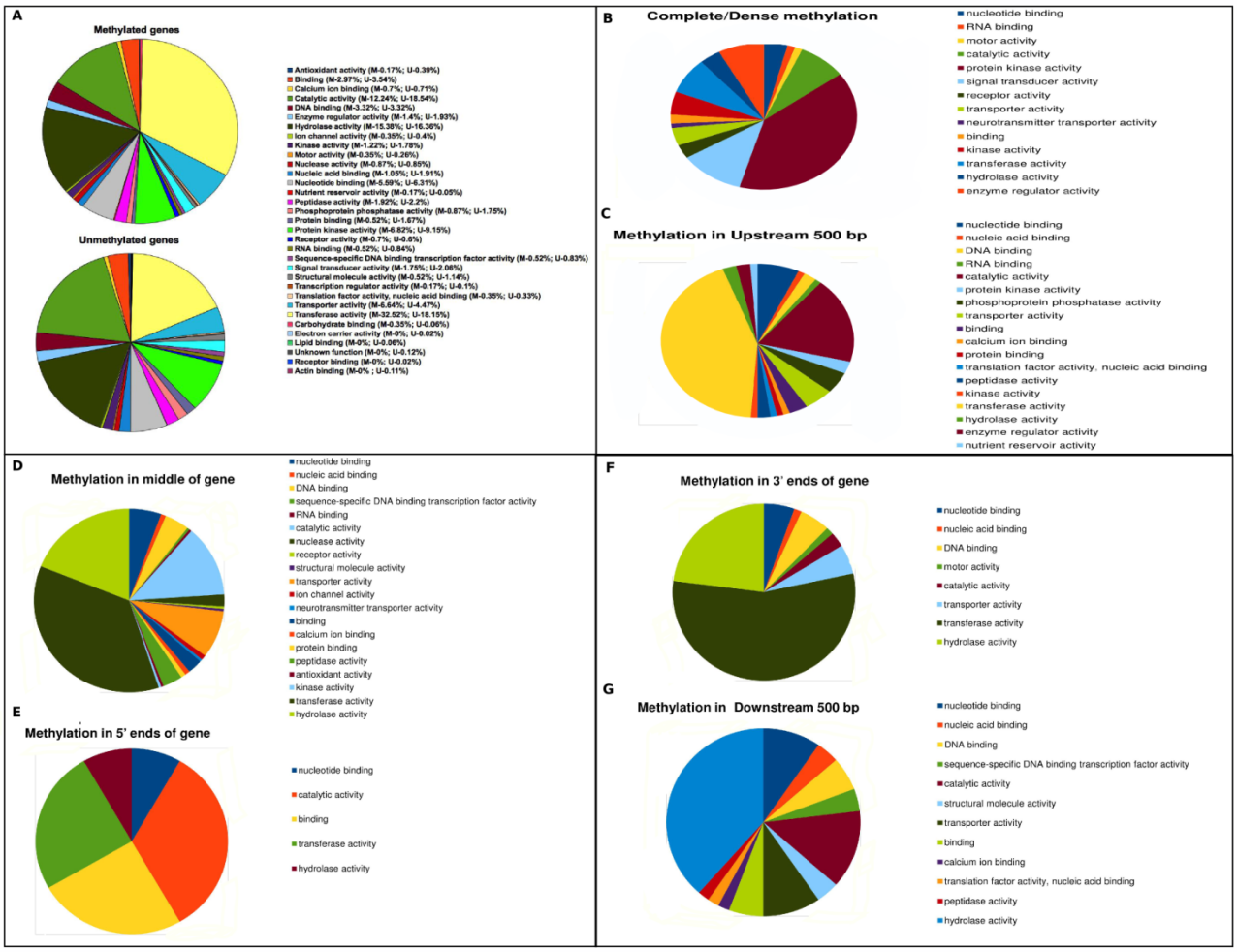


Figure S2.7 GO distribution of different classes of methylated genes. (A) Gene Ontology (GO) distribution of methylated and unmethylated gene sets. A total of 4722 genes were annotated using GO. Methylated genes (M) and Unmethylated genes (U) were grouped by the Molecular Function category of GO. Significant enrichment in the methylated gene set of categories 'transferase,' 'transporter,' 'carbohydrate binding,' and 'nutrient reservoir' activities is apparent. (B) Genes with complete/dense methylation, (C) Genes with methylation in upstream 500 bp region, (D) Genes with gene-body methylation, (E) Genes with methylation in 5' end, (F) Genes with methylation in 3' end, and (G) Genes with methylation in downstream 500 bp region.

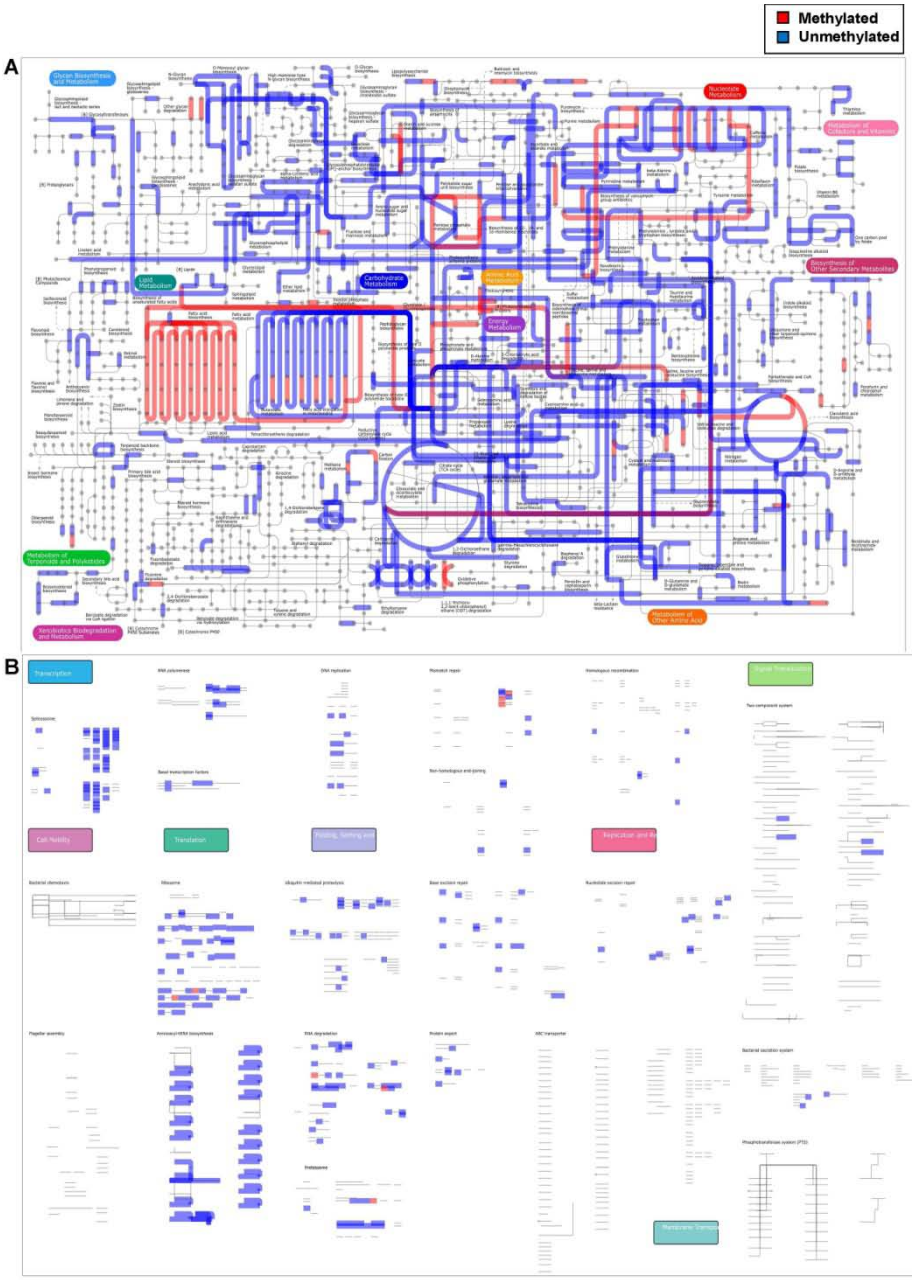


Figure S2.8 Metabolic pathways containing components encoded by methylated genes. KEGG-based overview of (A) metabolic pathways showing multiple functional branch points, and (B) Regulatory pathways.

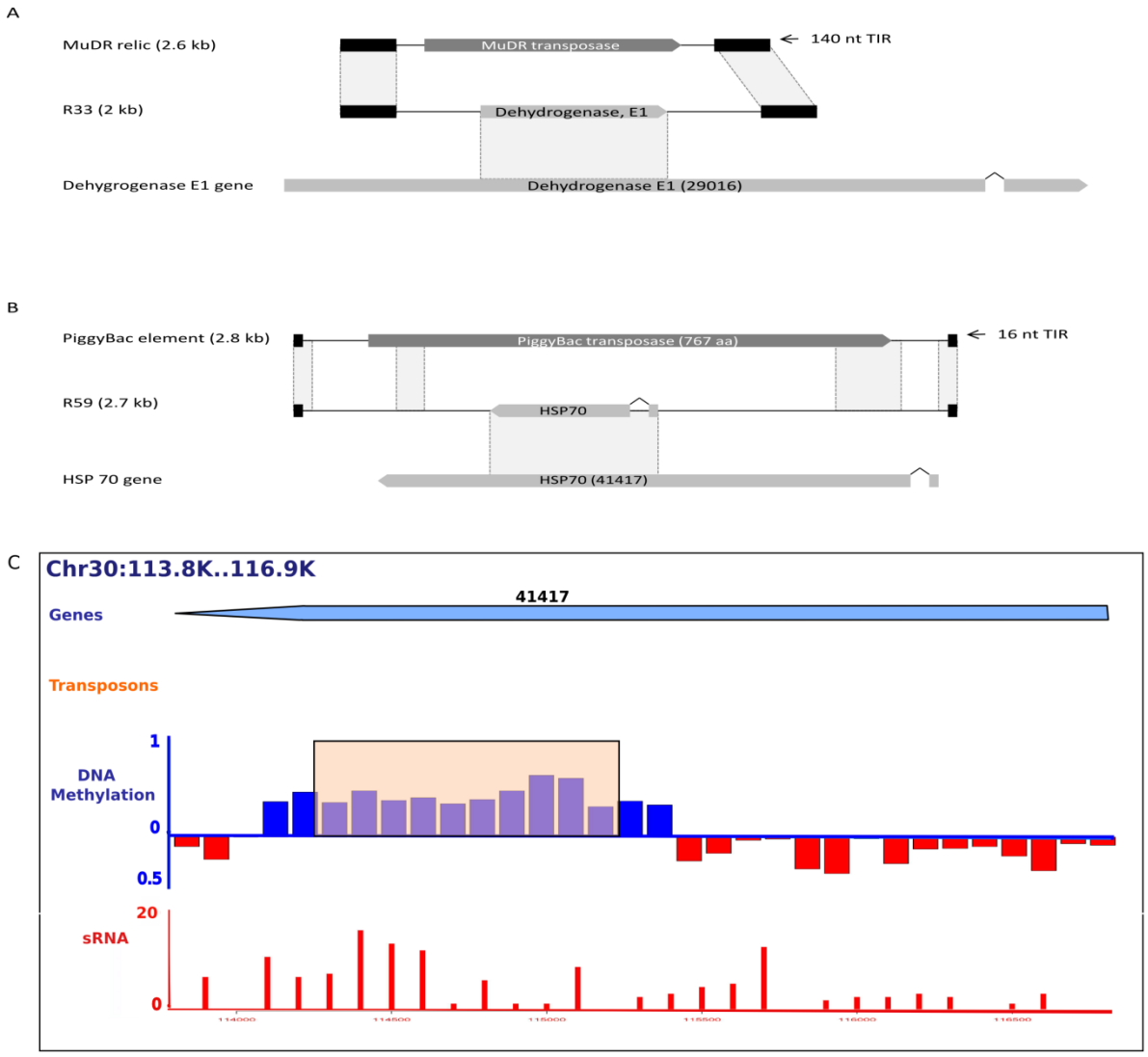


Figure S2.9 Non-autonomous Class II transposable elements with captured exons. Schematic alignment of the exon-containing non-autonomous Class II elements found in the *P. tricornutum* genome: R33 repeat (A), and R59 repeat (B), with respective autonomous element and genes from which captured exon originates. The frame indicates the inserted fragment of the protein in R59. Regions of high similarity between pairs of sequences are connected by grey shades. Terminal inverted repeats (TIRs) are indicated as black boxes. (C) Methylation status of the single copy heat shock protein *HSP70* gene on chromosome 30.

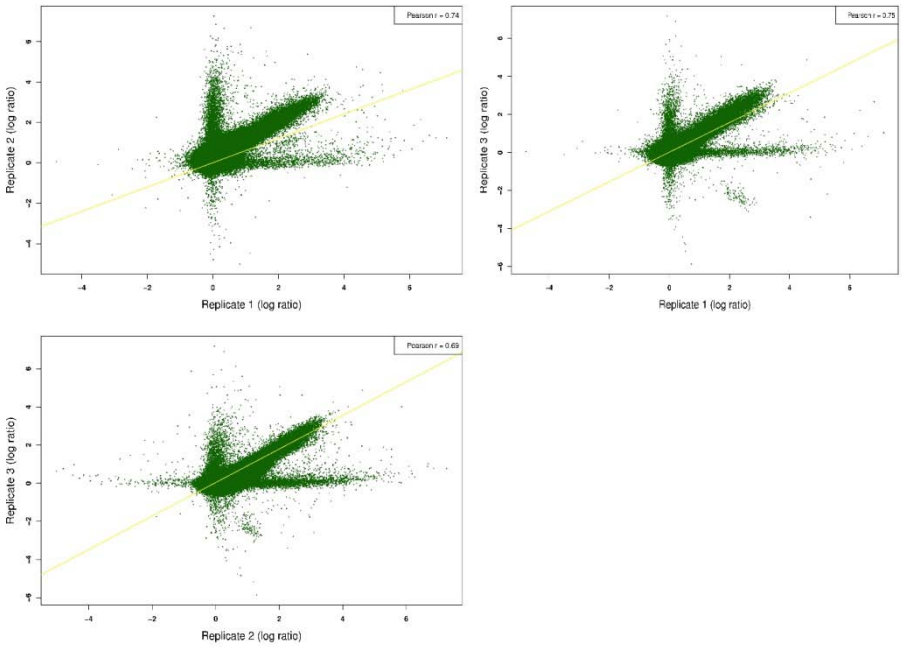


Figure S2.10 McrBC signal correlation between biological replicates. R-values are indicated at top right of each panel.

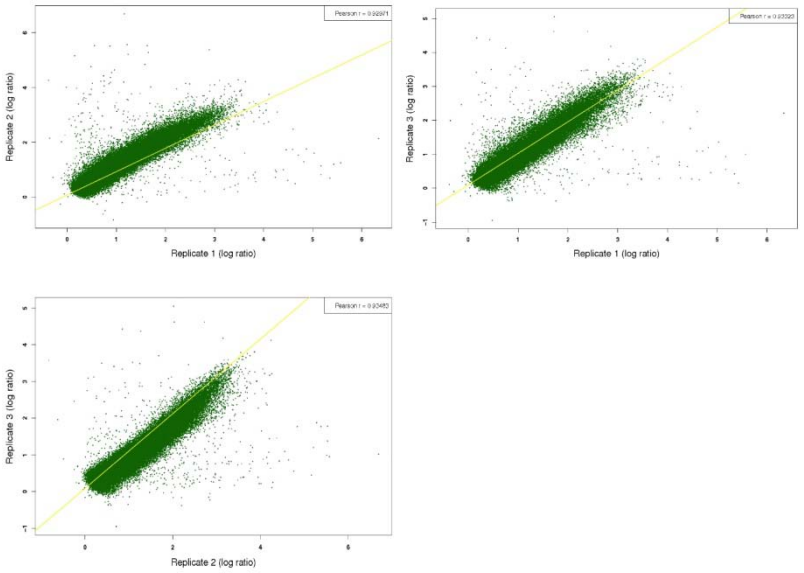


Figure S2.11 Replicate comparison. Correlation of three biological replicates after RINGO normalization with a p-value cutoff of 0.02. Outliers were removed and data shows a significant consistency. Pearson coefficients are indicated in top right of each panel.

Supporting Supplementry Tables

Table S2.1 Validation of methylated and unmethylated loci, including genes and TEs. Several loci from different locations on the genome were chosen for validation.

Protein ID	Location	McrBC methylation	Bisulfite validation
50240	Chr 28	unmethylated *(two sparse probe)	unmethylated
43549	Chr 2	unmethylated (single probe)	unmethylated
33493	Chr 3	unmethylated (single probe)	unmethylated
44978	Chr 5	unmethylated (single probe)	unmethylated
45772	Chr 8	unmethylated (single probe)	unmethylated
47265	Chr 13	unmethylated (single probe)	unmethylated
13557	Chr 12	unmethylated (single probe)	unmethylated
31408	Chr 1	unmethylated (single probe)	unmethylated
20757	Chr 10	unmethylated (single probe)	unmethylated
48392	Chr 17	unmethylated (three sparse probe)	unmethylated
44066	Chr 3	unmethylated (single probe)	unmethylated
47725	Chr 14	unmethylated (two sparse probe)	unmethylated
44551	Chr 4	unmethylated (single probe)	unmethylated
48135	Chr 16	unmethylated (two sparse probe)	unmethylated
47034	Chr 12	methylated (many overlapping probes)	methylated
44119	Chr 3	methylated (many overlapping probes)	methylated
47656	Chr 14	methylated (many overlapping probes)	methylated

47934	Chr 15	methylated (many overlapping probes)	methylated
47656	Chr 14	methylated (many overlapping probes)	methylated
47839	Chr 15	methylated (4 overlapping probes)	methylated
38262	Chr 15	methylated (many overlapping probes)	methylated
34576	Chr 5	methylated (many overlapping probes)	methylated
33360	Chr 3	methylated (many overlapping probes)	methylated
41417	Chr 30	methylated (many overlapping probes)	methylated
PTC30	TE	methylated	methylated
PTC5	TE	methylated	methylated
PTC8	TE	unmethylated	unmethylated

* To further strengthen the normalization approach that considered as methylated three overlapping probes single probe loci were also tested in the validation procedure and were indeed found to be unmethylated.

Table S2.2 Number of probes and distribution of HMRs over the genome.

Designed probes	2,171,824
HMRs	3,887
Probes in HMRs	98,080
HMR coverage in % of genome (bp)	1,412,473 bp of 27.4 Mb genome = 5.16 %
HMRs on TEs	2,806
TEs covered by HMRs	1,368
HMRs on TE inserted genes	604
HMRs on genes (without TEs)	505
HMR coverage/overlap on TEs (bp)	1,081,702 bp
HMR coverage/overlap on genes (bp)	121,932 bp

Table S2.3 Number of methylated genes and distribution of methylation over genes.

Methylated genes	326
Genes methylated on exons (May extend to introns)	217
Genes methylated on introns (exclusively on introns)	3
Genes methylated in 500 bp upstream region	77
Genes methylated at 5'-end (20% by gene length)	11
Complete/high methylation	23
Genes methylated in middle (60% by gene length)	171
Genes methylated at 3'-end (20% by gene length)	14
Genes methylated in 500 bp downstream region	30

Table S2.4 Distribution of total and methylated genes from groups of different origin according to their R value.

Gene category	R > 12 (Total genes = 7112)	R < 12 (Total genes = 703)	Number of methylated genes with R < 12	Number of methylated genes with R > 12	Total genes
Bacterial genes	397	35	1	8	571
Eukaryotic core genes	2417	311	12	76	2777
Diatom specific genes	900	89	5	24	1005
Pt specific Genes	2462	170	5	84	2664

2.8.2 Detailed Methods

Culture conditions

Cultures of *P. tricornutum* Bohlin clone Pt1 8.6 (CCMP2561) were obtained from the Provasoli-Guillard National Center for Culture of Marine Phytoplankton, Bigelow Laboratory for Ocean Sciences, USA. Cultures were grown in f/2 medium (2) made with 0.2-µm-filtered and autoclaved local seawater supplemented with f/2 vitamins and inorganic filter sterilized nutrients. Cultures were incubated at 19°C under cool white fluorescent lights at approximately 75 µmol.m-2.s-1 in 12h light: 12h dark conditions and maintained in exponential phase in semicontinuous batch cultures. Sterility was monitored by occasional inoculation into peptone-enriched media to assess the presence of bacteria (3).

DNA preparation for hybridization (Window McrBC Restriction)

Genomic DNA from three *P. tricornutum* cultures (biological replicates) was sonicated to an average size of 600 nt using the Bioruptor sonicator (UCD-200, Diagenode). After size-separation by gel electrophoresis, the gel window corresponding to the 500-700 nt DNA fragments was excised and DNA was recovered using QIAquick Gel Extraction kit. A total of 200 ng of size-filtered DNA was incubated overnight at 37°C in a reaction containing 10 units of McrBC enzyme (New England Biolabs), 1x McrBC buffer, 3 mM GTP, 300 ng/ml BSA, and water. In negative controls, GTP, which is the co-factor required for McrBC activity, was replaced by water. McrBC was then inactivated by heating the reactions for 20 minutes at 65°C. After size-separation by gel electrophoresis, DNA contained in the 500-700 nt gel window was recovered using QIAquick Gel Extraction kit. The efficiency and specificity of DNA digestion by McrBC was verified by semi-quantitative PCR by assessing the amount of DNA from PtC13 CoDi element and from the histone H4 gene which appeared unmethylated in all our set up experiments. For each replicate, we observed a significant and similar exclusion of methylated DNA in the samples compared to negative controls.

Primers: PtC13_Fw 5' TTG-CAA-ATT-TTC-AGC-AGC-AC 3' and PtC13_Rev 5' AGA-AGG-CTG-GGA-CAC-AGA-GA 3'; H4_Fw 5' AGG-TCC-TTC-GCG-ACA-ATA-TC 3' and H4_Rev 5' ACG-GAA-TCA-CGA-ATG-ACG-TT 3'.

Sample preparation and microarray hybridization

This protocol was performed according to Lippman et al. (4). Ten nanograms of (+/-) McrBC treated DNA were amplified using Sigma's WGA kit (St. Louis, MO). Three amplification reactions with and without GTP were performed in parallel and purified using Microcon YM-10 Columns (Millipore, Bedford, MA). One and half micrograms of amplified material were used for amplification and labeling with fluorescent dyes, Cy3 and Cy5, (GE HealthCare, UK) following NimbleGen's "NimbleChip Arrays User's Guide: DNA Methylation Analysis v2.0" (Roche NimbleGen, Germany). Microarray hybridization and washing were performed according to the NimbleGen "DNA Methylation Analysis v2.0" User Guide. Each array was scanned on the Axon4000B Scanner (Molecular Devices, Sunnyvale, CA) at 2 and 5 micron resolution. NimbleGen 2.1M Phaedoctylum tricornutum tiling arrays were designed based on the JGI Phatr2 genome. A total of 2.1 million probes are represented on this array. NimbleGen provided design and probe annotation.

Experimental validation

To test the sensitivity and validity of our WMR-chip and analysis pipeline, bisulfite-DNA sequencing was performed on 28 loci, genes and TEs included (Supplementary Table 1). Axenic *P. tricornutum* (CCMP632) cells were grown as described above. Genomic DNA was extracted as described previously (5) and sonicated for 7 cycles (20 seconds on and 30 seconds off) using a Bioruptor (UCD-200, Diagenode) to generate fragments between 500 and 1000 bp. EZ DNAMethylation-Direct™ Kit (ZYMO RESEARCH) was used to do bisulfite treatment. 200-500 ng of DNA in 20 ul water was added to 130 ul CT-conversion reagent solution. Samples were incubated in a thermal cycler performing 98 °C for 8 minutes, 64 °C for 3.5 hours and 4 °C storage for up to 20 hours. The samples were purified on columns using a series of buffers provided by the manufacturer. ZymoTaq™ PreMix (ZYMO RESEARCH) was used to amplify bisulfite-treated DNA. Primers for PCR were designed by on-line software Methprimer (http://www.urogene.org/methprimer/index1.html). After gel purification, PCR products were cloned into pGEM-T Easy Vector system (Promega ref: A1360) and sequenced. To confirm consistency, 10 clones were selected for each target sequence for sequencing. The sequences were analyzed by on-line software CYMATE (http://www.cymate.org).

RNA-seq preparation

P. tricornutum clone Pt1 8.6 cells were harvested at exponential phase and total RNA was extracted as described previously (5). 1 µg of total RNA was used for first strand cDNA synthesis followed by double strand cDNA using Mint Universal Kit from Evrogen (SK002). Double stranded DNA quality was monitored using agarose gels. cDNA was used for non-directional cloning and cDNA library construction for Illumina sequencing by Beckman Coulter Genomics. Sequencing was performed with a read length of 75 bp and sequencing coverage of 1.5 Gb.

Identification of methylated regions, distribution and expression analysis

Statistically significant probe bound regions (ChIP-enriched genomic regions) were detected using the RINGO (6) in R Bioconductor package. A cut-off P-value of 0.02 was set for normalization. Boundaries for methylated regions were defined as those with a minimum of three enriched overlapping enriched probes, using a moving window of 50 bp. Using TileHMM (7) and a hierarchical mixture model method using JAMIE (8) with an FDR value of 2% and 50 bp peak conditions yielded similar results (90 % agreement), although the boundaries differed by a few nucleotides. We found at least 98,080 enriched probes, which amounts to 4.5 % of probes covered. Normalization on the three biological replicates yielded robust consistency which was statistically validated using a Student t-test and showed a Pearson R-value between 0.92 – 0.93 (Figs. S10, S11). Functional analyses were done using GOSlim, GOEAST (9) and iPATH2 (10). Expression correlations with methylation were done using R-values derived from the EST sequences (1) and cDNA sequence data. Single-end Illumina reads (36 bp) from cDNA sequence data were mapped to the genome (PhatrV2 JGI) using TopHat v.1.1.3 (11). Relative abundances of transcripts were measured as Fragments Per Kilobase of exon per Million fragments mapped (FPKM) using Cufflinks (12). Data processing, analysis, and plotting were done using Python, R/Bioconductor, and CIRCOS (13). A genome browser based on Gbrowse is available to explore this methylome data (http://ptepi.biologie.ens.fr/cgi-bin/gbrowse/Pt_Epigenome).

2.8.3 Supplementary References

1. Maheswari U*, et al.* (2010) Digital expression profiling of novel diatom transcripts provides insight into their biological functions. *Genome Biol* 11(8):R85.

2. Guillard RRL (1975) Culture of phytoplankton for feeding marine invertebrates. *Culture of Marine Invertebrate Animals*, eds Smith W & Chanley MH (Plenum Press, New York, USA), pp 26-60.

3. Andersen RA, Morton SL, & Sexton JP (1997) Provasoli-Guillard National Center for Culture of Marine Phytoplankton 1997 list of strains. *J. Phycol.* 33((suppl)):1-75.

4. Lippman Z, Gendrel AV, Colot V, & Martienssen R (2005) Profiling DNA methylation patterns using genomic tiling microarrays. *Nat Methods* 2(3):219-224.

5. Siaut M*, et al.* (2007) Molecular toolbox for studying diatom biology in Phaeodactylum tricornutum. *Gene* 406(1-2):23-35.

6. Toedling J*, et al.* (2007) Ringo--an R/Bioconductor package for analyzing ChIP-chip readouts. *BMC Bioinformatics* 8:221.

7. Humburg P, Bulger D, & Stone G (2008) Parameter estimation for robust HMM analysis of ChIP-chip data. *BMC Bioinformatics* 9:343.

8. Wu H & Ji H (2010) JAMIE: joint analysis of multiple ChIP-chip experiments. *Bioinformatics* 26(15):1864-1870.

9. Zheng Q & Wang XJ (2008) GOEAST: a web-based software toolkit for Gene Ontology enrichment analysis. *Nucleic Acids Res* 36(Web Server issue):W358-363.

10. Yamada T, Letunic I, Okuda S, Kanehisa M, & Bork P (2011) iPath2.0: interactive pathway explorer. *Nucleic Acids Res* 39(Web Server issue):W412-415.

11. Trapnell C, Pachter L, & Salzberg SL (2009) TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics* 25(9):1105-1111.

12. Trapnell C*, et al.* (2010) Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat Biotechnol* 28(5):511-515.

13. Krzywinski M*, et al.* (2009) Circos: an information aesthetic for comparative genomics. *Genome Res* 19(9):1639-1645.

Chapter III:

Genome wide analysis of the histone marks H3K4me2, H3K9me2 and H3K27me3 in the model diatom *Phaeodactylum tricornutum*

Table of contents

Chapter III: 105

Genome wide analysis of the histone marks H3K4me2, H3K9me2 and H3K27me3 in the model diatom *Phaeodactylum tricornutum* 105

3.1 Abstract 108

3.2 Introduction 108

3.3 Results 110

3.3.1 Identification of genomic regions associated with H3K4me2, H3K9me2 and H3K27me3 110

3.3.2 Genome wide distribution of H3K4me2, H3K9me2 and H3K27me3 on genes 117

3.3.3 Genome wide distribution of H3K4me2, H3K9me2 and H3K27me3 on TEs 123

3.3.4 Combinatorial marking of H3K4me2, H3K9me2 and H3K27me3 on the same genomic regions 125

3.3.5 Combinatorial effect of histone modifications and DNA methylation on gene expression 127

3.4 Discussion 132

3.4.1 H3K4me2 mainly marks genes in *P. tricornutum* 132

3.4.2 H3K9me2 mainly marks TEs in *P. tricornutum* 133

3.4.3 The distribution of H3K27me3 in *P. tricornutum* is unorthodox 134

3.4.4 Methylated TEs and heavily methylated genes tend to be co-marked by H3K27me3 and H3K9me2 139

3.5 Perspectives 141

3.6 Material and Methods 142

3.6.1 Growth conditions 142

3.6.2 Peptide competition assay for antibody specificity test 142

3.6.3 Chromatin immunoprecipitation and sequencing 143

3.6.4 RNA sequencing 143

3.6.5 Computational analysis of histone modifications *P. tricornutum* by ChIP-sequencing 144

3.7 References 145

3.8 Supplementary information 157

3.8.1 Supplementary Figures 157

3.8.1 Supplementary tables 163

3.9 Annex-Chromatin immunoprecipitation protocol 166

Chromatin immunoprecipitation coupled to detection by quantitative real time PCR to study in vivo protein DNA interactions in two model diatoms *Phaeodactylum tricornutum* and *Thalasiossira pseudonana* 166

3.1 Abstract

Histone modifications constitute one of major epigenetic phenomena and thus play key roles in chromatin-level regulation in eukaryotes. A growing list of histone modifications is being described and the complexity of their functions is being explored. However, investigations on histone modifications are limited to only a few model species. Diatoms are thought to be the most successful group of eukaryotic phytoplankton in the modern ocean. The ability of diatoms to adapt rapidly to different environments implies not only DNA sequence-based regulation but also more reversible and flexible epigenetic changes. Studies on diatom epigenomes can therefore enhance our understanding of the mechanisms underlying diatom adaptation to the environment. Herein we report an epigenome map based on three histone modification marks (H3 lysine4 dimethyl, H3 lysine 9 dimethyl and H3 lysine 27 trimethyl) combined with genome wide DNA methylation, small RNA, and expression profile from a Stramenopile, the model diatom *Phaodactylum tricornutum*. We find that H3K4me2 is mainly associated with genes while both H3K9me2 and H3K27me3 target mostly transposable elements (TEs). The distribution of H3K27me3 is unusual and different from that profiled in other species in that it is mainly associated with TEs rather than genes. We also report that the genes marked by H3K27me3 tend to be lowly and differentially expressed. Furthermore, we find that H3K27me3 and H3K9me2 tend to co-mark not only methylated TEs but also heavily methylated genes, which appears to be important for maintaining the silencing of differentially expressed genes. The combinatorial analysis of different histone marks and DNA methylation in *P. tricornutum* provides an overview of diatom chromatin landscapes, and will help to define conserved structural and functional features.

3.2 Introduction

All genomes exert their functions in the context of chromatin. In eukaryotic organisms, nucleosomes constitute the basic unit of chromatin. They are composed of 146 bp of DNA and a histone core, which is an octamer consisting of two copies each of histones H2A, H2B, H3 and H4. Besides DNA that can be methylated, the residues of all histone proteins can be subject to different modifications, particularly their amino (N) terminal tails. So far more than 60 post-translational modifications have been detected (Bannister & Kouzarides, 2011). DNA methylation and histone modification are two major components of epigenetics, defined as the study of heritable changes other than changes in the underlying DNA sequence. The sum

information of genome wide distribution of DNA methylation and histone modifications is termed the epigenome. Different epigenetic marks and different combinations of them regulate transcription and other aspects of genome dynamics.

Profiling epigenomes from different organisms has deciphered how genomes are organized and regulated by different epigenetic marks (Bernstein, Meissner, & Lander, 2007; Liu et al., 2011; Millar & Grunstein, 2006; Rugg-Gunn, Cox, Ralston, & Rossant, 2010; Schübeler et al., 2004; Sinha, Wirén, & Ekwall, 2006; Wang, Schones, & Zhao, 2009; Zhao & Zhou, 2012; Zhou, Goren, & Bernstein, 2011; van Leeuwen & van Steensel, 2005). For multicellular organisms, the comparison of epigenomes at different stages during development has revealed the roles of epigenetic marks in cell differentiation and development (Lafos et al., 2011; Rugg-Gunn et al., 2010).

Diatoms are believed to be the most successful, abundant and diverse eukaryotic unicellular organisms in the modern oceans. They play vital roles in global ecosystems by contributing around 20% of global primary production. *Phaeodactylum tricornutum* is one of the two diatoms with a completed genome sequence and has been used as a model diatom species for decades, with established gene manipulation assays and molecular resources (De Riso et al., 2009; Maheswari et al., 2010; Siaut et al., 2007). Analysis of the *P. tricornutum* genome has revealed that the genome contains genes of different origins such as from red algae, green algae and bacteria, which have been obtained by endosymbiotic gene transfer and horizontal gene transfer (Bowler et al., 2008).

Besides genetic regulation, it has been proposed that reversible and flexible epigenetic mechanisms may also contribute to the ecological success of diatoms in a dynamic oceanic environment (Tirichine & Bowler, 2011). Another interesting question is to compare diatom epigenomes with those of other species to explore the origin and evolution of epigenetic regulation in eukaryotes. Epigenetic studies in model species such as the higher plant *Arabidopsis thaliana*, the fruit fly *Drosophila melanogaster*, the worm *Caenorhabditis elegans*, the budding yeast *Saccharomyces cerevisiae*, mouse and human have revealed the importance of epigenetic regulation in a wide range of biological processes (Ernst et al., 2011; Filion et al., 2010; Hawkins et al., 2010; Kharchenko et al., 2011; T. Liu et al., 2011; Marks et al., 2012; Millar & Grunstein, 2006; Roudier et al., 2011a).

Genome wide DNA methylation profiling of *P. tricornutum* based on McrBC-ChIP revealed that DNA methylation is low, with relatively extensive TE DNA methylation but only a few methylated genes that are strongly differentially expressed (Chapter II). Here, I continued to explore the epigenome of *P. tricornutum* by Illumina/Solexa sequencing after conducting chromatin immunoprecipitation (ChIP) using antibodies that target specific histone modifications (H3K4me2, H3K9me2 and H3K27me3). These three histone marks were chosen because they were shown in previous studies to be associated with distinct transcriptional activities and annotation features in other model species: H3K4me2 is mainly associated with genes; H3K9me2 is mainly associated with TEs and H3K27me3 is associated with repressed genes (Bessler, Andersen, & Villeneuve, 2010; Hawkins et al., 2010; Lafos et al., 2011; Lienert et al., 2011; T. Liu et al., 2011; Pauler et al., 2009a; Xiaoyu Zhang et al., 2007a; J. Zhou et al., 2010).

In order to investigate the correlations between histone modifications and gene expression, I further profiled gene expression patterns by sequencing cDNA synthesized from mRNA using Illumina sequencing (RNA-seq). In this chapter I describe an integrated genome-wide analysis of these three histone modifications and gene expression combined with previously published DNA methylation. The combinations of epigenetic marks studied here are distinct from those found in other organisms; in particular H3K27me3 has a novel distribution in *P. tricornutum* compared to what has been observed in other organisms.

3.3 Results

3.3.1 Identification of genomic regions associated with H3K4me2, H3K9me2 and H3K27me3

The general objective of this study was to generate epigenomic maps for *P. tricornutum* from the histone marks H3K4me2, H3K9me2 and H3K27me3 using chromatin extracted from cells in exponential phase followed by immunoprecipitation and high throughput Illumina sequencing (ChIP-seq). The peptide competition assay to test the specificity of antibodies for H3K4me2 (Millipore ref: 07-030) and H3K27me3 (Millipore ref: 07-449) in *P. tricornutum* was found to be satisfactory (**Figure 3.1**). No cross-reactivity was detected between H3K4me2 antibody and peptide H3K4me1 or peptide H3K4me3. The H3K27me3 antibody also did not show cross-reactivity with H3K27me2 nor H3K27me1 peptides. On the other

hand, the H3K9me2 antibody (Millipore ref: 17-648) did show some cross reactivity with H3K9me3, although this particular antibody showed better results among H3K9me2 antibodies that were tested.

Figure 3.1 The peptide competition assay using H3K4me2, H3K9me2, H3K27me3 antibody in *P. tricornutum*. Different ratios of antibody and peptide have been tested. Mono, Di and Tri-methylated peptides were used in different ratios as indicated. Western blot detection results were demonstrated here when the antibody was saturated by the peptide. The loading control is H4.

The ChIP-Seq resulted in 37,546,690 reads, 36,329,249 reads, 35,973,625 reads and 38,202,621 reads respectivelyfrom H3K4me2, H3K9me2, H3K27me3 and input control,

respectively. The average read length was 36 bp. The quality of sequencing was checked and filtering of raw reads performed by discarding reads with more than 2 bases with Phred Quality score of 30. Duplicate reads, i.e., reads mapping to the same genomic regions were avoided. ChIP-Seq peak signals identification using SICER resulted in 9,748 H3K4me2 modified domains covering 7,269,800 bp, 1,991 H3K9me2 modified domains covering 4,947,800 bp, and 865 H3K27me3 modified domains covering 2,903,600 bp (**Figure 3.2A**). The number of domains modified by H3K4me2 was therefore much higher than H3K9me2 and H3K27me3. The length of H3K9me2 (up to 20 kb) and H3K27me3 (up to and even longer than 20 kb) modified domains was significantly longer than that of H3K4me2 (most of them are less than 2 kb) (**SuppFigure 3.1**).

About half of the genome was marked by these histone modifications. H3K4me2, H3K9me2 and H3K27me3 marked regions covered 27.8%, 18.9% and 11.1% of the genome, respectively. Nearly half of the regions marked by H3K27me3 overlapped with H3K9me2. H3K4me2 marked regions slightly overlapped with H3K27me3 and H3K9me2 marked regions (**Figure 3.2B**). It is worth noting that 13.4% of H3K4me2 modified domains lie within intergenic regions. Beside this, for all three histone marks, based on the number of modified domains it seems that there are more enriched domains within genic regions than on TEs (**Figure 3.2C**). A systematic analysis of the locations of H3K4me2 marked regions revealed a very significant correlation between H3K4me2 and the presence of annotated genes. 81.94% of annotated genes (7,896 out of 9,636 total genes) were found to be associated with H3K4me2 while only 605 TEs (out of 3,493 TEs in total) were marked with H3K4me2. In stark contrast with H3K4me2, H3K9me2 was found associated principally with annotated TEs. A total of 1,350 (38.7%) of TEs were found to be marked with H3K9me2 while 1,350 genes were marked by it. Compared to H3K9me2, H3K27me3 was found to be even more highly enriched on TEs. 41% of annotated TEs (1,441 out of 3,493 TEs in total) were associated with H3K27me3 while only 586 annotated genes were marked with it.

The positive association between H3K4me2 and genes has been observed in plants, *C. elegans*, *Drosophila* and mammals (T. Liu et al., 2011; Pekowska, Benoukraf, Ferrier, & Spicuglia, 2010; Roudier et al., 2011a; Yin, Sweeney, Raha, Snyder, & Lin, 2011). H3K9me2 has been found mainly associated with TEs in heterochromatic regions in different organisms.

In *A. thaliana* H3K9me2 is associated with constitutive heterochromatin, transposons, repeat elements and involved in silencing of transposon activity (Lippman et al. 2004; Turck et al. 2007) (Bernstein et al., 2007; Brykczynska et al., 2010; Filion et al., 2010; X. Li et al., 2008; T. Liu et al., 2011; Roudier et al., 2011b; Wang et al., 2009; Xiaoyu Zhang, Bernatavichute, Cokus, Pellegrini, & Jacobsen, 2009). Furthermore, the distribution pattern of H3K27me3 on annotated features is distinct from that found in other organisms that have been profiled so far. In *A. thaliana*, *C. elegans*, *Drosophila* and mammals, H3K27me3 mainly correlates with genes not TEs, in *P. tricornutum* it significantly associates with TEs (**Figure 3.2D**).

A representative genomic region of chromosome 8 is shown in **Figure 3.3** to demonstrate the general distribution of these three histone marks. To validate the ChIP-seq results, I performed ChIP-qPCRpcr on a random collection of genomic loci including TEs and genes (**Sup Table 3.1**). This approach successfully validated most of the selected loci (82.4% for H3K4me2, 77.3% for H3K9me2 and 100% for H3K27me3).

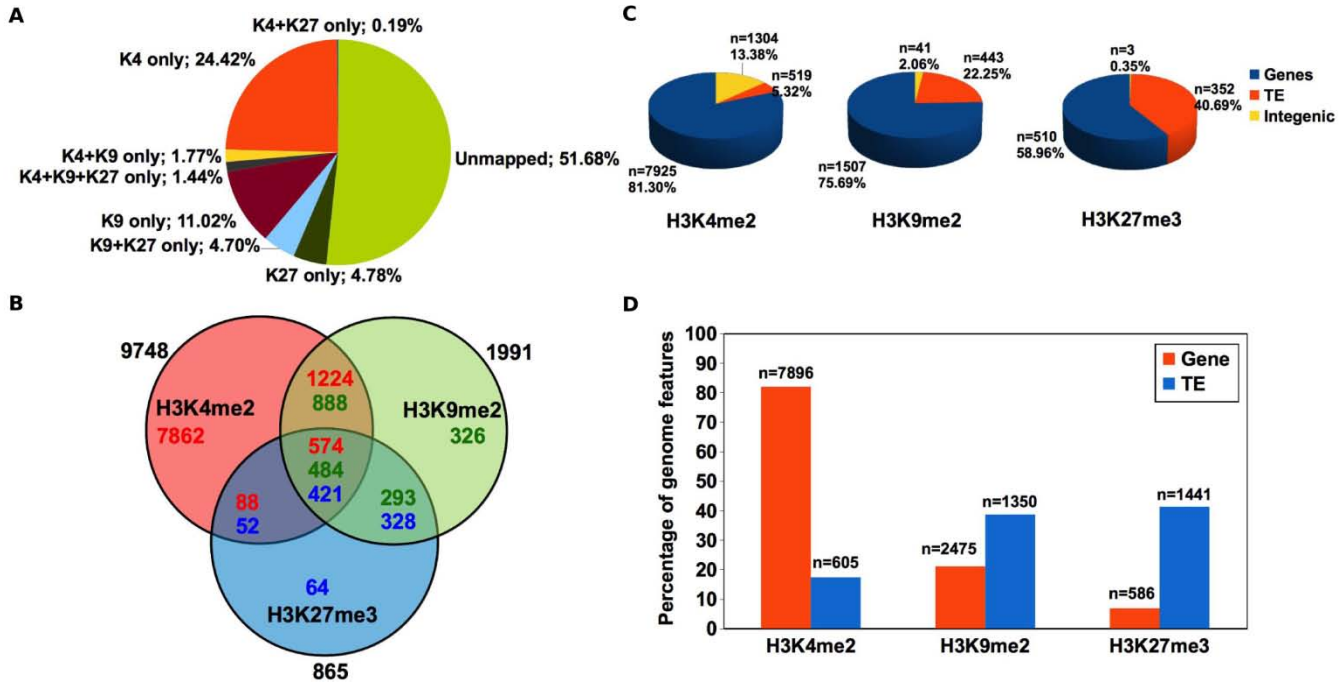


Figure 3.2 A. **ChIP-seq signal peaks distribution for H3K4me2, H3K9me2, and H3K27me3 in base pairs. B**. Distribution of Peak interval over the gene, transposable elements and intergenic regions. **C**. Venn diagram represents the number of peaks from each mark overlapping with the other. One interval can overlap with multiple interval of the other. Each number within the Venn diagram is exclusive. Peak distribution for each mark is calculated from the overlap between ChIP-seq mark and the gene or transposon using Peak Annotator. **D**. Number of genome features (Genes and Transposable elements) marked by these three histone marks. Percentage was derived using the known genome feature annotations.

Figure 3.3 A snapshot of 20 kb region of chromosome8 of *P. tricornutum* epigenomic browser. Average tag density is the number of reads. The H3K4me2 modification is found heavily on genes. Note the H3K27me3 and H3K9me2 marked genes are also heavily methylated and lack H3K4me2. H3K4me2 and H3K27me3 are antagonistic and mutually exclusive on most genomic regions.

3.3.2 Genome wide distribution of H3K4me2, H3K9me2 and H3K27me3 on genes

To investigate the enriched pattern of each histone modification on genes, we examined the average enrichment of H3K4me2, H3K9me2 and H3K27me3 on genes within the upstream 500 bp of coding sequence (CDS), the entire CDS, and the downstream 500 bp from the CDS. The enrichment of H3K4me2 significantly peaks near the 5'end of the CDS (**Figure 3.4A**). The peaks of H3K27me3 and H3K9me2 enrichment lie more in the middle of the transcribed regions and are not as sharp as H3K4me2. Enrichment of H3K9me2 and H3K27me3 in the flanking regions is almost even. H3K9me2 and H3K27me3 did not show any tendency to mark on preferentially longer or shorter genes (**Figure 3.4B**).

To explore the relationship between gene expression and each of these histone marks, cDNA from *P. tricornutum* cells grown in the same conditions as were used for the ChIP-seq experiments was generated and sequenced (RNA-seq). Based on comparison of RNA-seq and ChIP-seq data the influence of various combinations of epigenetic marks on the mRNA levels of different genes was assessed genome wide. The genes marked by H3K4me2 show the highest average expression levels and the largest variation in expression (**Figure 3.5A**). Strikingly, genes marked by H3K27me3 displayed the lowest gene expression level, indicating that H3K27me3 associates with repressed genes in *P. tricornutum*, consistent with previous studies in other organisms (Lafos et al., 2011; Young et al., 2011; Xiaoyu Zhang et al., 2007a). To some extent, H3K9me2 also displays a moderate repressing effect on genes. Generally the genes marked by H3K9me2 also have lower expression levels than the genes marked by H3K4me2 but it is not as significant as with H3K27me3. Genes marked by H3K4me2 shows at the mid position shows higher expression level, while H3K9me2, H3K27me3 marks within the gene-body correlated with reduced represses expression (**Figure 3.5B**). Taken together these observations suggest that H3K4me2 is the general mark for expressed genes, whereas H3K27me3 and H3K9me2 appear to associate with repressed genes.

We further examined whether there are correlations between any of these histone modifications and differential gene expression. The degree of differential expression of *P. tricornutum* genes across 16 cDNA libraries was previously estimated by calculating the statistical significance of differential mRNA levels in specific conditions compared with a

random distribution (Maheswari et al., 2009). R-values indicate the degree of differential expression: constitutively expressed genes have low R-values while genes significantly over-represented in specific growth conditions (i.e., differentially expressed) have high R-values. We found that genes marked by H3K27me3 have significantly increased R-values compared to other groups of genes (**Figure 3.5C**). Conversely, genes marked with H3K9me2 also show increased R-values albeit not as significant as H3K27me3. These observations indicate that genes marked by either H3K9me2 or H3K27me3, in particular H3K27me3, are under tighter transcriptional control.

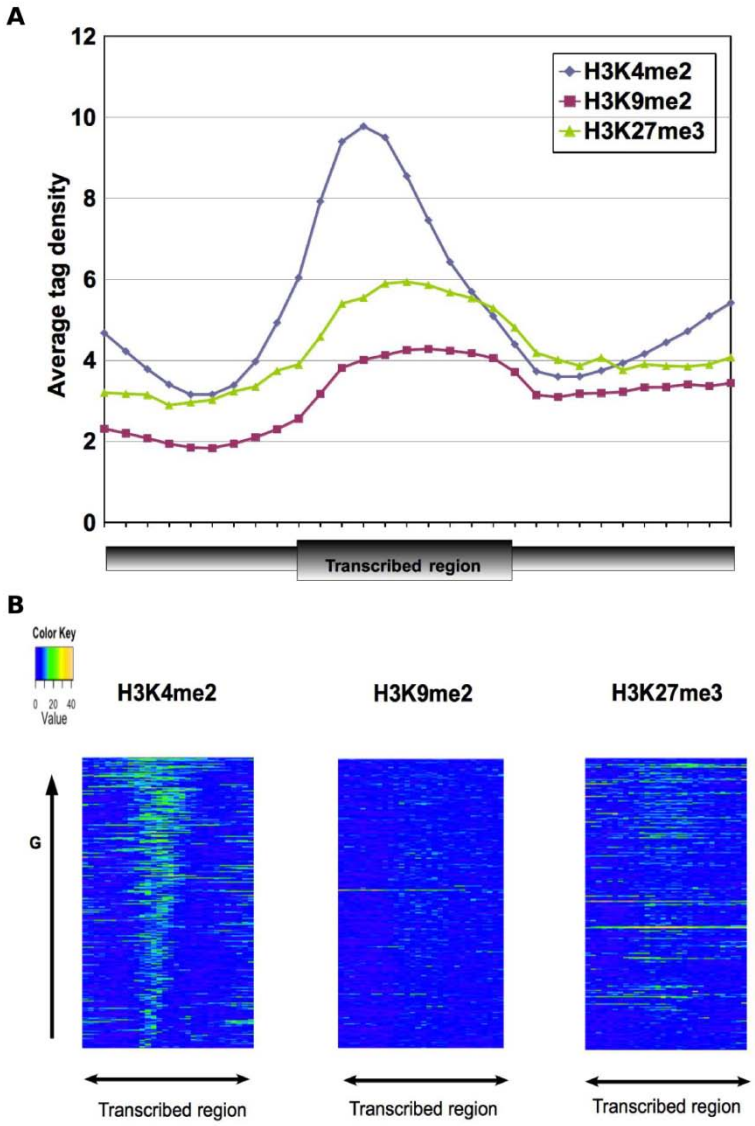


Figure 3.4 A. **Enrichment profile of H3K4me2, H3K9me2 and H3K27me3 along genes (upstream 500bp, coding region, downstream 500bp).** Average tag density is the number of sequence reads per gene. A sharp rise in enrichment is seen in the flanking regions (both at 5' and 3' ends), which indicates genes nearby are marked by H3K4me2. **B. Heat map of enrichment peaks of H3K4me2, H3K9me2 and H3K27me3 within the coding sequence region of different lengths of genes.** G indicates the gene length. Even distributions of reads along the gene start to end were seen in both H3K27me3 and H3K9me2, while H3Kme2 shows a sharp peak towards the 5' end.

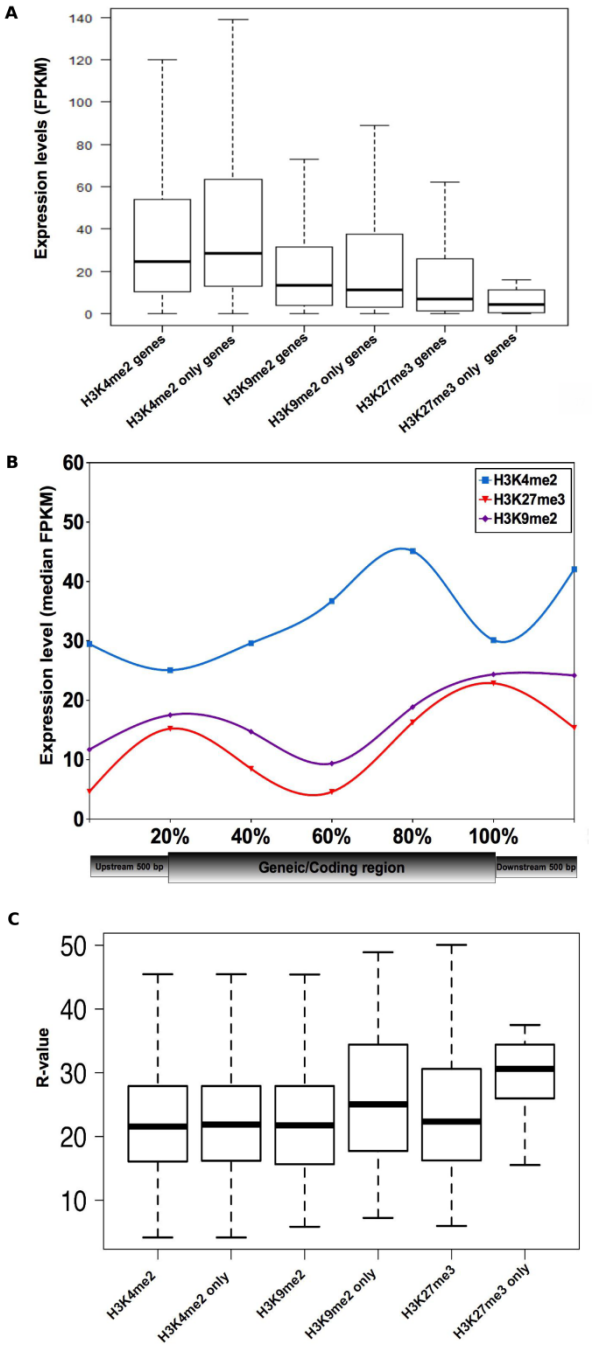


Figure 3.5 Expression profiles of genes marked by H3K4me2, H3K9me2, and H3K27me3. Gene expression was quantified in standard growth conditions using RNA-seq data. About 85 % of genes are expressed and quantified as fragments per kilobase of exons per million reads mapped (FPKM). **A.** Genes that carry H3 K4 dimethylation labels achivrly expressed genes while H3 K9 dimethylation and H3 K27 trimethylation, tend to mark genes with inactive gene expression states. **B.** H3 Lysine-4 dimethylation of genes shows higher

transcriptional activity, particularly genes with marking at the 200 bp mid position of coding region. A similar but opposite trend was seen for genes with H3K9me2 or H3K27me3 modifications. **C.** Boxplots showing differential expression profiles of genes marked by H3K4me2, H3K9me2, and H3K27me3. Genes with R-values below 12 are considered to be constitutively expressed.

To explore whether certain histone modifications correlate with particular gene orthologies, we analyzed the distribution of H3K4me2, H3K9me2 and H3K27me3 marked genes based on the orthology classification described by Bowler et al 2008 and Maheswari et al 2010 (*P. tricornutum* specific, diatom specific, or predicted to have been acquired from bacteria by horizontal gene transfer). Genes marked by these three histone modification marks appeared to be distributed evenly among genes belonging to these different orthologous groups **(Supp Figure 3.2**). This same general phenomenon was observed previously for genes displaying DNA methylation (Chapter II).

To investigate the functions of genes marked with specific histone modifications, we performed a gene ontology (GO) analysis to assess whether they are enriched in certain functional categories. The genes marked by H3K4me2 and one of the other two histone marks (7,896; 81.9% of annotated genes) are enriched in oxidoreductase activity GO category while the genes only marked by H3K4me2 (6,047; 62.8% of annotated genes) are enriched in structural constituent of ribosome GO category. The genes marked by H3K9me2 and one of the other histone marks (2,475; 25.7% of annotated genes) are enriched in hydrolase, ATPase, inorganic cation transmembrane transport, nucleoside tri-phosphatase activity, helicases, and structural constituent of cytoskeleton GO categories, while the functions of the genes only marked by H3K9me2 (218; 2.3% of annotated genes) was found to mark genes with functions similar to that marked by H3K27me3. The genes marked by H3K27me3 (586; 6.1% of annotated genes) and H3K9me2 are enriched in ATP binding, helicase, glucosidase activity while the genes marked by H3K27me3 alone (113; 1.2% of annotated genes) are enriched in protein kinase activity, cAMP dependent protein kinase, phosphotransferase and Diamine N-acetyl transferase GO caterogies (**Supp Figure 3.3**).

Because H3K4me2 is enriched on genes, its genome wide profile can help us to improve gene annotation. The analyses of RNA-seq data and genome wide H3K4me2 distribution have confirmed most of the predicted genes, and also revealed some novel genes. Based on current gene annotations of the *P. tricornutum* genome, 7,896 out of 10,402 genes are marked by H3K4me2, while 1,304 H3K4me2 enriched domains were detected within intergenic regions. The H3K4me2 enriched domains in intergenic regions are potential genes which have been missed by gene prediction. Separately, RNA-seq expression profiling showed that 721 novel

genes appear in the “intergenic regions” (**Supp Figure 3.4A**). We further investigated the novel gene regions that were overlapping between RNA-seq and H3K4me2 enriched genes. Potentially 493 genes were detected and annotated (**Supp Figure 3.4B**).

3.3.3 Genome wide distribution of H3K4me2, H3K9me2 and H3K27me3 on TEs

The *P. tricornutum* TEs contain Class I elements including LTR-RTs (Copia), relics of non LTR-RT retrotransposon-like elements (RTE), and a few copies of Class II transposons including Piggybac, Tpase-like, and MuDR-like elements. Among them, 1,350 TEs (38.7%) and 1,441 TEs (41.3%) were found marked by H3K9me2 and H3K27me3, respectively. Among these marked TEs, most of them belong to Copia type: 1040 Copia TEs and 1,173 Copia TEs were found marked by H3K9me2 and H3K27me3, respectively **(Figure 3.6**). As for H3K4me2, only 605 TEs are marked by it and most of them are simple repeats (n=322). Overall, a significant fraction of potentially active Copia TEs were found associated with H3K9m2 and H3K27me3 which implies that these marks may regulate the activation of TEs, especially Copia type TEs which have recently amplified in the genome in *P. tricornutum* (Maumus et al., 2009a).

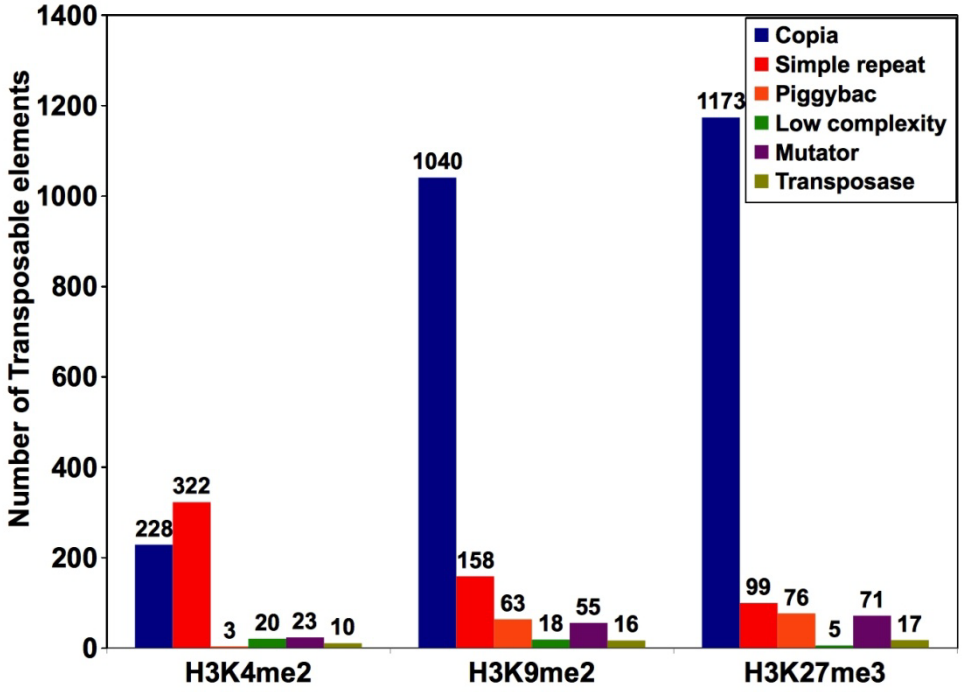


Figure 3.6 Different groups of TEs marked by H3K4me2, H3K9me2 and H3K27me3. About 41% of the transposable elements annotated in *P. tricornutum* are marked by H3K27me3, particularly Copia type elements which are well annotated unlike other TEs. A similar proportion (1,350 TEs, i.e., around 39%) of TE groups is seen in H3K9me2 marked transposable elements.

3.3.4 Combinatorial marking of H3K4me2, H3K9me2 and H3K27me3 on the same genomic regions

The $2^3 = 8$ theoretically possible combinations were all observed in *P. tricornutum* (**Figure 3.7A**). The PCA analysis of histone modification marks show that H3K9me2 and H3K27me3 are highly correlated (**Figure 3.7B**) because the regions marked by them are highly overlapping. Among 865 H3K27me3 marked domains, 749 were found to overlap with 777 H3K9me2 marked domains (the total number of domains marked by H3K9me2 is 1,991). The regions marked by both H3K9me2 and H3K27me3 contain 182 genes and 942 TEs. Most H3K4me2 modified regions are H3K4me2 unique regions (7,862 domains out of 9,748 domains). We also observed that a fraction of H3K9me2 marked genes overlap with H3K4me2: 1,224 H3K4me2 marked domains were found to overlap with 888 H3K9me2 marked regions, comprising 1,260 genes and 43 TEs. Considering the possible cross reactivity of H3K9me2 antibody with H3K9me3 epitopes in *P. tricornutum*, a fraction of the genes marked by H3K9me2 may in fact correspond to marking with H3K9me3. Regions marked by all three H3K4me2, H3K9me2 and H3K27me3 marks were also identified, corresponding to 360 genes and 273 TEs.

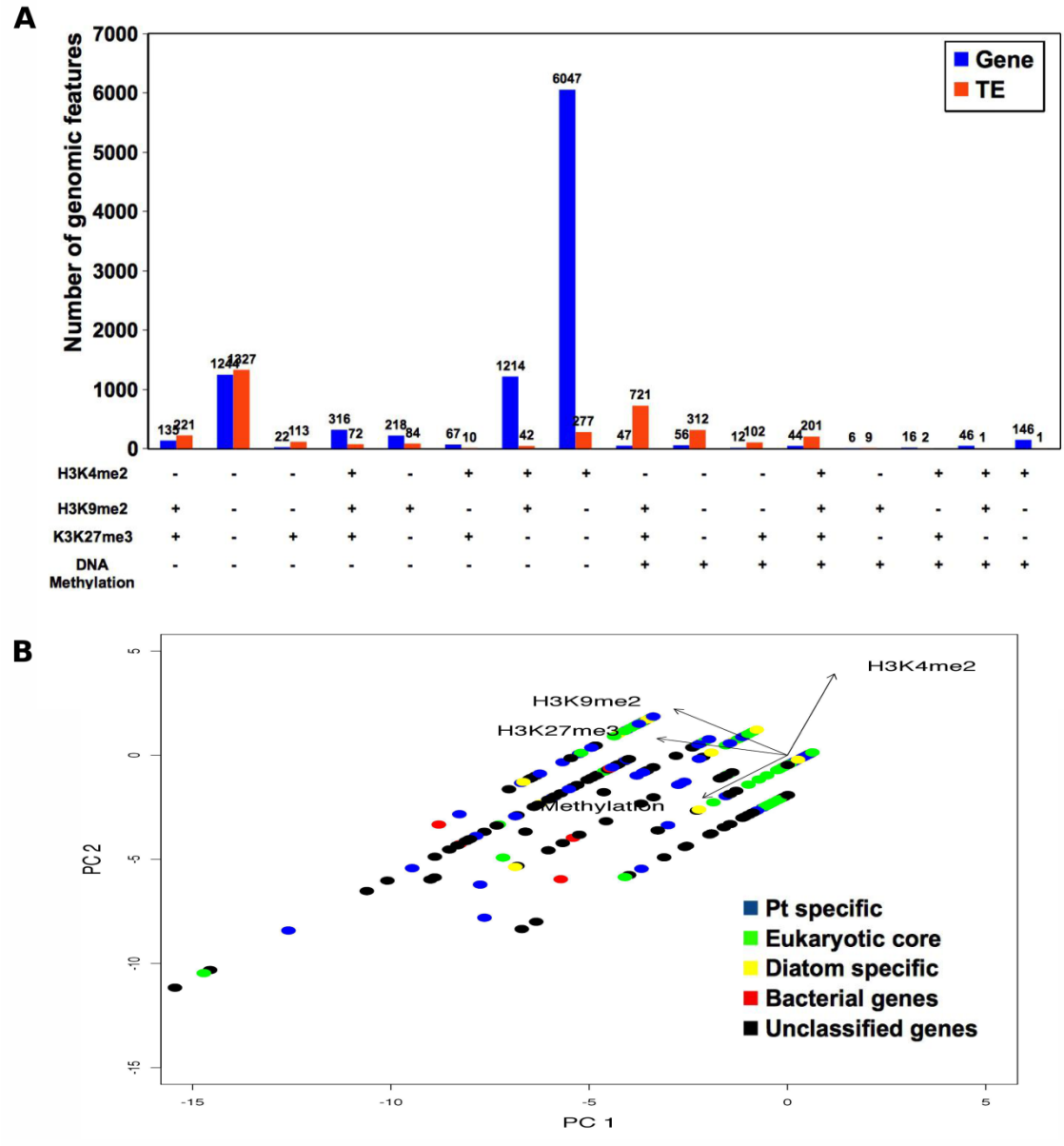


Figure 3.7 Combinatorial localization patterns of the three chromatin marks and DNA methylation. A. A significant number of DNA methylated genomic features are associated with both H3K27me3 and H3K9me2. **B.** PCA analysis showing that H3K9me2 and H3K27me3 co-occur in most genomic features. Also seen in the plot are the orthologous gene groups which do not correlate with any particular histone modification.

3.3.5 Combinatorial effect of histone modifications and DNA methylation on gene expression

The correlation between different histone marks and DNA methylation was also investigated. DNA methylation data was generated by McrBC-ChIP (Chapter II). The regions which were identified as being co-marked by H3K9me2, H3K27me3 and DNA methylation contain 721 TEs (942 TEs were found co-marked by H3K9me2 and H3K27me3) which indicated that the TE regions marked by both H3K9me2 and H3K27me3 also highly overlap with DNA methylation modified regions. Combinatorial analysis of DNA methylation, H3K9me2 and H3K27me3 on TEs also showed that most methylated TEs tend to be co-marked by H3K27me3 and H3K9me2 (**Figure 3.8**). Compared to H3K9me2, DNA methylation is however more associated with H3K27me3. In *A. thaliana*, the correlation of H3K9me2 and DNA methylation in a CHG context has been observed (Bernatavichute, Zhang, Cokus, Pellegrini, & Jacobsen, 2008). Here only 6 genes and 9 TEs were found in the regions marked by DNA methylation and H3K9me2, demonstrating that the direct correlation between H3K9me2 and DNA methylation found in *A. thaliana* does not exist in *P. tricornutum.*

I further investigated the association of DNA methylation patterns on genes and histone modifications. Interestingly, most of the heavily and completely methylated genes were found to be marked by both H3K9me2 and H3K27me3. For the other categories of methylated genes, such a preference was not detected. The methylated genes only marked by H3K27me3 were found to belong to just three categories: 5' end methylated genes, gene body methylated genes (middle of genes) and completely methylated genes, while methylated genes only marked by H3K9me2 were not found (**Figure 3.9**). Only heavily methylated genes tend to be co-marked by H3K27me3 and H3K9me2 while other groups of methylated genes (500bp upstream, 5' end, gene body, 3'end and 500bp downstream) do not display any particular patterns.

For the methylated TEs, it appears that DNA methylation tends to be excluded from simple repeats. Otherwise there is no obvious preference on particular groups of TEs. It is obvious that the cross-talk mechanisms of DNA methylation and H3K27me3 in *P. tricornutum* are different from what have been described in multicellular organisms. Overall, not only methylated TEs but also the heavily methylated genes tend to be co-modified by H3K9me2

and H3K27me3. It is interesting that these heavily methylated genes appear to be important for maintaining the silencing of differentially expressed genes (Chapter II).

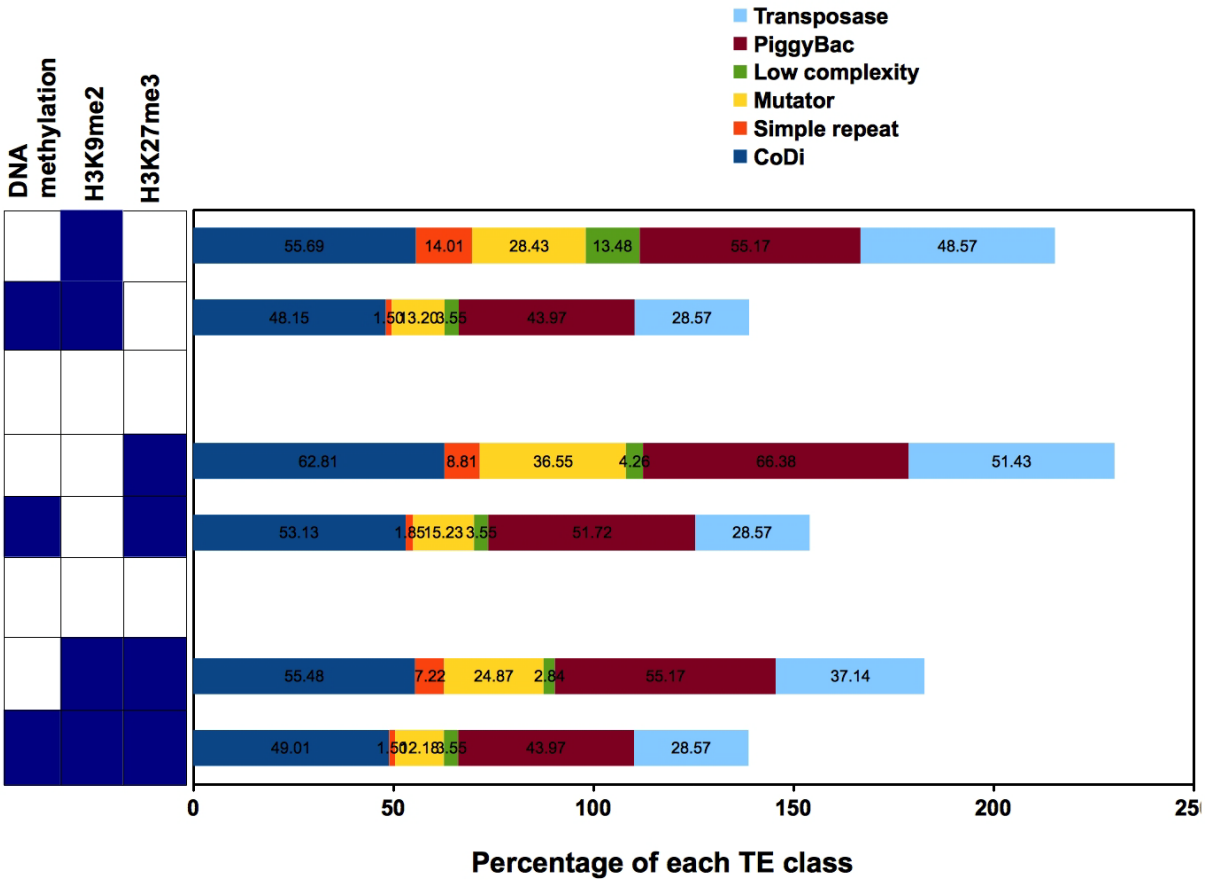


Figure 3.8 Association pattern of DNA methylation and three different histone marks (H3K4me2, H3K9me2 and H3K27me3) on different classes of TEs. Second and fourth rows show a significant level of deposition of H3K9me2 on methylated TEs and H3K27me3 on methylated elements. Tendency of H3K9me2 and H3K27me3 to co-mark on a similar proportion of all groups of methylated TE is clearly seen in the sixth row.

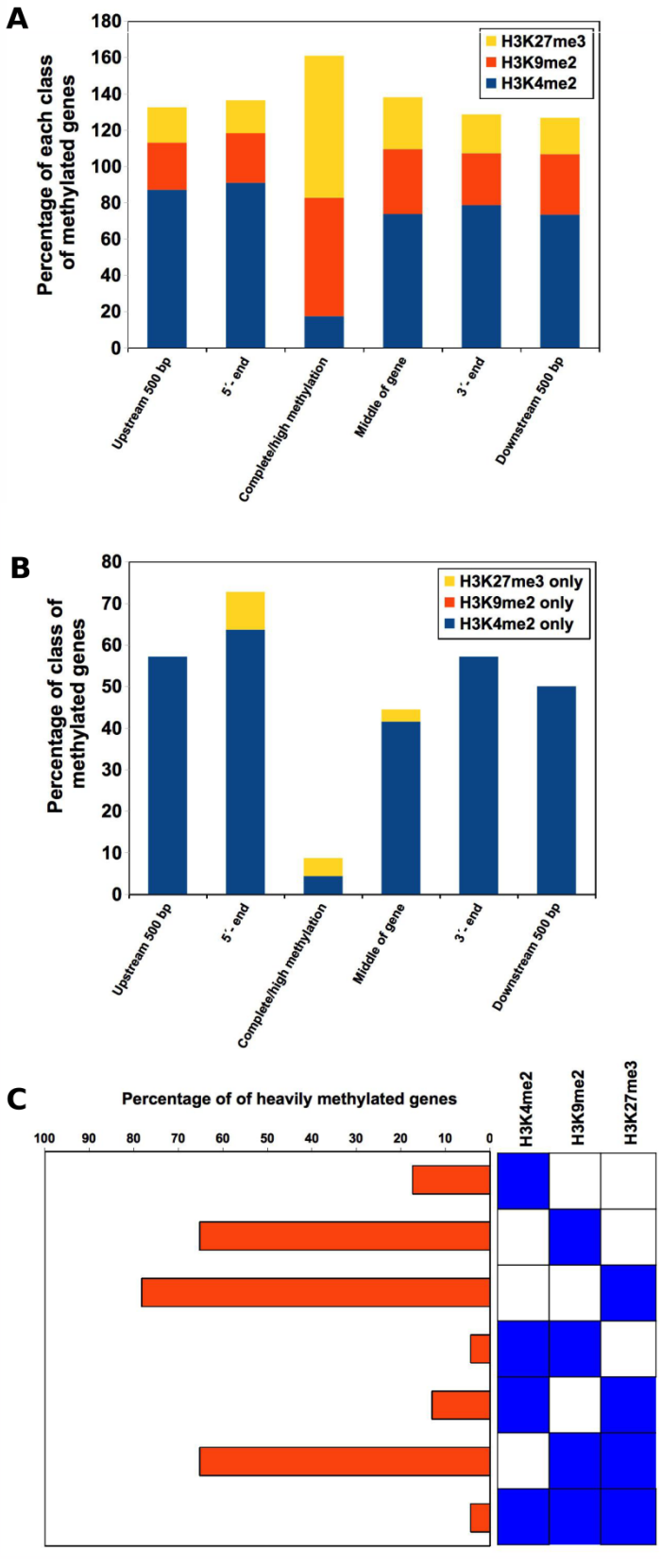

A
Percentage of each class of methylated genes
H3K27me3
H3K9me2
H3K4me2
Upstream 500 bp
5`- end
Complete/high methylation
Middle of gene
3`- end
Downstream 500 bp
B
Percentage of class of methylated genes
H3K27me3 only
H3K9me2 only
H3K4me2 only
C
Percentage of of heavily methylated genes
H3K4me2
H3K9me2
H3K27me3

Figure 3.9 Combination of histone modification marks H3K4me2, H3K9me2 and H3K27me3 on different classes of methylated genes. Methylated genes are grouped according to their localization along the gene or upstream or downstream of the gene. **A.** Heavily methylated genes are co-marked by both H3K9me2 and H3K9me3. **B.** Non-overlapping histone modification pattern showing most methylated genes are co-marked by H3K9me2 and H3K27me3. This trend is observed in other genomic features. **C.** Among the 25 heavily methylated or completely methylated genes, most of them are acquired by H3K9me2 and H3K27me3. These heavily methylated genes are shown correlated to repression of gene expression. A combination of H3K9me2 and H3K9me3 could repress or differentially express such genes. In the first three coloums, when a gene is marked by a histone mark, whether it is also marked by other histone marks is not considered. The last four coloums demonstrate the combination of different histone marks on heavily methylated genes.

3.4 Discussion

The work reorted here has led to the generation of an integrated epigenomic map of three histone methylation marks together with transcriptional outputs in the diatom model species *P. tricornutum.* Combined with previous genome wide DNA methylation data, comprehensive and combinatorial analyses revealed some conserved and some specific epigenetic features in *P. tricornutum* which have not been observed in other organisms.

3.4.1 H3K4me2 mainly marks genes in *P. tricornutum*

In *P. tricornutum,* H3K4me2 was found extensively associated with genes with large range of expression, demonstrating that its distribution pattern is consistent with previous findings in other organisms. Generally, H3K4me2 can be present on active genes as well as on poised and inactive genes and is not influenced by expression level. In *A. thaliana,* H3K4me2 does not appear to index genes in relation to their expression level and may not be directly involved in transcriptional activation (Xiaoyu Zhang et al., 2009). In the yeast *S. cerevisiae,* H3K4me2 is also present on both inactive and active euchromatic genes (Millar & Grunstein, 2006). In human cells, H3K4me2 is associated with genes independently of their expression level suggesting its implication in fine tuning of tissue-specific gene expression (Pekowska et al., 2010).

In *P. tricornutum,* the enrichment of H3K4me2 on genes peaks in the gene body and is slightly skewed towards the transcription start site. Enrichment for H3K4me2 within gene bodies is a characteristic trait of expressed genes in yeast (B. Li, Carey, & Workman, 2007). In *A. thaliana,* H3K4me2 is also enriched within the gene body but the peak of enrichment is very near to the transcription start site (Roudier et al., 2011a). In human CD4+ T cells, five enrichment profiles of H3K4me2 have been identified. The genes with high levels of H3K4me2 within the gene body were characterized as a subset of tissue-specific genes (Pekowska et al., 2010).. The enrichment profile of these genes is therefore similar with that found in *P. tricornutum.*

Because H3K4me2 is the mark for both active and inactive genes, analysis of H3K4me2 modified genomic regions can help us to improve *P. tricornutum* genome annotation. Indeed, analyses of H3K4me2 distribution combined with RNA-seq expression data identified 493

novel genes in *P. tricornutum* which proves that epigenetic studies can be useful for aiding in improvement of genome annotation. 562 genes are not expressed and not marked by H3K4me2. Carrying out genome wide analysis of other histone marks such as H3K36me3, which is well documented to be associated with transcribed regions (Kolasinska-zwierz et al., 2009) can further improve gene annotation in the future. Another good candidate is H3K4me3. It is known that H3K4me3 is a general mark for active genes in plants and mammals accumulating predominantly on promoter regions (Guenther, Levine, Boyer, Jaenisch, & Young, 2007; Xiaoyu Zhang et al., 2009). ChIP-seq with H3K4me3 in *P. tricornutum* will probably bring new insights into the correlation between chromatin states and gene regulation and help to refine genome annotation and gene expression analyses.

3.4.2 H3K9me2 mainly marks TEs in *P. tricornutum*

In *P. tricornutum*, H3K9me2 was found mainly associated with TEs which is consistent with what has been observed in plants and animals where this mark has been profiled. H3K9me2 is a hallmark for repressive heterochromatin mainly enriched in TEs in different organisms (Ernst et al., 2011; Filion et al., 2010; T. Liu et al., 2011; Roudier et al., 2011b). In *P. tricornutum*, the genes marked by H3K9me2 have lower expression levels compared to the ones marked by H3K4me2 which indicates that H3K9me2 is associated with gene silencing, a notion which is also consistent with previous studies (Roudier, Ahmed, Bérard, Sarazin, Mary-Huard, Cortijo, Bouyer, Caillieux, Duvernois-Berthet, Al-Shikhley, Giraut, Després, Drevensek, Barneche, Dèrozier, Brunaud, Aubourg, Schnittger, Bowler, Martin-Magniette, Robin, Caboche, & Colot, 2011b; J. Zhou et al., 2010; J. Zhou et al., 2010). However, the gene repression effect caused by H3K9me2 modification on genes is not as significant as H3K27me3. Furthermore, the profile of H3K9me2 should be taken with caution considering the eventual cross reactivity of the antibody used. Thus, the enrichment of H3K9me2 in this study might be partly due to H3K9me3 modified regions. In *A. thaliana*, H3K9me3 tends to mark highly expressed genes not TEs. H3K9me2 associated genes appear generally to be repressed in *P. tricornutum*. If we believe that a subset of genes marked by H3K9me2 is due to cross activity of H3K9me3 which is associated with active genes like in *A. thaliana*, we therefore cannot explain the fact that H3K9me2 marked genes in our study are moderately repressed genes. In *C. elegans* and human, H3K9me2 and H3K9me3 tend to appear together

in repressive heterochromatic regions (T. Liu et al., 2011; Wang et al., 2009). In our study, the regions marked by H3K9me2 do not contain a large percentage of genes which implies that the distribution of H3K9me2 and H3K9me3 in *P. tricornutum* may resemble that in animals. The observed overlap in the mapping of the two marks, presumably due to a spatial proximity of the two epitopes in the *P. tricornutum* genome, needs to be further investigated to distinguish the regions that are truly marked by one or the other modifications. It is however important to keep in mind the possible occurrence of two marks on the same locus or region. The above hypothesis has to be proven by carrying out H3K9me3 ChIP-seq in *P. tricornutum.*

Based on the current genome assembly, there are 33 chromosomes or scaffolds that are identified in *P. tricornutum.* However, among 33 chromosomes only 12 are "real chromosomes" containing two telomeric regions at either end. Based on known features of centromeric regions in other organisms such as characteristic sequences and motifs, the centromeres could not be identified in *P. tricornutum*. TEs are almost randomly distributed and not concentrated in particular segments of the chromosome which would have given an indication for the presence of centromeres. Research in yeast also showed that H3K9me2 is enriched in peri-centromeric heterochromatic regions but absent in centromeres (Sinha et al., 2006). In *A. thaliana*, H3K9me2 was found to be localized in the heterochromatic knob region of chromosome 4 (Lippman et al, 2004) and it showed a higher than average level in pericentromeric/centromeric regions (Bernatavichute et al., 2008). If this is also the case in *P. tricornutum*, it can help us to identify peri-centromeric and heterochromatic regions in *P. tricornutum.* However, this study has shown that H3K9me2 enriched regions are almost randomly distributed along chromosomes (data not shown).

3.4.3 The distribution of H3K27me3 in *P. tricornutum* is unorthodox

H3K27me3 has been found associated with repressed genes and it plays a key role in cell differentiation and development in both plants and animals (Roh, Cuddapah, Cui, & Zhao, 2006; Xiaoyu Zhang et al., 2007a; X. D. Zhao et al., 2007). In higher multicellular organisms, H3K27me3 was found to target mainly genic regions and its presence is mainly associated with gene silencing. In *A. thaliana* seedlings, more than 4,000 H3K27me3 targeted regions are largely protein-coding genes, TEs being mostly excluded. In *D. melanogaster*, *C. elegans*

and human cells, H3K27me3 is also found associated with repressed genes (Filion et al., 2010; T. Liu et al., 2011; Roh et al., 2006).

To our surprise, in *P. tricornutum* H3K27me3 predominantly deposits on TEs not genes. The distribution of H3K27me3 in *P. tricornutum* therefore implies that the functions and mechanisms of H3K27me3 in single celled diatoms may be different from multicellular organisms. It is worth noting that a recent study showed that a high number of TEs were found targeted by 15% of H3K27me3 in meristem cells which were not detected in leaves. Furthermore, it was shown that H3K27me3 instead of DNA methylation regulates the activation of these TEs in meristematic cells (Lafos et al., 2011). This result suggests the functions of H3K27me3 are not confined to genes but also affect TEs. Since most H3K27me3 targets TEs in *P. tricornutum*, we can assume that it has a major function for TE regulation in this organism. In the future, more genome wide profiles of H3K27me3 from other unicellular organisms will help better understand its functions in single celled organisms.

Previously, it was believed that H3K27me3 does not exist in unicellular organisms because H3K27me3 and its related proteins, Polycomb group proteins (PcG), were not detected in model unicellular organisms such as budding yeast *S. cerevisiae* and *Schizosaccharomyces pombe.* PcG represented by repressive complex 1 (PRC1) and 2 (PRC2) were initially discovered in *D. melanogaster* (Raphaël Margueron & Reinberg, 2011). PcG can catalyze H3 methylation and generate both H3K27me2 and H3K27me3. Together with H3K27me3 deposition, PcG plays vital roles in multicelluar development, stem cell biology and cancer (Jaenisch & Young, 2008; Jones & Baylin, 2007; Lechner, Boshoff, & Beck, 2010; Rajasekhar & Begemann, 2007). Analysis of the distribution of E(z) and other polycomb 2 (PRC2) components in eukaryotic organisms demonstrated that PRC2 is present in various unicellular eukaryotes including another diatom species *Thalasiossira pseudonana* (Shaver, Casas-Mollano, Cerny, & Cerutti, 2010). In *P. tricornutum* the putative E(z) protein likely to be responsible for catalyzing H3K27me3 could be found (Pt32817) as well as other putative PRC2 components. This provides the genetic support for the existence of H3K27me3 in *P. tricornutum.* It further supports the hypothesis that H3K27me3 existed in the last common unicellular ancestor, but was lost at certain times during eukaryotic evolution, as exemplified

by the cases of *S. pombe* and *S. cerevisiae* (Raphaël Margueron & Reinberg, 2011; Shaver et al., 2010).

Like H3 lysine 4 and lysine9, lysine 27 can be either mono-, di- or tri-methylated. These three histone states (H3K27me1, H3K27me2 and H3K27me3) have distinct distributions and impacts on transcriptional activity. In *A. thaliana*, H3K27me1 is prevalent over silent TEs in pericentromeric regions, where it is thought to prevent over-replication while H3K27me2 is associated with H3K27me3 depositing on repressed genes as well as on TEs (Jacob, Stroud, et al., 2010; Jacob, Feng, et al., 2010; Roudier et al., 2011b). In *Drosophila*, H3K27me1 and H3K27me2 are associated with euchromatin and heterochromatin. Here the question is whether a big proportion of TEs marked by H3K27me3 in my study is due to the H3K27me1 and H3K27me2 epitopes affinity for the H3K27me3 antibody. The peptide competition assays demonstrated that the H3K27me3 antibody I used has no cross-reactivity with H3K27me1 and H3K27me2 epitopes in *P. tricornutum* which indicates that the majority of TEs marked by H3K27me3 are not likely due to cross-reactivity of H3K27me1 and H3K27me2. Furthermore, H3K27me3 antibody doesn't have affinity to H3K27me1 peptide as well as H3K27me2 peptide *in vitro*. Though it is unlikely that H3K27me3 antibody can bind H3K27me1 and H3K27me2 epitopes in *P. tricornutum* even in the absence of H3K27me3 modification, it is still necessary to analyze histone modifications in *P. tircornutum* by mass spectrometry to resolve this issue definitively. This work will provide information about the number of existing histone modifications and their proportions in *P. tricornutum*and will pave the way for future epigenetic research in *P. tricornutum*..

On the other hand, even though H3K27me3 mainly marks TEs, the H3K27me3 modified regions over large domains resemble the enriched profiles of H3K27me3 in animals. H3K27me3 modified regions typically form broad domains including over multiple genes (>5kb) in animals (Bernstein et al., 2006; Pauler et al., 2009b; Young et al., 2011). Recently, another two types of enrichment profiles have been identified in Mouse C57BL/6 Bruce 4 (B4) ES cells (mECSs): a peak of enrichment around the transcription start site (TSS) is commonly associated with 'bivalent' genes, where H3K4me3 also marks the TSS, and another surprising enrichment profile with a peak in the promoter of genes that is associated with active transcription (Young et al., 2011). In plants, the mark covers much shorter regions (typically

<1kb) which tend to be restricted to the coding regions of single genes (Elling & Deng, 2009; Xiaoyu Zhang et al., 2007b). Taken together, these observations suggest that the establishment, spreading and maintenance of H3K27me3 are diverse in plants and animals. For *P. tricornutum,* probably the H3K27me3 deposition and spreading mechanisms on TEs resemble those in animals on genes because both of them are over large domains.

Although H3K27me3 marked regions are mainly restricted to TEs, there is still a significant subset of genes that are marked by H3K27me3. As expected, such genes have lower expression levels. In particular, genes that are only marked by H3K27me3 have the lowest expression levels compared to other groups, suggesting that H3K27me3 has a similar function on gene repression in *P. tricornutum* as it does in animals and plants (Deal & Henikoff, 2010; T. Liu et al., 2011; Oh, Park, & van Nocker, 2008; Turck et al., 2007; Xiaoyu Zhang et al., 2007b). We also found genes marked by H3K27me3 with a tendency to be differentially expressed under specific conditions. In other words, H3K27me3 marked genes are more responsive to external stimuli compared to other genes. In *A. thaliana,* it was also demonstrated that H3K27me3 is dynamically regulated in response to developmental or environmental cues: H3K27me3 targeted genes in seedlings are enriched for genes with tissue-specific expression or induced by abiotic and biotic stresses (Lafos et al., 2011; Xiaoyu Zhang et al., 2007b). It is striking that a large number of developmentally important transcription factor genes in *A. thaliana* were found marked by H3K27me3. According to the GO analyses, the genes marked only by H3K27me3 in *P. tricornutum* are enriched in Diamine N-acetyl transferase, Protein kinase, cAMP dependent Protein kinase and Phosphotransferase which are involved in signal transduction, biosynthesis and amino metabolism. It is very interesting that Diamine N-acetyl transferase is involved in urea cycle which was initially considered only to exist in metazoans but was surprisingly found in diatoms (Allen et al., 2011; Armbrust et al., 2004).

In *D. melanogaster* and mammals, H3K27me3 locates to facultative heterochromatin (Bannister & Kouzarides, 2011; Brinkman et al., 2006; Lin et al., 2011) while in *A. thaliana,* rice and maize it is associated with repressed genes in euchromatic regions (Cheutin & Cavalli, 2012; Shi & Dawe, 2006; Turck et al., 2007). Although H3K27me3 locates differently in heterochromatin and euchromatin in metazoans and plants, respectively, its role

as a dynamic regulator of cell differentiation and development is conserved in these multicellular organisms. In *Drosophila* and *C. elegans*, the correlation of H3K27me3 dynamic changes and development was revealed (Oktaba et al., 2008; Schuettengruber et al., 2009; Yuzyuk, Fakhouri, Kiefer, & Mango, 2009). H3K27me3 was also found to be involved in mammalian embryonic stem cells differentiation (Hawkins et al., 2010; Marks et al., 2012; Mikkelsen et al., 2007). In *Xenopus* embryos, it was found predominantly deposited upon subsequent spatial restriction of repression of transcriptional regulators (Akkers et al., 2009). In *A. thaliana*, it was also demonstrated that protein coding genes that showed strong expression differences between meristem and leaf were enriched for H3K27me3 (Lafos et al., 2011). All these studies suggest conserved mechanisms wherein H3K27me3 confers developmental dynamics to gene expression patterns.

The roles of H3K27me3 in unicellular organisms have not been explored extensively yet. The only research on unicellular organisms is a report of immunofluorescence staining with an H3K27me3 antibody in *Tetrahymena* where it was shown that H3K27me3 is associated with all three heterochromatic structures (Y. Liu et al., 2007). In another unicellular alga *Chlamydomonas reinhardtii*, analysis of histone modifications by mass spectrometry has been carried out. H3K27me1 and H3K27me2 were detected whereas it was difficult to confirm the existence of H3K27me3 in *C. reinhardtii* because of the difficulty in distinguishing H3K27ace and H3K27me3. The *C. reinhardtii* H3 sequences around “K27” position are not conserved compared to human and *A. thaliana* so the commercial H3K27me3 antibody is not applicable in *C. reinhardtii*. In this case, the existence of H3K27me3 also cannot be confirmed by western blotting and ChIP experiments in *C. reinhardtii* (Shaver et al., 2010). Thus the roles of H3K27me3 in unicellular organisms needs to be further explored. The ChIP-seq with H3K27me3 in *P. tricornutum* reported here is the first genome wide H3K27me3 distribution study in a unicellular organism. Based on the fact that H3K27me3 is associated with repressed genes and TEs, it is tempting to speculate that H3K27me3 can define a facultative heterochromatin state in *P. tricornutum*. All together H3K27me3 has a novel distribution in unicellular organisms and probably has distinct functions from animals and plants.

3.4.4 Methylated TEs and heavily methylated genes tend to be co-marked by H3K27me3 and H3K9me2

In *P. tricornutum*, we surprisingly found that H3K27me3 and H3K9me2 tend to co-mark heavily methylated genes and TEs. It is worth noting that only heavily methylated genes, not other groups of methylated genes, tend to be co-marked by H3K27me3 and H3K9me2. For the methylated TEs, it appears that all groups of methylated TEs except simple repeats tend to be co-marked by H3K9me2 and H3K27me3. The combination of DNA methylation, H3K9me2 and H3K27me3 on heavily methylated genes and methylated TEs is due to the tendency of co marking effect of H3K27me3 and H3K9me2 because DNA methylation and H3K9me2 alone do not show the tendency to appear in the same location (neither on genes nor TEs) in *P. tricornutum*.

In muticellular organisms, DNA methylation and H3K27me3 are usually antagonistic on genes. It was found that in postnatal neural stem cells (NSCs) in mouse, *de novo* DNA methylation at non promoter regions by Dnmt3a is required to counteract PcG repression. In other words, *de novo* DNA methylation and H3K27me3 are antagonists in non promoter regions (Wu et al., 2010). For the promoter regions of mammalian stem cells, H3K27me3 is associated with hypomethylated DNA (Brunner et al., 2009; Hawkins et al., 2010; Mikkelsen et al., 2007). As for plants, H3K27me3 and 5mC are mutually exclusive in seedlings of *A. thaliana* and rice (X. Li et al., 2008; Mathieu, Probst, & Paszkowski, 2005; Weinhofer, Hehenberger, Roszak, Hennig, & Ko, 2010). In *Xenopus* embryos DNA methylation is also absent in large H3K27me3 domains (Bogdanovic et al., 2011). It appears that the antagonism of H3K27me3 and DNA methylation plays a key role in cell differentiation (Meissner, 2010; Mikkelsen et al., 2007; Mohn et al., 2008). In mouse embryonic stem cells (ESCs), most H3K27me3 marked regions are also co-marked by the active mark H3K4me3. As differentiation proceeds, genes become enriched with either of the two opposing marks. It turns out that most genes, which are repressed and not activated during lineage commitment, are *de novo* DNA methylated during differentiation.

The antagonism of H3K27me3 and DNA methylation also seems to involve imprinting. At the imprinted locus *Rasgrf1* in mouse, DNA methylation was found on the paternal allele and H3K27me3 was on the maternal allele. These two marks are mutually exclusive and

antagonize by blocking the placement of the other (Lindroth et al., 2008). In *P. tricornutum*, DNA methylation and H3K27me3 tend to comark the same regions so it is possible that these two marks antagonize each other to regulate allele specific expression. Lines in which DNA methyltransferses and Ez (which is responsible for H3K27me3 deposition) have been inactivated could be used to test this hypothesis. In mutant strains the status of H3K27me3 and DNA methylation of several loci which are marked by these two marks in wild type can be checked. In this way, whether DNA methylation and H3K27me3 affect each other can be revealed. If it is the case cooperation between these two marks, it is worth to be further explored.

The positive correlation of H3K27me3 and DNA methylation was detected in the promoter regions of cancer-specific methylated genes in tumor cells (Easwaran et al., 2012; Ohm et al., 2007; Schlesinger et al., 2007). In human ESCs, H3K27me3 is also positively associated with DNA methylation across a broad range of mCG outside of promoter regions (Hawkins et al., 2010).

In *P. tricornutum* the correlation of H3K9me2 and H3K27me3 on the same regions is also unusual. In *A. thaliana*, *C. elegans*, and *Drosophila*, the regions marked by H3K9me2 and H3K27me3 tend to be dissimilar (Filion et al., 2010; Guo, Zhou, Elling, Charron, & Deng, 2008; T. Liu et al., 2011; J. Zhou et al., 2010). The cross-talk between H3K9me2 and H3K27me3 has not been reported in other organisms yet, and so its function in *P. tricornutum* will be important to resolve.

The direct correlation between H3K9 methyltransferases (HKMT) and DNA methyltransferases (DNMTs) has been detected in fungi, plants and animals (Jackson et al., 2002; Lehnertz et al., 2003; Tamaru & Selker, 2001). In *A. thaliana*, small double-stranded RNAs (dsRNA's) are processed to guide methylation to complementary DNA loci by the RdDM (RNA-directed DNA methylation) pathway (Mahfouz, 2010). These dsRNAs are then processed to direct histone 3 lysine 9(H3K9) methylation via Ago4 and the SUVH (Suppressor of Variegation Homolog) histone methyltransferase family (Enke, Dong, & Bender, 2011). This H3K9 dimethylation is then putatively bound by the cytosine methyltransferase CMT3, which methylates CHG methylation (Cao et al., 2003). A high correlation between H3K9m2 and CHG DNA methylation was detected in *A. thaliana*. The

coding regions of genes that are associated exclusively with methylation in a CG context did not contain H3K9m2 (Bernatavichute et al., 2008). However, in *P. tricornutum*, H3K9me2 does not have a direct correlation with DNA methylation, and it seems rather that H3K27me3 connects H3K9me2 and DNA methylation. The crosstalk between H3K27me3, H3K9me2 and DNA methylation is therefore likely to be different from what is known in plant and animal systems.

It is interesting that most of the heavily and completely methylated genes are co marked by H3K9me2 and H3K27me3. Among them, some genes have no expression under normal conditions but are expressed under other specific conditions. For example, gene Pt12452, encoding a putative plasma membrane hydrogen ATPase, which is heavily methylated and co-marked by H3K9me2 and H3K27me3, is not expressed under normal conditions but is specifically induced when *P. tricornutum* cells switch from fusiform to triradiate morphotype (**Figure 3.3**). Gene Pt12452 may therefore encode a protein crucial for the morphotype transition and it is possible that the cooperative regulation of H3K27me3, H3K9me2 and extensive DNA methylation is also implicated in it. So it will be very interesting to check the status of H3K27me3, H3K9me2 and DNA methylation of gene Pt12452 of *P. tricornutum* in triradiate morphytype cells.

3.5 Perspectives

The comparison of DNA methylation and histone modification profiles and further combinatorial studies with gene expression and small RNAs at a genome wide scale between *P. tricornutum*, animal and plant model systems will help us better understand the epigenetic components in an evolutionary view and evaluate the impact of epigenetic components studied in this work on adaptive evolutionary processes. Furthermore, *P. tricornutum* can transit among four different forms: fusiform, oval, round, and triradiate. Interestingly, specific environmental cues can trigger these morphotype transitions (De Martino et al., 2011). Therefore, *P. tricornutum* appears as an excellent model to study the correlation between epigenomic dynamics and morphological variations in response to external stimuli.

Besides histone methylation, other histone modifications such as acetylation and phosphorylation also have great impacts on chromatin and further influence the transcriptional activity of genes (Bannister & Kouzarides, 2011). In the future, genome wide distribution of

more histone marks should be generated. A more comprehensive mapping of the combinations of different histone marks, DNA methylation and small RNAs profiles will draw a chromatin landscape on the whole genome, and may define structural and functional domains on chromosomes such as centromeres. Because diatoms are so successful in ocean environments, it is possible that epigenetic regulation, which is reversible and more flexible compared to genetic regulation, plays a significant role in their adaptation to different environments. It will be very intriguing to compare epigenomes under different stress conditions such as high CO_2 and nitrate or iron depletion with normal conditions. In this way, we can elucidate the dynamics of different epigenetic marks under different conditions. The comparison of diatom chromatin landscapes with those of other model species will shed light on the mechanisms underlying the ecological success of this evolutionary distinct eukaryotic group in the context of epigenetic regulation.

3.6 Material and Methods

3.6.1 Growth conditions

Cultures of *P. tricornutum* Bohlin clone Pt1 8.6 (CCMP2561) were obtained from the Provasoli-Guillard National Center for Culture of Marine Phytoplankton, Bigelow Laboratory for Ocean Sciences, USA. Cultures were grown in artificial sea water medium (Vartanian, M., et al 2009). Cultures were incubated at 19°C under cool white fluorescent lights at approximately 75 μmol.m-2.s-1 in 12h light: 12h dark conditions and maintained in exponential phase in semi continuous batch cultures.

3.6.2 Peptide competition assay for antibody specificity test

The peptide competition assay (PCA) is a recommended procedure for confirming the specific band reactivity of an antibody. Antibody is pre-incubated with the peptide prior to use in immunoblotting assays. Different amounts (depends on different peptides and antibody) of peptides were added to a 10 ml BSA solution containing antibody and incubated under light agitation for 4 h at room temperature and additional 1 hour at 4 degree before immunoblotting assays. The antibodies and peptides used in this work include H3K4me2 (Millipore Ref: 07-030), H3K9me2 (Millipore, Ref: 17-681), H3K27me3 (Millipore, Ref: 07-449), H3K4me1 peptide (Abcam Ref: ab1340), H3K4me2 peptide (Abcam Ref: ab7768), H3K4me3 peptide

(Abcam Ref: ab1342), H3K9me1 peptide (Millipore, Ref: 12-569), H3K9me2 peptide (Millipore, Ref: 12-430), H3K9me3 peptide (Millipore, Ref**:** 12-568), H3K27me1 peptide (Millipore, Ref: 12-567), H3K27me2 peptide (Millipore, Ref: 12-566) and H3K27me3 peptide (Millipore, Ref: 12-565). Different concentrations of antibody and peptide concentration were tried. The nuclear enriched protein used for immunoblotting was extracted following chromatin extraction protocol with minor modifications: the culture was not fixed by formaldehyde and sonication was not needed.

3.6.3 Chromatin immunoprecipitation and sequencing

The culture at exponential phase was fixed with formaldehyde (the final concentration in culture is 1%). The fixation was stopped by adding glycine into the culture (the final concentration in culture is 0.125M) followed by 5min of shaking. The cells were washed with PBS solution by centrifugation at 4 °C. The chromatin extraction and immunoprecipitation were carried out following the protocol modified from *Arabidsopsis thaliana* chromatin protocol (see Annexes). H3K4me2 (Millipore Ref: 07-030), H3K9me2 (Millipore, Ref: 17-681) and H3K27me3 (Millipore, Ref: 07-449) were used for immunoprecipitation. Before sending samples for sequencing, quantitative PCR on the DNA recovered from immunoprecipitation and input DNA were conducted to guarantee the quality of DNA from immunoprecipitation. The calculation and analyses of ChIP- qPCR were done following protocol described in material and methods section. DNA recovered from immunoprecipitation by three histone modification antibodies and input DNA (chromatin "input" without immunoprecipitation) was sent for Illumina sequencing by Beckman Coulter Genomics. Sequencing was performed with a read length of 36 bp and sequencing coverage of 5.6-5.9G. Annex part of this chapter shows the detailed protocol.

3.6.4 RNA sequencing

P. tricornutum clone Pt1 8.6 cells were harvested at exponential phase and total RNA was extracted by Trizol. The culture for RNA extraction is the same batch for chromatin extraction and immunoprecipitation. After being treated by Dnase (Invitrogen), 1 µg of total RNA was used for first strand cDNA synthesis followed by double strand cDNA using Mint Universal Kit from Evrogen (SK002). The quality of double stranded DNA was monitored using

agarose gels. cDNA was used for non directional cloning and cDNA library was constructed for Illumina sequencing by Beckman Coulter Genomics. Sequencing was performed with a read length of 75 bp and sequencing coverage of 1.5 Gb.

3.6.5 Computational analysis of histone modifications *P. tricornutum* by ChIP-sequencing

Single-end sequencing of the three ChIP samples and Input in Illumina GAIIx with a read length of 36 bp, yielded an average of approximately 37 million reads (H3K4me2:37,546,690; H3K9me2:36,329,249; H3K27me3:35,973,625, Input:38,202,621). Reads obtained were quality controlled with standardized procedure using FASTQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/). Data for all three ChIP samples and Input were of good quality with mean quality score of 39, with 50 % mean GC content and high duplication. We studied the distribution of these three histone modifications on genes and TEs based on current genome annotation (http://genome.jgi-psf.org/Phatr2/Phatr2.home.html). The reads were mapped onto the *P. tricornutum* genome V2.0 using Bowtie with mismatch permission of 2 bp along the reads. Unique mapping of reads resulted in 62%, 43%, 52%, 84% for H3K4me2, H3K9me2, H3K27me3 and Input respectively. Even distribution of reads in the genomic region was found. To identify regions that were significantly enriched, we used MACS and SICER (http://home.gwu.edu/~wpeng/Software.htm). MACS was designed to detect sharp peaks and SICER works well on mapped reads with diffused peaks. MACS and SICER shown peaks with similar peak-range with comparable overlapped genomic region. But for H3K9me2 and H3K27me3, MACS showed very fewer peaks and in these two cases SICER showed significant diffused peaks which overlaps on repeat regions. Enriched regions from SICER were adopted for further analysis. Visualizing and analysis of genome-wide enrichment profile were done with IGV. Peaks annotation on genes and transposons were performed using PeakAnalyzer . Finally, GO based functional analysis on enriched genes were performed using BLAST2GO with a significant False Discovery Rate (FDR) cut-off of 0.05% probability level. R and python were extensively used for data analysis. Results of the analysis were made available on Gbrowse based genome browser at: http://ptepi.biologie.ens.fr/ptgb.html.

3.7 References

Akkers, R. C., van Heeringen, S. J., Jacobi, U. G., Janssen-Megens, E. M., Françoijs, K.-J., Stunnenberg, H. G., & Veenstra, G. J. C. (2009). A hierarchy of H3K4me3 and H3K27me3 acquisition in spatial gene regulation in Xenopus embryos. Developmental cell, 17(3), 425–34. doi:10.1016/j.devcel.2009.08.005

Allen, A. E., Dupont, C. L., Oborník, M., Horák, A., Nunes-Nesi, A., McCrow, J. P., Zheng, H., et al. (2011). Evolution and metabolic significance of the urea cycle in photosynthetic diatoms. *Nature*, *473*(7346), 203–7. doi:10.1038/nature10074

Armbrust, E. V., Berges, J. a, Bowler, C., Green, B. R., Martinez, D., Putnam, N. H., Zhou, S., et al. (2004). The genome of the diatom Thalassiosira pseudonana: ecology, evolution, and metabolism. *Science (New York, N.Y.)*, *306*(5693), 79–86. doi:10.1126/science.1101156

Bannister, A. J., & Kouzarides, T. (2011). Regulation of chromatin by histone modifications. *Cell research*, *21*(3), 381–95. doi:10.1038/cr.2011.22

Bernatavichute, Y. V., Zhang, X., Cokus, S., Pellegrini, M., & Jacobsen, S. E. (2008). Genome-wide association of histone H3 lysine nine methylation with CHG DNA methylation in Arabidopsis thaliana. *PloS one*, *3*(9), e3156. doi:10.1371/journal.pone.0003156

Bernstein, B. E., Meissner, A., & Lander, E. S. (2007). The mammalian epigenome. *Cell*, *128*(4), 669–81. doi:10.1016/j.cell.2007.01.033

Bernstein, B. E., Mikkelsen, T. S., Xie, X., Kamal, M., Huebert, D. J., Cuff, J., Fry, B., et al. (2006). A bivalent chromatin structure marks key developmental genes in embryonic stem cells. *Cell*, *125*(2), 315–26. doi:10.1016/j.cell.2006.02.041

Bessler, J. B., Andersen, E. C., & Villeneuve, A. M. (2010). Differential localization and independent acquisition of the H3K9me2 and H3K9me3 chromatin modifications in the Caenorhabditis elegans adult germ line. *PLoS genetics*, *6*(1), e1000830. doi:10.1371/journal.pgen.1000830

Bogdanovic, O., Long, S. W., van Heeringen, S. J., Brinkman, A. B., Gómez-Skarmeta, J. L., Stunnenberg, H. G., Jones, P. L., et al. (2011). Temporal uncoupling of the DNA methylome and transcriptional repression during embryogenesis. *Genome research*, *21*(8), 1313–27. doi:10.1101/gr.114843.110

Bowler, C., Allen, A. E., Badger, J. H., Grimwood, J., Jabbari, K., Kuo, A., Maheswari, U., et al. (2008). The Phaeodactylum genome reveals the evolutionary history of diatom genomes. *Nature*, *456*(7219), 239–44. doi:10.1038/nature07410

Brinkman, A. B., Roelofsen, T., Pennings, S. W. C., Martens, J. H. a, Jenuwein, T., & Stunnenberg, H. G. (2006). Histone modification patterns associated with the human X chromosome. *EMBO reports*, *7*(6), 628–34. doi:10.1038/sj.embor.7400686

Brunner, A. L., Johnson, D. S., Kim, S. W., Valouev, A., Reddy, T. E., Neff, N. F., Anton, E., et al. (2009). Distinct DNA methylation patterns characterize differentiated human embryonic stem cells and developing human fetal liver. *Genome research*, *19*(6), 1044–56. doi:10.1101/gr.088773.108

Brykczynska, U., Hisano, M., Erkek, S., Ramos, L., Oakeley, E. J., Roloff, T. C., Beisel, C., et al. (2010). Repressive and active histone methylation mark distinct promoters in human and mouse spermatozoa. *Nature structural & molecular biology*, *17*(6), 679–87. doi:10.1038/nsmb.1821

Cao, X., Aufsatz, W., Zilberman, D., Mette, M. F., Huang, M. S., Matzke, M., & Jacobsen, S. E. (2003). Role of the DRM and CMT3 Methyltransferases in RNA-Directed DNA Methylation. *Current Biology*, *13*(24), 2212–2217. doi:10.1016/j.cub.2003.11.052

Cheutin, T., & Cavalli, G. (2012). Progressive polycomb assembly on H3K27me3 compartments generates polycomb bodies with developmentally regulated motion. *PLoS genetics*, *8*(1), e1002465. doi:10.1371/journal.pgen.1002465

De Martino, A., Bartual, A., Willis, A., Meichenin, A., Villazán, B., Maheswari, U., & Bowler, C. (2011). Physiological and molecular evidence that environmental changes elicit

morphological interconversion in the model diatom Phaeodactylum tricornutum. *Protist*, *162*(3), 462–81. doi:10.1016/j.protis.2011.02.002

De Riso, V., Raniello, R., Maumus, F., Rogato, A., Bowler, C., & Falciatore, A. (2009). Gene silencing in the marine diatom Phaeodactylum tricornutum. *Nucleic acids research*, *37*(14), e96. doi:10.1093/nar/gkp448

Deal, R. B., & Henikoff, S. (2010). A simple method for gene expression and chromatin profiling of individual cell types within a tissue. *Developmental cell*, *18*(6), 1030–40. doi:10.1016/j.devcel.2010.05.013

Easwaran, H., Johnstone, S., Vanneste, L., Ohm, J., Mosbruger, T., Wang, Q., Aryee, M. J., et al. (2012). *A DNA hypermethylation module for the stem/progenitor cell signature of cancer. Genome research*. doi:10.1101/gr.131169.111

Elling, A. a, & Deng, X. W. (2009). Next-generation sequencing reveals complex relationships between the epigenome and transcriptome in maize. *Plant signaling & behavior*, *4*(8), 760–2. doi:10.1105/tpc.109.065714

Enke, R. a., Dong, Z., & Bender, J. (2011). Small RNAs Prevent Transcription-Coupled Loss of Histone H3 Lysine 9 Methylation in Arabidopsis thaliana. (G. P. Copenhaver, Ed.)*PLoS Genetics*, *7*(10), e1002350. doi:10.1371/journal.pgen.1002350

Ernst, J., Kheradpour, P., Mikkelsen, T. S., Shoresh, N., Ward, L. D., Epstein, C. B., Zhang, X., et al. (2011). Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature*, *473*(7345), 43–9. doi:10.1038/nature09906

Filion, G. J., van Bemmel, J. G., Braunschweig, U., Talhout, W., Kind, J., Ward, L. D., Brugman, W., et al. (2010). Systematic protein location mapping reveals five principal chromatin types in Drosophila cells. *Cell*, *143*(2), 212–24. doi:10.1016/j.cell.2010.09.009

Guenther, M. G., Levine, S. S., Boyer, L. a, Jaenisch, R., & Young, R. a. (2007). A chromatin landmark and transcription initiation at most promoters in human cells. *Cell*, *130*(1), 77–88. doi:10.1016/j.cell.2007.05.042

Guo, L., Zhou, J., Elling, A. a, Charron, J.-B. F., & Deng, X. W. (2008). Histone modifications and expression of light-regulated genes in Arabidopsis are cooperatively influenced by changing light conditions. *Plant physiology*, *147*(4), 2070–83. doi:10.1104/pp.108.122929

Hawkins, R. D., Hon, G. C., Lee, L. K., Ngo, Q., Lister, R., Pelizzola, M., Edsall, L. E., et al. (2010). Distinct epigenomic landscapes of pluripotent and lineage-committed human cells. *Cell stem cell*, *6*(5), 479–91. doi:10.1016/j.stem.2010.03.018

Jackson, J. P., Lindroth, A. M., Cao, X., & Jacobsen, S. E. (2002). Control of CpNpG DNA methylation by the KRYPTONITE histone H3 methyltransferase. *Nature*, *416*(6880), 556–60. doi:10.1038/nature731

Jacob, Y., Feng, S., Leblanc, C. A., Bernatavichute, Y. V., Cokus, S., Johnson, L. M., Pellegrini, M., et al. (2010). ATXR5 and ATXR6 are novel H3K27 monomethyltransferases required for chromatin structure and gene silencing. *Nature structural & molecular biology*, *16*(7), 763–768. doi:10.1038/nsmb.1611.ATXR5

Jacob, Y., Stroud, H., Leblanc, C., Feng, S., Zhuo, L., Caro, E., Hassel, C., et al. (2010). Regulation of heterochromatic DNA replication by histone H3 lysine 27 methyltransferases. *Nature*, *466*(7309), 987–91. doi:10.1038/nature09290

Jaenisch, R., & Young, R. (2008). Stem cells, the molecular circuitry of pluripotency and nuclear reprogramming. *Cell*, *132*(4), 567–82. doi:10.1016/j.cell.2008.01.015

Jones, P. a, & Baylin, S. B. (2007). The epigenomics of cancer. *Cell*, *128*(4), 683–92. doi:10.1016/j.cell.2007.01.029

Kharchenko, P. V., Alekseyenko, A. a, Schwartz, Y. B., Minoda, A., Riddle, N. C., Ernst, J., Sabo, P. J., et al. (2011). Comprehensive analysis of the chromatin landscape in Drosophila melanogaster. *Nature*, *471*(7339), 480–5. doi:10.1038/nature09725

Kolasinska-zwierz, P., Down, T., Latorre, I., Liu, T., Liu, X. S., & Ahringer, J. (2009). Differential chromatin marking of introns and expressed exons by H3K36me3. *Nature genetics*, *41*(3), 376–381. doi:10.1038/ng.322.

Lafos, M., Kroll, P., Hohenstatt, M. L., Thorpe, F. L., Clarenz, O., & Schubert, D. (2011). Dynamic regulation of H3K27 trimethylation during Arabidopsis differentiation. *PLoS genetics*, *7*(4), e1002040. doi:10.1371/journal.pgen.1002040

Lechner, M., Boshoff, C., & Beck, S. (2010). *Cancer epigenome. Advances in genetics* (1st ed., Vol. 70, pp. 247–76). Elsevier Inc. doi:10.1016/B978-0-12-380866-0.60009-5

Lehnertz, B., Ueda, Y., Derijck, A. A. H. A., Braunschweig, U., Perez-burgos, L., Kubicek, S., Chen, T., et al. (2003). Suv39h-Mediated Histone H3 Lysine 9 Methylation Directs DNA Methylation to Major Satellite Repeats at Pericentric Heterochromatin. *Current*, *13*, 1192–1200. doi:10.1016/S

Li, B., Carey, M., & Workman, J. L. (2007). The role of chromatin during transcription. *Cell*, *128*(4), 707–19. doi:10.1016/j.cell.2007.01.015

Li, X., Wang, X., He, K., Ma, Y., Su, N., He, H., Stolc, V., et al. (2008). High-resolution mapping of epigenetic modifications of the rice genome uncovers interplay between DNA methylation, histone methylation, and gene expression. *The Plant cell*, *20*(2), 259–76. doi:10.1105/tpc.107.056879

Lienert, F., Mohn, F., Tiwari, V. K., Baubec, T., Roloff, T. C., Gaidatzis, D., Stadler, M. B., et al. (2011). Genomic prevalence of heterochromatic H3K9me2 and transcription do not discriminate pluripotent from terminally differentiated cells. *PLoS genetics*, *7*(6), e1002090. doi:10.1371/journal.pgen.1002090

Lin, N., Li, X., Cui, K., Chepelev, I., Tie, F., Liu, B., Li, G., et al. (2011). A barrier-only boundary element delimits the formation of facultative heterochromatin in Drosophila melanogaster and vertebrates. *Molecular and cellular biology*, *31*(13), 2729–41. doi:10.1128/MCB.05165-11

Lindroth, A. M., Park, Y. J., McLean, C. M., Dokshin, G. a, Persson, J. M., Herman, H., Pasini, D., et al. (2008). Antagonism between DNA and H3K27 methylation at the imprinted Rasgrf1 locus. *PLoS genetics*, *4*(8), e1000145. doi:10.1371/journal.pgen.1000145

Liu, T., Rechtsteiner, A., Egelhofer, T. a, Vielle, A., Latorre, I., Cheung, M.-S., Ercan, S., et al. (2011). Broad chromosomal domains of histone modification patterns in C. elegans. *Genome research*, *21*(2), 227–36. doi:10.1101/gr.115519.110

Liu, Y., Taverna, S. D., Muratore, T. L., Shabanowitz, J., Hunt, D. F., & Allis, C. D. (2007). RNAi-dependent H3K27 methylation is required for heterochromatin formation and DNA elimination in Tetrahymena. *Genes & Development*, *21*(12), 1530–1545. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/17575054

Maheswari, U., Jabbari, K., Petit, J.-L., Porcel, B. M., Allen, A. E., Cadoret, J.-P., De Martino, A., et al. (2010). Digital expression profiling of novel diatom transcripts provides insight into their biological functions. *Genome biology*, *11*(8), R85. doi:10.1186/gb-2010-11-8-r85

Maheswari, U., Mock, T., Armbrust, E. V., & Bowler, C. (2009). Update of the Diatom EST Database: a new tool for digital transcriptomics. *Nucleic acids research*, *37*(Database issue), D1001–5. doi:10.1093/nar/gkn905

Mahfouz, M. M. (2010). RNA-directed DNA methylation: mechanisms and functions. *Plant signaling & behavior*, *5*(7), 806–16. Retrieved from http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3115029&tool=pmcentrez&rendertype=abstract

Margueron, R., & Reinberg, D. (2011). The Polycomb complex PRC2 and its mark in life. *Nature*, *469*(7330), 343–9. doi:10.1038/nature09784

Marks, H., Kalkan, T., Menafra, R., Denissov, S., Jones, K., Hofemeister, H., Nichols, J., et al. (2012). The Transcriptional and Epigenomic Foundations of Ground State Pluripotency. *Cell*, *149*(3), 590–604. doi:10.1016/j.cell.2012.03.026

Mathieu, O., Probst, A. V., & Paszkowski, J. (2005). Distinct regulation of histone H3 methylation at lysines 27 and 9 by CpG methylation in Arabidopsis. *The EMBO journal*, *24*(15), 2783–91. doi:10.1038/sj.emboj.7600743

Maumus, F., Allen, A. E., Mhiri, C., Hu, H., Jabbari, K., Vardi, A., Grandbastien, M.-A., et al. (2009). Potential impact of stress activated retrotransposons on genome evolution in a marine diatom. *BMC Genomics*, *10*(1), 624. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/20028555

Meissner, A. (2010). Epigenetic modifications in pluripotent and differentiated cells. *Nature biotechnology*, *28*(10), 1079–1088. doi:10.1038/nbt1684

Mikkelsen, T. S., Ku, M., Jaffe, D. B., Issac, B., Lieberman, E., Giannoukos, G., Alvarez, P., et al. (2007). Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. *Nature*, *448*(7153), 553–60. doi:10.1038/nature06008

Millar, C. B., & Grunstein, M. (2006). Genome-wide patterns of histone modifications in yeast. *Nature reviews. Molecular cell biology*, *7*(9), 657–66. doi:10.1038/nrm1986

Mohn, F., Weber, M., Rebhan, M., Roloff, T. C., Richter, J., Stadler, M. B., Bibel, M., et al. (2008). Lineage-specific polycomb targets and de novo DNA methylation define restriction and potential of neuronal progenitors. *Molecular cell*, *30*(6), 755–66. doi:10.1016/j.molcel.2008.05.007

Oh, S., Park, S., & van Nocker, S. (2008). Genic and global functions for Paf1C in chromatin modification and gene expression in Arabidopsis. *PLoS genetics*, *4*(8), e1000077. doi:10.1371/journal.pgen.1000077

Ohm, J. E., McGarvey, K. M., Yu, X., Cheng, L., Schuebel, K. E., Cope, L., Mohammad, H. P., et al. (2007). A stem cell-like chromatin pattern may predispose tumor suppressor genes to DNA hypermethylation and heritable silencing. *Nature genetics*, *39*(2), 237–42. doi:10.1038/ng1972

Oktaba, K., Gutiérrez, L., Gagneur, J., Girardot, C., Sengupta, A. K., Furlong, E. E. M., & Müller, J. (2008). Dynamic regulation by polycomb group protein complexes controls pattern formation and the cell cycle in Drosophila. *Developmental cell*, *15*(6), 877–89. doi:10.1016/j.devcel.2008.10.005

Pauler, F. M., Sloane, M. a, Huang, R., Regha, K., Koerner, M. V., Tamir, I., Sommer, A., et al. (2009a). H3K27me3 forms BLOCs over silent genes and intergenic regions and specifies a histone banding pattern on a mouse autosomal chromosome. *Genome research*, *19*(2), 221–33. doi:10.1101/gr.080861.108

Pauler, F. M., Sloane, M. a, Huang, R., Regha, K., Koerner, M. V., Tamir, I., Sommer, A., et al. (2009b). H3K27me3 forms BLOCs over silent genes and intergenic regions and specifies a histone banding pattern on a mouse autosomal chromosome. *Genome research*, *19*(2), 221–33. doi:10.1101/gr.080861.108

Pekowska, A., Benoukraf, T., Ferrier, P., & Spicuglia, S. (2010). A unique H3K4me2 profile marks tissue-specific gene regulation. *Genome research*, *20*(11), 1493–502. doi:10.1101/gr.109389.110

Rajasekhar, V. K., & Begemann, M. (2007). Concise review: roles of polycomb group proteins in development and disease: a stem cell perspective. *Stem cells (Dayton, Ohio)*, *25*(10), 2498–510. doi:10.1634/stemcells.2006-0608

Roh, T.-Y., Cuddapah, S., Cui, K., & Zhao, K. (2006). The genomic landscape of histone modifications in human T cells. *Proceedings of the National Academy of Sciences of the United States of America*, *103*(43), 15782–7. doi:10.1073/pnas.0607617103

Roudier, F., Ahmed, I., Bérard, C., Sarazin, A., Mary-Huard, T., Cortijo, S., Bouyer, D., et al. (2011a). Integrative epigenomic mapping defines four main chromatin states in Arabidopsis. *The EMBO journal*, *30*(10), 1928–38. doi:10.1038/emboj.2011.103

Roudier, F., Ahmed, I., Bérard, C., Sarazin, A., Mary-Huard, T., Cortijo, S., Bouyer, D., et al. (2011b). Integrative epigenomic mapping defines four main chromatin states in Arabidopsis. *The EMBO journal*, *30*(10), 1928–38. doi:10.1038/emboj.2011.103

Rugg-Gunn, P. J., Cox, B. J., Ralston, A., & Rossant, J. (2010). Distinct histone modifications in stem cell lines and tissue lineages from the early mouse embryo. *Proceedings of the National Academy of Sciences of the United States of America*, *107*(24), 10783–90. doi:10.1073/pnas.0914507107

Schlesinger, Y., Straussman, R., Keshet, I., Farkash, S., Hecht, M., Zimmerman, J., Eden, E., et al. (2007). Polycomb-mediated methylation on Lys27 of histone H3 pre-marks genes for de novo methylation in cancer. *Nature genetics*, *39*(2), 232–6. doi:10.1038/ng1950

Schuettengruber, B., Ganapathi, M., Leblanc, B., Portoso, M., Jaschek, R., Tolhuis, B., van Lohuizen, M., et al. (2009). Functional anatomy of polycomb and trithorax chromatin landscapes in Drosophila embryos. *PLoS biology*, *7*(1), e13. doi:10.1371/journal.pbio.1000013

Schübeler, D., MacAlpine, D. M., Scalzo, D., Wirbelauer, C., Kooperberg, C., van Leeuwen, F., Gottschling, D. E., et al. (2004). The histone modification pattern of active genes revealed through genome-wide chromatin analysis of a higher eukaryote. *Genes & development*, *18*(11), 1263–71. doi:10.1101/gad.1198204

Shaver, S., Casas-Mollano, J. A., Cerny, R. L., & Cerutti, H. (2010). Origin of the polycomb repressive complex 2 and gene silencing by an E(z) homolog in the unicellular alga Chlamydomonas. *Epigenetics : official journal of the DNA Methylation Society*, *5*(4), 301–12. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/20421736

Shi, J., & Dawe, R. K. (2006). Partitioning of the maize epigenome by the number of methyl groups on histone H3 lysines 9 and 27. *Genetics*, *173*(3), 1571–83. doi:10.1534/genetics.106.056853

Siaut, M., Heijde, M., Mangogna, M., Montsant, A., Coesel, S., Allen, A., Manfredonia, A., et al. (2007). Molecular toolbox for studying diatom biology in Phaeodactylum tricornutum. *Gene*, *406*(1-2), 23–35. doi:10.1016/j.gene.2007.05.022

Sinha, I., Wirén, M., & Ekwall, K. (2006). Genome-wide patterns of histone modifications in fission yeast. *Chromosome research: an international journal on the molecular, supramolecular and evolutionary aspects of chromosome biology*, *14*(1), 95–105. doi:10.1007/s10577-005-1023-4

Tamaru, H., & Selker, E. U. (2001). A histone H3 methyltransferase controls DNA methylation in Neurospora crassa. *Nature*, *414*(6861), 277–83. doi:10.1038/35104508

Tirichine, L., & Bowler, C. (2011). Decoding algal genomes: tracing back the history of photosynthetic life on Earth. *The Plant journal: for cell and molecular biology* , *66*(1), 45–57. doi:10.1111/j.1365-313X.2011.04540.x

Turck, F., Roudier, F., Farrona, S., Martin-Magniette, M.-L., Guillaume, E., Buisine, N., Gagnot, S., et al. (2007). Arabidopsis TFL2/LHP1 specifically associates with genes marked by trimethylation of histone H3 lysine 27. *PLoS genetics*, *3*(6), e86. doi:10.1371/journal.pgen.0030086

Wang, Z., Schones, D. E., & Zhao, K. (2009). Characterization of human epigenomes. *Current opinion in genetics & development*, *19*(2), 127–34. doi:10.1016/j.gde.2009.02.001

Weinhofer, I., Hehenberger, E., Roszak, P., Hennig, L., & Ko, C. (2010). H3K27me3 Profiling of the Endosperm Implies Exclusion of Polycomb Group Protein Targeting by DNA Methylation, *6*(10), 1–14. doi:10.1371/journal.pgen.1001152

Wu, H., Coskun, V., Tao, J., Xie, W., Ge, W., Yoshikawa, K., Li, E., et al. (2010). Dnmt3a-dependent nonpromoter DNA methylation facilitates transcription of neurogenic genes. *Science (New York, N.Y.)*, *329*(5990), 444–8. doi:10.1126/science.1190485

Yin, H., Sweeney, S., Raha, D., Snyder, M., & Lin, H. (2011). A high-resolution whole-genome map of key chromatin modifications in the adult Drosophila melanogaster. *PLoS genetics*, *7*(12), e1002380. doi:10.1371/journal.pgen.1002380

Young, M. D., Willson, T. a, Wakefield, M. J., Trounson, E., Hilton, D. J., Blewitt, M. E., Oshlack, A., et al. (2011). ChIP-seq analysis reveals distinct H3K27me3 profiles that correlate with transcriptional activity. *Nucleic acids research*, 1–13. doi:10.1093/nar/gkr416

Yuzyuk, T., Fakhouri, T. H. I., Kiefer, J., & Mango, S. E. (2009). The polycomb complex protein mes-2/E(z) promotes the transition from developmental plasticity to differentiation in C. elegans embryos. *Developmental cell*, *16*(5), 699–710. doi:10.1016/j.devcel.2009.03.008

Zachary Lippman, Anne-Vale´ rie Gendrel, M. B., Matthew W. Vaughn, Neilay Dedhia1, W. Richard McCombie, Kimberly Lavine, Vivek Mittal, Bruce May, Kristin D. Kasschau, James C. Carrington, Rebecca W. Doerge, V. C., & Martienssen1, & R. (2004). Role of transposable elements in heterochromatin and epigenetic control, *2*. doi:10.1038/nature02724.1.

Zhang, X., Bernatavichute, Y. V., Cokus, S., Pellegrini, M., & Jacobsen, S. E. (2009). Genome-wide analysis of mono-, di- and trimethylation of histone H3 lysine 4 in Arabidopsis thaliana. *Genome biology*, *10*(6), R62. doi:10.1186/gb-2009-10-6-r62

Zhang, X., Clarenz, O., Cokus, S., Bernatavichute, Y. V., Pellegrini, M., Goodrich, J., & Jacobsen, S. E. (2007a). Whole-genome analysis of histone H3 lysine 27 trimethylation in Arabidopsis. *PLoS biology*, *5*(5), e129. doi:10.1371/journal.pbio.0050129

Zhang, X., Clarenz, O., Cokus, S., Bernatavichute, Y. V., Pellegrini, M., Goodrich, J., & Jacobsen, S. E. (2007b). Whole-genome analysis of histone H3 lysine 27 trimethylation in Arabidopsis. *PLoS biology*, *5*(5), e129. doi:10.1371/journal.pbio.0050129

Zhao, X. D., Han, X., Chew, J. L., Liu, J., Chiu, K. P., Choo, A., Orlov, Y. L., et al. (2007). Whole-genome mapping of histone H3 Lys4 and 27 trimethylations reveals distinct genomic

compartments in human embryonic stem cells. *Cell stem cell*, *1*(3), 286–98. doi:10.1016/j.stem.2007.08.004

Zhao, Y., & Zhou, D.-X. (2012). Epigenomic modification and epigenetic regulation in rice. *Journal of Genetics and Genomics*, 1–9. doi:10.1016/j.jgg.2012.02.009

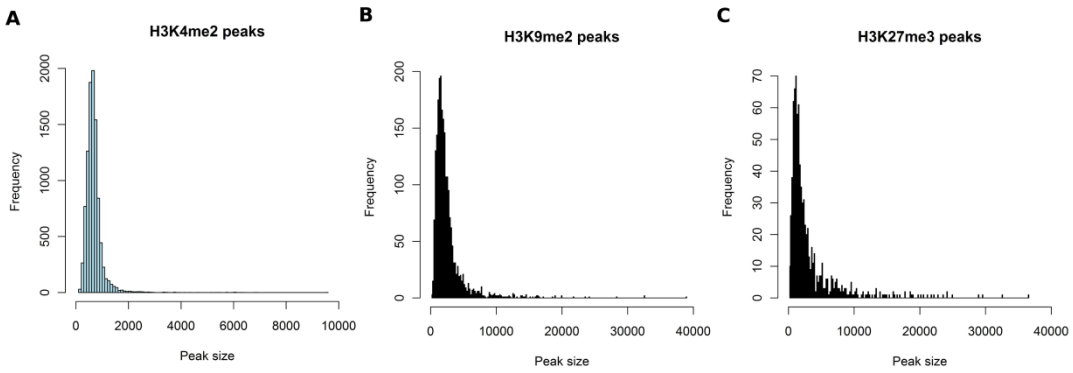
Zhou, J., Wang, X., He, K., Charron, J.-B. F., Elling, A. a, & Deng, X. W. (2010). Genome-wide profiling of histone H3 lysine 9 acetylation and dimethylation in Arabidopsis reveals correlation between multiple histone marks and gene expression. *Plant molecular biology*, *72*(6), 585–95. doi:10.1007/s11103-009-9594-7

Zhou, V. W., Goren, A., & Bernstein, B. E. (2011). Charting histone modifications and the functional organization of mammalian genomes. *Nature reviews. Genetics*, *12*(1), 7–18. doi:10.1038/nrg2905

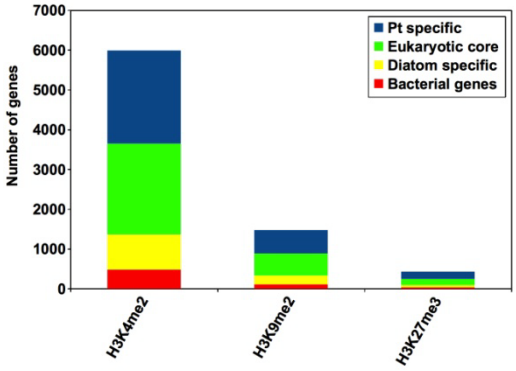
van Leeuwen, F., & van Steensel, B. (2005). Histone modifications: from genome-wide maps to functional insights. *Genome biology*, *6*(6), 113. doi:10.1186/gb-2005-6-6-113

3.8 Supplementary information

3.8.1 Supplementary Figures

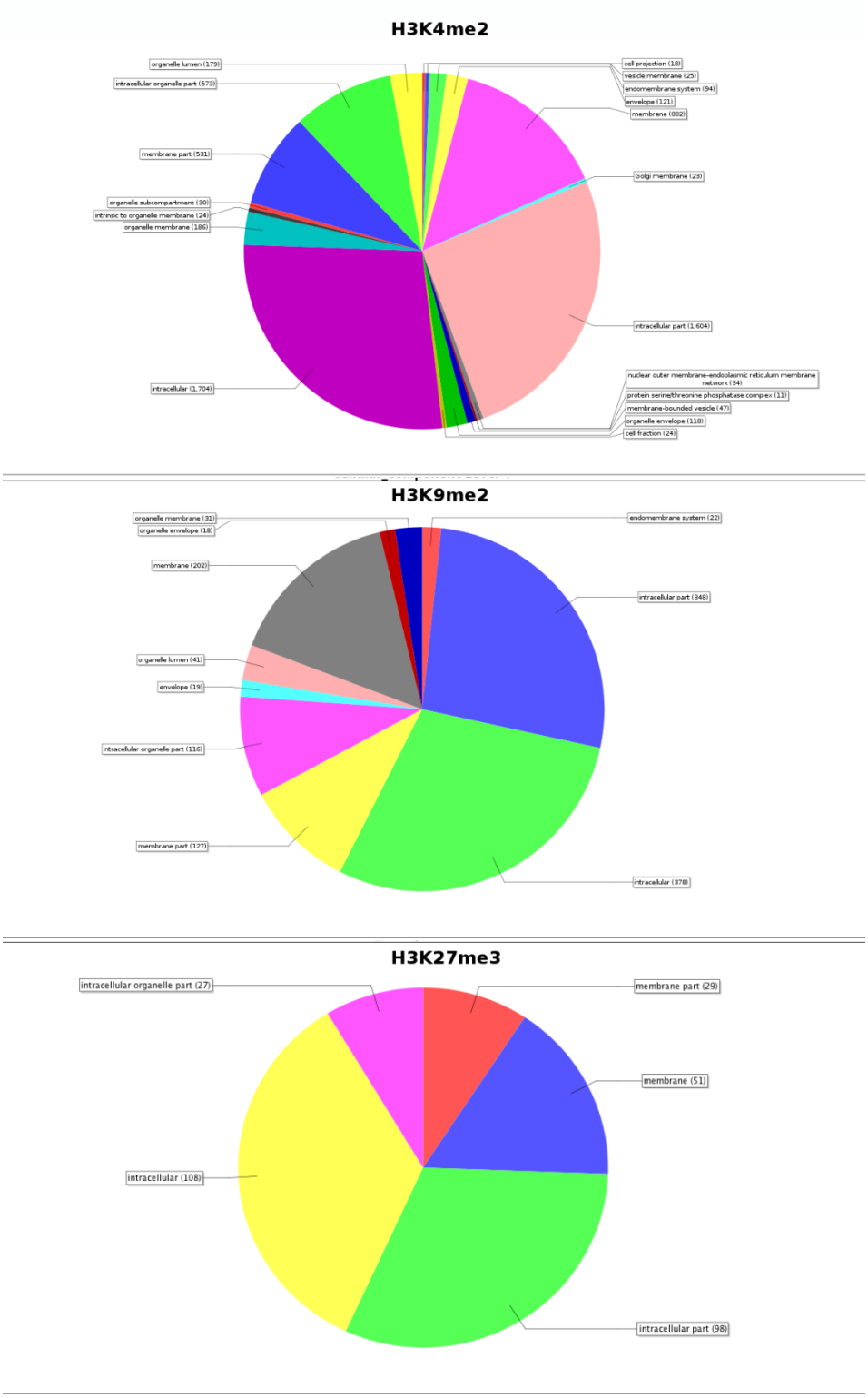


Supp Fig 3.1 ChIP-seq signal peak size of enriched domains from three sets of histone modification ChIP-Seq data (H3K4me2, H3K9me2 and H3K27me3 marks). Peaks were detected using SICER, with a window of 200 bp and a gap of 1 window for H3K4me2 and 3 windows for H3K9me2 and H3K27me3 (although change in window size does not affect peak size). Average peak size is 175-250 bp for H3K4me2 and ~900bp-1kb for H3K9me2 and H3K27me3.

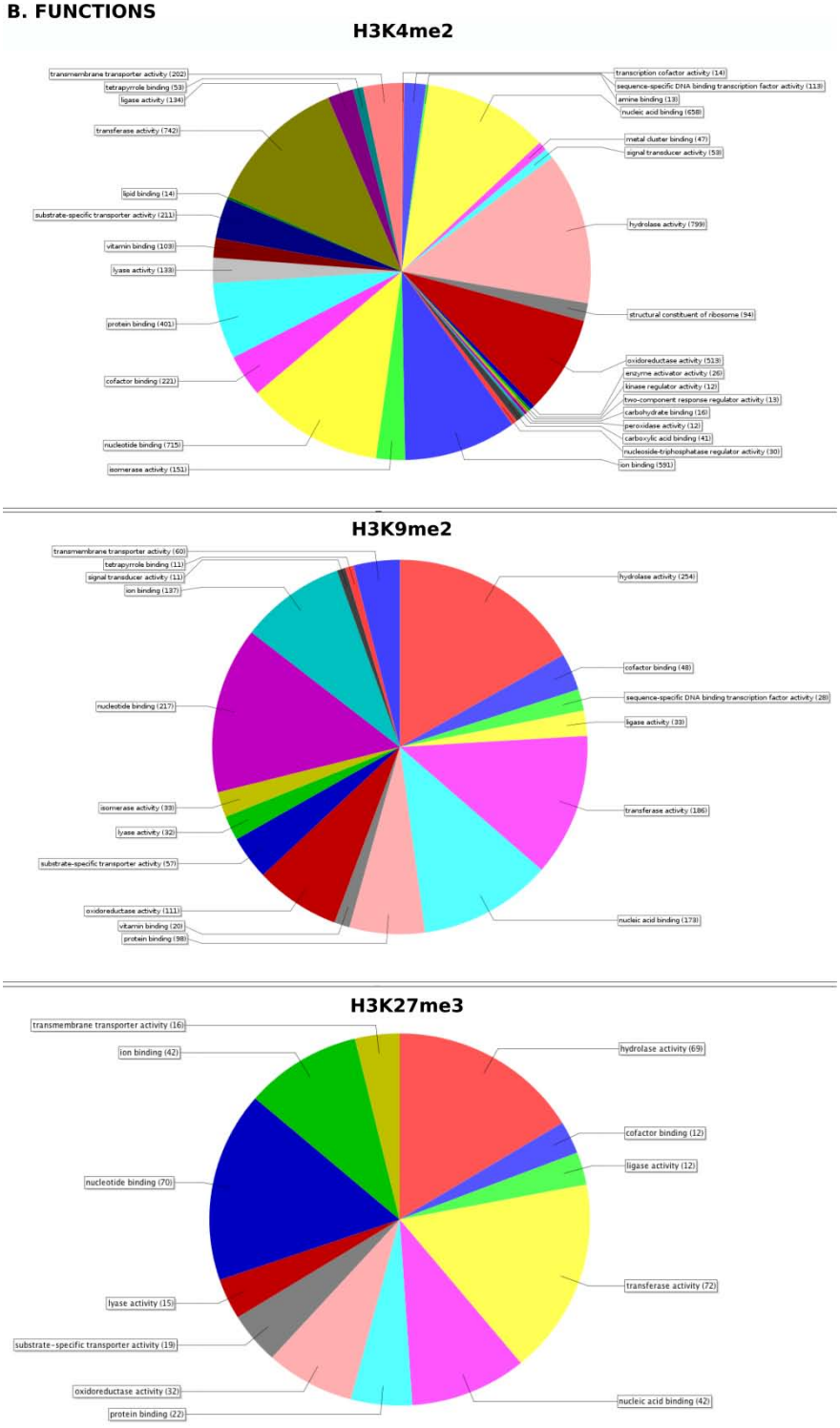


Supp Fig 3.2. Three different histone methylation modifications on Orthologous gene groups in *P.tricornutum*. Four orthologous groups have been defined in *P.tricornutum* by genome comparison (2887 Pt-specific putative encoding genes are present only in this diatom, 2,913 Core eukaryotic putative encoding genes are shared in all eukaryotes, 1,083 Diatom specific putative encoding genes which are present in both sequenced diatoms but in no other eukaryote, 587 Bacterial proteins). Each histone modification is found in equal proportion on all four orthologous gene groups.

A. CELLULAR COMPONENT

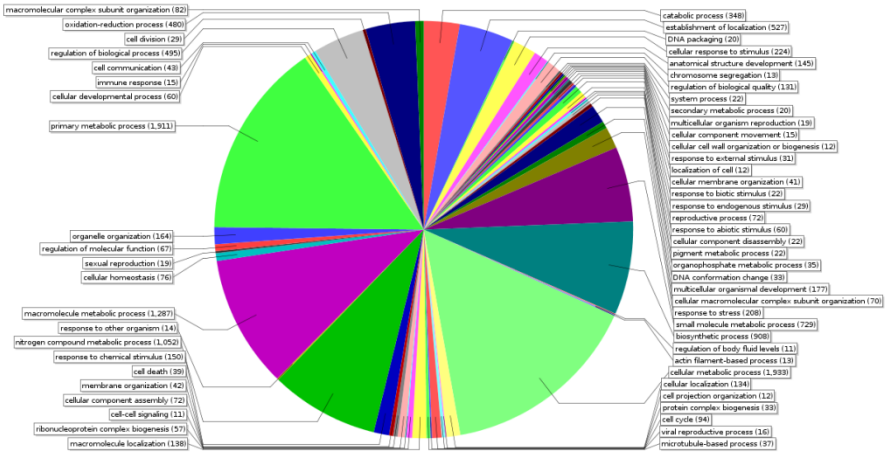


B. FUNCTIONS

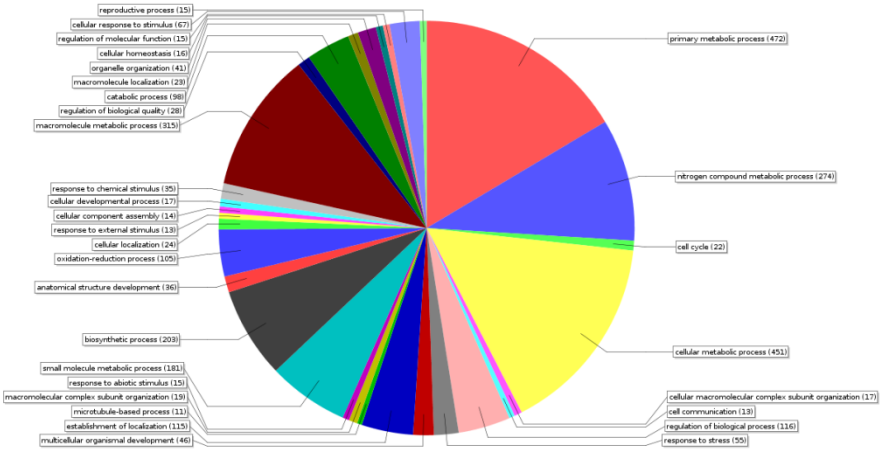

H3K4me2
transmembrane transporter activity (202)
tetrapyrrole binding (53)
ligase activity (134)
transferase activity (742)
lipid binding (14)
substrate-specific transporter activity (211)
vitamin binding (103)
lyase activity (133)
protein binding (401)
cofactor binding (221)
nucleotide binding (715)
isomerase activity (151)
transcription cofactor activity (14)
sequence-specific DNA binding transcription factor activity (113)
amine binding (13)
nucleic acid binding (658)
metal cluster binding (47)
signal transducer activity (53)
hydrolase activity (799)
structural constituent of ribosome (94)
oxidoreductase activity (513)
enzyme activator activity (26)
kinase regulator activity (12)
two-component response regulator activity (13)
carbohydrate binding (16)
peroxidase activity (12)
carboxylic acid binding (41)
nucleoside-triphosphatase regulator activity (30)
ion binding (591)
H3K9me2
transmembrane transporter activity (60)
tetrapyrrole binding (11)
signal transducer activity (11)
ion binding (137)
nucleotide binding (217)
isomerase activity (33)
lyase activity (32)
substrate-specific transporter activity (57)
oxidoreductase activity (111)
vitamin binding (20)
protein binding (98)
hydrolase activity (254)
cofactor binding (48)
sequence-specific DNA binding transcription factor activity (28)
ligase activity (33)
transferase activity (186)
nucleic acid binding (173)
H3K27me3
transmembrane transporter activity (16)
ion binding (42)
nucleotide binding (70)
lyase activity (15)
substrate-specific transporter activity (19)
oxidoreductase activity (32)
protein binding (22)
hydrolase activity (69)
cofactor binding (12)
ligase activity (12)
transferase activity (72)
nucleic acid binding (42)

C.CELLULAR PROCESSES

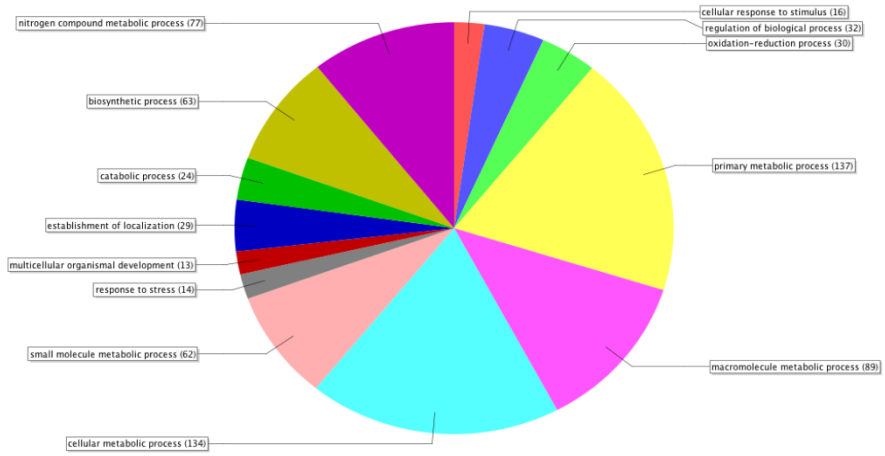
H3K4me2



H3K9me2



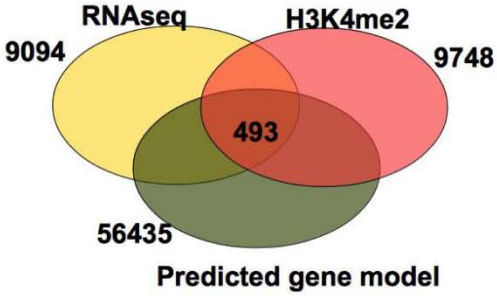
H3K27me3



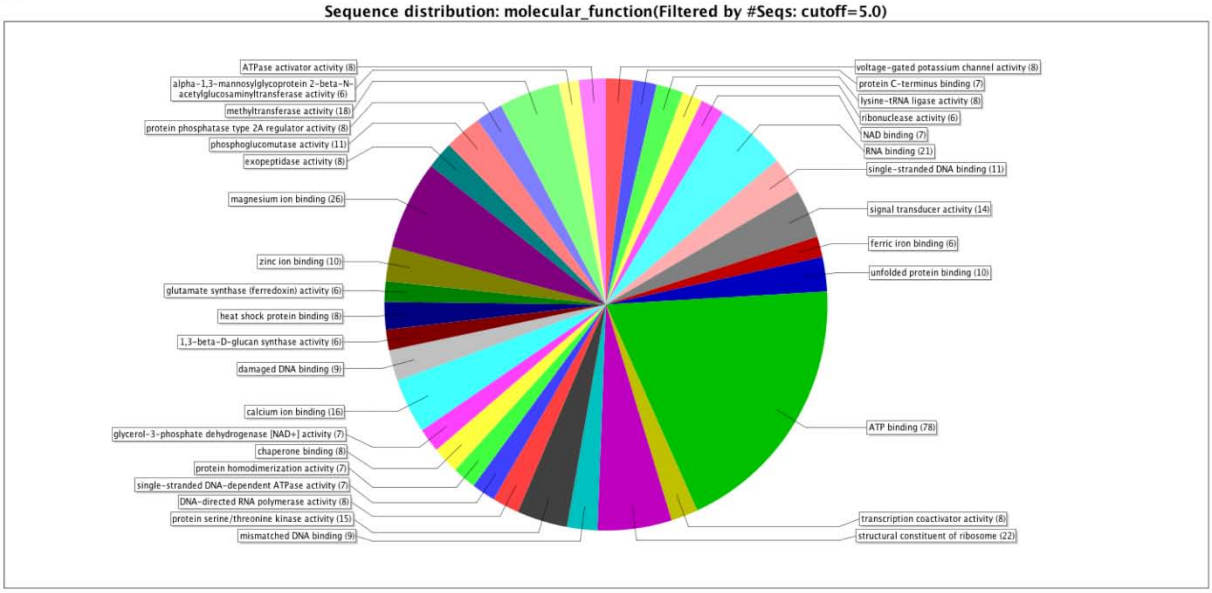
Supp Fig3.3 Gene Ontology (GO) categories of genes marked by the three different histone modifications. **A.** Cellular component category of gene ontology. **B.** Gene ontology analysis of genes within this functional category reveals that H3K9me2 and H3K27me3 are localized in genes with similar functions. **C.** GO cellular processes category shows a distinct difference between these three histone methylation modified genes. H3K27me3 is limited to a few genes within this category.

A

Category	Number of known gene model	Number of novel proteins
Histone mark (H3K4me2)	7896/10402	1852 (9748-7896)
Expression profile	9094/10402	721



B



Supp Fig 3.4 Genome annotation for novel genes predicted. A. Comparison of gene models, genes predicted using RNAseq data and H3K4me2 localized genes from ChIP-seq data. At least 721 proteins could be predicted from RNAseq. 423 proteins from all three methods could be further annotated. **B**. BLAST2GO annotation of the predicted proteins.

3.8.1 Supplementary tables

Table S3.1 Validation of enriched and unenriched loci, including genes and TEs. Several loci from different locations on the genome were chosen for validation. (a) H3K4me2 ChIP-Qpcr validation. (b)H3K9me2 ChIP-Qpcr validation. (c) H3K27me3 ChIP-Qpcr validation

(a) H3K4me2 ChIP-Qpcr validation.

loci for validation	enrichment in ChIP-seq data	enrichment in ChIP-qpcr
37885	yes	yes
40162	yes	yes
44850	yes	yes
45961	yes	yes
47483	yes	yes
49580	yes	yes
49558	yes	yes
55090	yes	yes
21198	yes	yes
40447	yes	yes
46440	yes	yes
10524	no	yes low
12909	no	yes low
43826	no	yes low
45975	no	yes
39237	no	yes
9828	no	yes

(b) H3K9me2 ChIP-Qpcr validation

loci for validation	enrichment in ChIP-seq data	enrichment in ChIP-qpcr
TE chr_1:6090-6340	yes	yes
TE chr_4:50601-50802	yes	yes
TE chr_5:1074130-1074350	yes	yes
TE chr_5:17100-17390	yes	yes

TE chr_7:15366-15571	yes	yes
TE chr_8:999190-999440	yes	yes
40338	yes	yes
48357	yes	yes
12366	yes	yes
52547	yes	yes
TE chr_25:443170-443400	yes	yes
38048	yes	yes
36992	yes	yes
12909	yes	yes
10524	no	yes
9828	no	yes
21988	no	yes
TE chr_12:900460-900640	no	yes
TE chr_18:178040-178240	no	yes
14212	no	yes
34971	no	no
54330	no	no

(c) H3K27me3 ChIP-Qpcr validation

loci for validation	**enrichment in ChIP-seq data**	**enrichment in ChIP-qpcr**
40474	no but yes in MACS	yes
34469	yes	yes
11472	yes	yes
TE Chr_1 :6050-6350	yes	yes
TE chr_5:1074500-1074740	yes	yes
TE chr_8:22199-22354	yes	yes
TE chr_3:94200-94500	yes	yes
chr_7:15380-15670	yes	yes
39885	yes	yes
TE chr_12:900460-900640	yes	yes
TE chr_18:178040-178240	yes	yes
39237	yes	yes
45975	yes	yes

chr_5:17100-17390	yes	yes
chr_7:15366-15571	yes	yes
45041	no	no

3.9 Annex-Chromatin immunoprecipitation protocol

Chromatin immunoprecipitation coupled to detection by quantitative real time PCR to study in vivo protein DNA interactions in two model diatoms *Phaeodactylum tricornutum* and *Thalasiossira pseudonana*

Xin Lin[1], Leïla Tirichine[1*] and Chris Bowler[1*]

Ecole Normale Supérieure, Institut de Biologie de l'ENS, IBENS, Paris, F-75005 France. CNRS, UMR 8197, Paris, F-75005. 46 rue d'Ulm, 75005 Paris, France.

[*]Corresponding author

Abstract

In this report we describe a chromatin immunoprecipitation (ChIP) protocol for two fully sequenced model diatom species *Phaeodactylum tricornutum* and *Thalasiossira pseudonana*. This protocol allows the extraction of satisfactory amounts of chromatin and gives reproducible results. We coupled the ChIP assay with real time quantitative PCR. Our results reveal that the two major histone marks H3K4me2 and H3K9me2 exist in *P. tricornutum* and *T. pseudonana.* As in other eukaryotes, H3K4me2 marks active genes whereas H3K9me2 marks transcriptionally inactive transposable elements. Unexpectedly however, *T. pseudonana* housekeeping genes also show a relative enrichment of H3K9me2. We also discuss optimization of the procedure, including growth conditions, cross linking and sonication. Validation of the protocol provides a set of genes and transposable elements that can be used as controls for studies using ChIP in each diatom species. This protocol can be easily adapted to other diatoms and eukaryotic phytoplankton species for genetic and biochemical studies.

Introduction

Diatoms are a group of eukaryotic phytoplankton with a wide distribution and a large diversity in marine and fresh water ecosystems. It is estimated that there are between 10,000 and 100,000 extant diatom species (Round, F. E., 1992). Diatoms play essential roles in global biogeochemical cycles because they are believed to be responsible for 20% of global carbon fixation and 40% of marine primary productivity (David M.Nelson, Paul Tréguer, Mark A. Brzezinski, Aude Leynaert, 1995). Diatoms capture CO_2 through photosynthesis and act as a critical buffer against global warming by sequestering organic carbon in the ocean interior. They can live under different conditions in all oceans from the poles to the tropics, and are often the first group of phytoplankton to benefit from sporadic nutrient upwelling events, indicating their intrinsic adaptability to changing environments.

Completed whole genome sequences from three species, the centric diatom *Thalasiossira pseudonana,* and the pennate diatoms *Phaeodactylum tricornutum* and *Fragilariopsis cylindrus* are now available (Armbrust et al., 2004), (Bowler et al., 2008)(http://genome.jgi-psf.org/Fracy1/Fracy1.home.html). Another pennate species, the toxic *Pseudo-nitzschia multiseries* is also being sequenced and will provide an additional source for comparative genomics. *T. pseudonana* is widely distributed in marine environments and is of significant ecological importance. *P. tricornutum* on the other hand is considered to be of little ecological relevance but is the representative of pennate diatoms because of a long history of physiological experiments and the availability of a wide range of tools for reverse genetics (Siaut et al., 2007)(De Riso et al., 2009). Furthermore, a digital gene expression database (http://www.diatomics.biologie.ens.fr/EST3/index.php) is available for both species.

The whole genome sequences have revealed a wealth of information about diatom genes. It was shown for example that diatoms have acquired genes both from their endosymbiotic ancestors and by horizontal gene transfer from prokaryotes. But while DNA is the substrate for mutations upon which natural selection can act, DNA sequence in itself may not explain adequately their ability to adapt to changing environments, and more flexible mechanisms based on epigenetic processes could provide additional control. These changes include DNA methylation and histone tail post-translational modifications that alter chromatin structure.

Study of diatom epigenomes can therefore provide a more in depth look at the regulatory mechanisms underlying their natural phenotypic adaptability to environmental changes.

Several tools for studying chromatin have been developed. Among them, chromatin immuno-precipitation (ChiP) has become a powerful tool to detect in vivo interactions between a DNA-associated protein and DNA fragments. ChIP combined with microarray or massively parallel sequencing is used to study gene regulatory networks active during development and/or in response to the environment. ChIP is also a valuable tool for mapping genome wide epigenetic modifications such as histone marks, and has been used to characterize several eukaryotic genomes (Gendrel, Lippman, Martienssen, & Colot, 2005) (Grably M, 2010). However, no ChIP protocol has been reported for marine phytoplankton. *T. pseudonana* and *P. tricornutum* were therefore chosen to set up a ChIP protocol in diatoms using two histone marks known to characterize active and repressive chromatin states in other organisms.

The principle of a ChIP procedure includes: (1) Cross linking of DNA and protein with formaldehyde to covalently combine DNA and attached proteins *in vivo*, (2) Fragmentation of the fixed chromatin to an average size of 500 bp ranging from 200 to 1000 bp, (3) Chromatin extraction by a succession of extraction buffers, (4) Immunoprecipitation with specific antibodies, (5) Purification of DNA after immunoprecipitation and reverse crosslinking, and (6) Analysis of bound DNA by PCR which involves comparison of the intensity of PCR signals from the precipitated template with positive and negative controls. Standard PCR on immunoprecipitated DNA from a specific genomic region provides a direct assessment of protein association with that region, whereas quantitative PCR can assess not only whether a protein binds to that region, but also further compare the relative abundance at different genomic regions.

The protocol described in this work was optimized for each of the steps described above. It is an adaptation of ChIP protocols used for yeast and *Arabidopsis* (Saleh, Alvarez-Venegas, & Avramova, 2008)(Nelson, Denisenko, & Bomsztyk, 2006). It is therefore not new in its principles but takes into consideration features inherent to diatoms such as their siliceous frustule, a component that can make the simple extraction of chromatin a considerable challenge.

Results and discussion

Chromatin extraction and processing

ChIP is a very powerful technique for revealing association of specific DNA regions with proteins of interest. However, it is not a trivial technique, and highly specific antibodies against the protein of interest or a particular histone modification are required. Furthermore, false-negative signals may originate from inefficient antibody binding, and the beads used in ChIP can bind non-specific sites and cause background noise in negative controls. The starting materials for immunoprecipitation are also critical, and sample dilution can effectively decrease background noise. To avoid saturation of antibody in ChIP assays, calibration curves should be built before precipitation with the antibody to determine the optimal amount of antibody to use.

The protocol described herein (**Fig. 1**) has taken into consideration these issues as well as the particularities of diatoms. Diatoms have the unique ability to precipitate soluble silicic acid into a finely patterned cell wall built from amorphous silica, and they also possess photosynthetic chloroplasts. Chromatin extraction buffers modified from *Arabidopsis* ChIP protocols were used for *P. tricornutum* and *T. pseudonana* which have plant features, such as chloroplasts and a cell wall. It is important in this protocol to use artificial sea water instead of natural sea water in order to control the different components of the medium such as silica. Extraction of chromatin from diatom species can be very difficult because of the silica based rigid cell wall (Kim et al., 2012), which can interfere with chromatin extraction because it binds DNA. A compromise therefore needs to be found between an optimal growth of cells permitted by low concentration of silica and satisfactory amounts of extracted chromatin. In this study, we have chosen species that have different requirements for silicon. *P. tricornutum* has a facultative requirement for silicic acid, whereas *T. pseudonana* needs a silicon source to grow. A growth medium without silicic acid was used for *P. tricornutum* while *T. pseudonana* grew in medium containing a low concentration (0.025 g/l).

To preserve cell integrity, diatom cells were fixed in the growth medium prior to any handling. Formaldehyde was used to cross-link the proteins to the DNA. Cross linking is a time dependent procedure and our trials have established 10 minutes as an optimal time for both *P. tricornutum* and *T. pseudonana*. Excessive cross-linking might reduce antigen accessibility

and sonication efficiency. For a different diatom species, we recommend 5 minutes for a start, proceed with sonication, reverse cross-link and run a gel to see how much DNA is recovered and whether the size is optimal. If not, we recommend trying longer or shorter times for cross linking.

Sonication time was determined after trying three different times, 6 cycles of 30s ON and 1 minute OFF, 9 and 12 cycles. Nine cycles gave the best range of DNA sizes which is between 200 and 1000 bp with a maximum of DNA fragments at 500 bp (data not shown).

Peptide competition assay

The nuclear-enriched protein used for immunoblotting was extracted following a chromatin extraction protocol with minor modifications (see Materials and Methods). Antibodies for two histone marks H3K4me2 and H3K9me2 were used for validation of the ChiP protocol. A peptide competition assay was performed to confirm the specific band reactivity of the antibodies. This is an important issue especially for genome wide studies as unspecific antibody binding will lead to non-specific signals increasing background noise and the occurrence of false positives. Both antibodies were pre-incubated with three different concentrations of the corresponding peptides (see Materials and Methods) prior to immunobloting. For the H3K4me2 histone mark, the peptide competition assay did not detect the presence of a band for peptide H3K4me2 while a band was seen for the two other peptides, indicating the absence of competition with the other modifications of lysine 4 (**Fig**. 2A). Three different references of H3K9me2 antibodies from Millipore were tested in this assay. Two of them were discarded because of a lack of specificity while the third one gave better results (**Fig**. 2B).

H3K4me2 is enriched in genes in *P. tricornutum* and *T. pseudonana*

Quantitative real time transcriptase polymerase chain reaction (QRT-PCR) coupled with the ChIP protocol described herein was used to investigate the enrichment of H3K4 and H3K9 dimethylation on two genes, histone H4 and diatom phytochrome (Dph), and four transposable elements (TEs) in *P. tricornutum* and *T. pseudonana*. A clear enrichment of genes in H3K4me2 shown by a set of primers spanning promoter and gene body was demonstrated by QRT-PCR (**Fig**. 3A). Both histone H4 and Dph show significant differences

in the enrichment in H3K4me2 between immunoprecipitated sample and mock which is the no antibody control. However, no significant differences were observed for the five TEs chosen in this study. Likewise, in *T. pseudonana*, the Dph and histone H4 show enrichment in H3K4me2 in both promoter and gene body, whereas the four chosen TEs (C12, C4, C19 and G1) show no enrichment in H3K4me2 (**Fig**. 3B). Additional genes and TEs were tested and showed similar results (data not shown). Altogether, our data show a conservation of H3K4me2 location between the two diatom species and multicellular organisms, because several genome wide studies in plants and mammals have indeed shown the presence of H3K4me2 on genes (C. L. Liu et al., 2005)(Mikkelsen et al., 2007)(Xiaoyu Zhang et al., 2009).

H3K9me2 shows a different enrichment profile in *P. tricornutum* and *T. pseudonana*

ChiP analysis of H3K9me2 in *P. tricornutum* revealed that TEs contain a significant enrichment for this mark. We particularly focused on two previously characterized diatom-specific copia-like retrotransposable elements known as *Blackbeard* (*Bkb*) and *Surcouf* (*Scf*) (Maumus et al., 2009b) *Bkb* was particularly enriched in H3K9me2 while *Scf* was the least enriched. Intermediate levels were observed among the remaining TEs (**Fig**. 4A). On the other hand, the two protein-coding genes histone H4 and Dph were clearly depleted of H3K9me2 in both promoter and gene body regions (**Fig**. 4A). Both genes were shown by us and others to be transcriptionally active ((Siaut et al., 2007) unpublished results). It was previously shown that *Bkb* and *Scf* are also transcriptionally inactive in normal growth conditions (Maumus et al., 2009b). This is consistent with the primary function of H3K9me2 in repressing TEs in order to maintain genome stability. In *P. tricornutum*, the ChIP assay shows that repressed TEs are marked by H3K9me2 while active genes are depleted in this mark which is similar to what has been observed in plants and mammals (J. Zhou et al., 2010)(Tachibana et al., 2002).

In *T. pseudonana*, a similar pattern of enrichment of TEs by H3K9me2 was observed (**Fig**. 4B). The four tested TEs were highly enriched for H3K9me2, indicating a conserved profile for this mark among both diatoms. Surprisingly, some genes also showed a significant enrichment for H3K9me2, particularly at promoter regions (**Fig** 4B). This unusual association of H3K9me2 with active genes can be due to an intrinsic feature of the centric diatom *T. pseudonana* which is believed to have diverged from its distantly related pennate diatom *P.*

tricornutum 90 million years ago. Comparative genomics and analysis of molecular divergence has shown indeed that both genomes are as different as those of mammals and fish (Bowler et al., 2008). Furthermore, combinatorial patterns of antagonistic chromatin marks are known to occur (Bapat et al., 2010; Bernstein et al., 2006; Roudier et al., 2011b; Weishaupt, Sigvardsson, & Attema, 2010). In *Drosophila* S2 cells, clusters of transcriptionally active genes were reported to be enriched in H3K9me2 (Riddle et al., 2011). Similarly, differentiated mouse ES cells were reported to contain large domains of H3K9me2 (Wen, Wu, Shinkai, Irizarry, & Feinberg, 2009). A hypothesis is that combinatorial chromatin marks can poise genes for transcription, creating more flexible chromatin states ready to adjust for subtle changes in the microenvironment of regulated genes(Riddle et al., 2011)(Lanzuolo, Sardo, Orlando, & Drosophila, 2012). The presence of chromatin marks known to be silent on euchromatin regions or active genes is intriguing and future analysis will be particularly important for elucidating this question.

Conclusions

The ChIP protocol described herein provides reproducible results with two different diatom species grown in different media. The quality of the procedure monitored by the no antibody control is high as no or insignificant background noise was observed. This protocol is also rapid, and can be completed within 3 days. The quantity and quality of eluted DNA from immunoprecipitation are also satisfactory. Using the described ChIP method combined with real time quantitative PCR, we have demonstrated the existence in diatoms of two types of histone modifications, H3K4me2 and H3K9me2. The H3K4me2 mark is associated with transcriptionally active genes in *P. tricornutum* and *T. pseudonana*. This result is consistent with the distribution of H3K4me2 in plants and mammals. H3K9me2 binds TEs whereas no association with genes was detected in *P. tricornutum*. In *T. pseudonana*, H3K9me2 also correlated significantly with TEs, although it also appears to bind protein-coding genes. This is different from the distribution pattern of H3K9me2 in *P. tricornutum*. The differences of H3K9me2 distribution pattern between *P. tricornutum* and *T. pseudonana* may be inherent to the genetic and/or epigenetic background of the two species which belong to pennate and centric diatoms, respectively. Gene enrichment with H3K9me2 could be further confirmed by a genome wide study of H3K9me2 distribution. Our results show that our experimental and data analysis approach are indeed highly sensitive to detect differences between the two

species if they do occur. The ChIP protocol we describe can be combined with microarray or massively parallel sequencing for further genome-wide studies. This protocol has been indeed successfully combined with Illumina sequencing for studies of global histone modifications in *P. tricornutum* (unpublished results). Furthermore, our ChIP assay can likely be easily adapted to other eukaryotic phytoplankton species for in vivo protein DNA interaction studies.

Materials and Methods

Harvesting cells and cross-linking

P. tricornutum and T. *pseudonana* cells were grown in 400 ml of artificial sea water (Vartanian, Desclés, Quinet, Douady, & Lopez, 2009) under 12/12 light dark period at 19°C until cell density reached around 1 million cells/ml. 11.27 ml of 36.5% of formaldehyde (Fluka cat. No. 200018) were added to the culture to get a final concentration of 1% in the medium. Cultures were then shaken for 10 min. Fixation was stopped by adding 2M glycine (final concentration 0.125M, Sigma cat. No. 241261) followed by incubation for 5 min at room temperature. Cells were washed twice with PBS solution. The supernatant was removed after centrifugation at 4000 rpm for 5 min at 4°C. At this stage, the fixed pellet can be stored at – 80°C for several months.

Chromatin extraction and sonication

5 ml of Extraction buffer I (0.4 M sucrose, 1 mini tablet Roche per 50 ml (Roche cat. No. 11873 580 001), 10 mM $MgCl_2$, 5mM 2-mercaptoethanol, 10 mM Tris-HCl pH 8) were added to 50 ml culture pellet. Tubes were left on ice for 5 min followed by centrifugation at 4000 rpm for 20 minutes at 4°C. The supernatant was removed gently and the pellet was suspended in 1 ml of Extraction Buffer II (0.25M sucrose, 10mM Tris-HCl pH8, 10 mM MgCl2, 1% Triton X-100, 1 mini tablet Roche diluted in 1 ml (for 10 ml), 5mM 2- mercaptoethanol) and centrifuged at 10,000 rpm for 10 minutes at 4°C. The supernatant was removed and the pellet was resuspended in 300 µl of Extraction Buffer III (1.7 M sucrose, 1 minitablet diluted in 1ml (for 10 ml), 0.15% Triton X-100, 2 mM MgCl2, 5mM 2-mercaptoethanol, 10 mM Tris HCl pH8). 300µl of Extraction Buffer III were added to a clean Eppendorf tube, and 300µl solution (resuspended pellet from last step) was carefully added on the top of the clean 300µl

of Extraction Buffer III. The Eppendorf tubes were then centrifuged at 13,000 rpm for 1 hour at 4 degrees.

The supernatant was removed and the chromatin pellet was resuspended in 300µl (or 200µl if small pellet) of Nuclei Lysis Buffer (50 mM Tris HCl pH 8, 10 mM EDTA, 1 mini tablet of protease inhibitors diluted in 1ml (for 10 ml), 1% SDS). The pellet should be resuspended by pipetting up and down and vortexing (keep solution cold between vortexing).

The chromatin solution was sonicated for 9 cycles, 30 seconds ON and 1 minute OFF for each cycle on full power. 5 µl were kept for DNA extraction to check sonication efficiency.

DNA was extracted after reverse cross-linking and checked on a 1% agarose gel. The DNA fragment should be around 200 bp-1000 bp. Sonicated chromatin can be frozen at –80°C for several months or can be used directly for immunoprecipitation.

Immunoprecipitation and reverse cross-linking

The chromatin solution was centrifuged at 13,000 rpm for 5 minutes at 4 degrees to pellet debris, and the supernatant was transferred to a new tube. At this step, 20 µl were removed for input control. The remaining volume of sonicated chromatin was measured and the volume brought to up to 3ml with ChIP Dilution Buffer (1% triton, 1.2mM EDTA, 167 mM NaCl, 16.7 mM Tris HCl pH8). The chromatin solution was split into 3 tubes (1ml each). Tube 1 contains DNA A beads labeled H3K4me2, tube 2 contains DNA B beads labeled H3K9me2, tube 3 contains DNA C beads labeled No Antibody (mock).

For each IP, 45 µl of DNA beads A (Invitrogen cat. No. 100.02D) and 45 µl DNA beads G (Invitrogen cat. No. 100.04D) were mixed in siliconized tubes, then washed twice with ChIP dilution buffer and resuspended in 90 µl Chip dilution buffer.

For Ig capturing, 5 µl of each antibody were added to the siliconized Eppendorf tube that contained the mixed 60 µl of beads. For preclearing, the diluted chromatin solutions were added to the tubes that contained 30 µl mixed beads. All tubes were left with gentle rotation at 4°C for 2 hours. The beads-Ig complexes were washed once with 1 ml ChIP dilution buffer and resuspended in 60 µl ChIP dilution buffer. The precleared chromatin was transferred into the beads-Ig complexes and tubes were left with gentle rotation at 4°C overnight. Four washes

of DNA beads-Ig-antigen complexes were performed in sequence using Low Salt Wash (150mM NaCl, 0.1% SDS, 20mM Tris-HCl pH8, 2mM EDTA, 1% Triton X-100) Buffer, High Salt Wash Buffer (500mM NaCl, 0.1% SDS, 1% Triton X-100, 20mM Tris-HCl pH8, 2mM EDTA), LiCl Wash Buffer (0.25M LiCL, 1% IGEPAL CA-630, 10Mm Tris-HCL pH8, 1mM EDTA, 1% sodium deoxycholate) and TE Buffer (10Mm Tris-HCl pH8, 1Mm EDTA). Washes were done twice with 1 ml of each buffer. The first wash was quick without agitation and the second one was for 5 min with gentle rotation at 4°C.

Immunoprecipitated complexes were eluted by adding 250µl of Elution Buffer (1% SDS, 0.1M $NaHCO_3$) to the washed beads, followed by a brief vortex for mixing and incubated at 65 °C for 15 min (tubes were mixed during incubation). Tubes were then put on the magnet DynaMag (Invitrogen) and the eluate was carefully transferred to another Eppendorf tube. The elution was repeated and elaute combined to obtain 500 µl. Reverse cross-linking was performed by adding 20 µl of 5M NaCl to the eluate including the total DNA (input) and incubation at 65°C overnight.

DNA recovery

To resume reverse cross-linking, 10 µl of 0.5M EDTA, 20µl Tris-HCl 1M (pH 6.5), 2µl of 10mg/ml proteinase K, and 1 µl of 10mg/ml RNase were added to the eluate and incubated for one hour at 45°C. DNA was recovered by phenol and chloroform extraction (phenol; phenol/chloroform 1:1; chloroform). DNA was precipitated with ethanol (2 volumes and 1/2) and 1/10 volume sodium acetate (3M pH 5.3). 2µl of glycogen (20mg/ml) were added to the ethanol precipitation step before incubation for 2 hours at -20°C. The tubes were centrifuged at 4°C at 13,000 rpm for 30 min and pellets were washed with 70% ethanol, dried at room temperature, and resuspended in 50µl of distilled water.

Quantitative PCR

For both diatoms, specific primers were designed for two genes and a set of TEs. The input DNA pulled out from ChIP was diluted 10 times before q-PCR. Quantitative PCR was performed using a Roche LightCycler® 480 machine on IP, input and mock DNA which were mixed to 5 µl LightCycler® DNA Master SYBR Green I 2X, 3 µl forward/reverse primers 1

μM, and 1 μL H2O. The PCR program was performed as follows: 10 minutes at 95°C; 45 cycles of 95°C for 15 seconds and 60°C for 1 minute.

Data analysis

The Ct value (number of cycles required for the fluorescent signal to cross the threshold) is recorded in the experimental report after analysis by Roche LightCycler® 480 software. The Ct values of the duplicates should show minimal variability, indicating that samples were properly handled (ideally, it should be below 0.2). Ct values were used for performing the calculation which consists on evaluating the fold difference between experimental sample and normalized input.

ΔCt (normalized to the input samples) value for each sample. ΔCt [normalized ChIP] = (Ct [ChIP] - (Ct [Input] - Log2 (Input Dilution Factor).

Where Input Dilution Factor = (fraction of the input chromatin saved)$^{-1}$ × Input dilution factor before q PCR. Here the fraction of Input chromatin saved is 20μl and the fraction for each IP is 90 μl. The IP fraction is 4.5 times the input fraction. For QPCR runs Input was diluted 10 times which makes the final dilution factor of the Input fraction (Input Dilution Factor) = 4.5x 10 = 45. Then the equation above is as follows: ΔCt [normalized ChIP] = (Ct [ChIP] - (Ct [Input] - Log2 (45). Finally, the percentage (Input %) value for each sample is calculated as follows: Input %=100/ 2 $^{\Delta Ct\ [normalized\ ChIP]}$. The “Input %” value represents the enrichment of certain histone modification on specific region.

Peptide competition assay and western blotting

The antibody was pre-incubated with the peptide prior to use in immunoblotting assays. Different amounts (depends on different peptides and antibody) of peptides were added to a 10 ml BSA solution containing antibody and incubated under gentle agitation for 4 h at room temperature and an additional 1 hour at 4°C before the immunoblotting assays. The antibodies and peptides used in this work include H3K4me2 (Millipore Ref: 07-030), H3K9me2 (Millipore, Ref: 17-681), H3K4me1 peptide (Abcam Ref: ab1340), H3K4me2 peptide (Abcam Ref: ab7768), H3K4me3 peptide (Abcam Ref: ab1342), H3K9me1 peptide (Millipore, Ref: 12-569), H3K9me2 peptide (Millipore, Ref: 12-430), H3K9me3 peptide (Millipore, Ref**:** 12-568). Different concentrations of antibody and peptide concentration were compared. The

nuclear enriched protein used for immunoblotting was extracted following the chromatin extraction protocol with minor modifications: the culture was not fixed by formaldehyde and sonication was not needed.

References

1. Round, F. E. RMC et al (1992) *The Diatoms: biology & morphology of the genera* (Cambridge [England] ; New York, Cambridge University Press).

2. David M.Nelson, Paul Tréguer, Mark A. Brzezinski, Aude Leynaert and BQ (1995) Production and dissolution of biogenic silica in the ocean: revised global estimates, comparison with regional data and relationship to biogenic sedimentation. *GLOBAL BIOGEOCHEMICAL CYCLE* 9:359-372.

3. Armbrust EV et al. (2004) The genome of the diatom Thalassiosira pseudonana: ecology, evolution, and metabolism. *Science (New York, N.Y.)* 306:79-86.

4. Bowler C et al. (2008) The Phaeodactylum genome reveals the evolutionary history of diatom genomes. *Nature* 456:239-44.

5. Siaut M et al. (2007) Molecular toolbox for studying diatom biology in Phaeodactylum tricornutum. *Gene* 406:23-35.

6. De Riso V et al. (2009) Gene silencing in the marine diatom Phaeodactylum tricornutum. *Nucleic acids research* 37:e96.

7. Gendrel A-V, Lippman Z, Martienssen R, Colot V (2005) Profiling histone modification patterns in plants using genomic tiling microarrays. *Nature methods* 2:213-8.

8. Grably M ED (2010) A detailed protocol for chromatin immunoprecipitation in the yeast Saccharomyces cerevisiae. *Methods Mol Biol.* 638:211-224.

9. Saleh A, Alvarez-Venegas R, Avramova Z (2008) An efficient chromatin immunoprecipitation (ChIP) protocol for studying histone modifications in Arabidopsis plants. *Nature protocols* 3:1018-25.

10. Nelson JD, Denisenko O, Bomsztyk K (2006) Protocol for the fast chromatin immunoprecipitation (ChIP) method. *Nature protocols* 1:179-85.

11. Kim B-H et al. (2012) Simple, Rapid and Cost-Effective Method for High Quality Nucleic Acids Extraction from Different Strains of Botryococcus braunii. *PLoS ONE* 7:e37770.

12. Liu CL et al. (2005) Single-nucleosome mapping of histone modifications in S. cerevisiae. *PLoS biology* 3:e328.

13. Mikkelsen TS et al. (2007) Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. *Nature* 448:553-60.

14. Zhang X, Bernatavichute YV, Cokus S, Pellegrini M, Jacobsen SE (2009) Genome-wide analysis of mono-, di- and trimethylation of histone H3 lysine 4 in Arabidopsis thaliana. *Genome biology* 10:R62.

15. Maumus F et al. (2009) Potential impact of stress activated retrotransposons on genome evolution in a marine diatom. *BMC Genomics* 10:624.

16. Zhou J et al. (2010) Genome-wide profiling of histone H3 lysine 9 acetylation and dimethylation in Arabidopsis reveals correlation between multiple histone marks and gene expression. *Plant molecular biology* 72:585-95.

17. Tachibana M et al. (2002) G9a histone methyltransferase plays a dominant role in euchromatic histone H3 lysine 9 methylation and is essential for early embryogenesis. 3:1779-1791.

18. Bernstein BE et al. (2006) A bivalent chromatin structure marks key developmental genes in embryonic stem cells. *Cell* 125:315-26.

19. Weishaupt H, Sigvardsson M, Attema JL (2010) Epigenetic chromatin states uniquely define the developmental plasticity of murine hematopoietic stem cells. *Blood* 115:247-56.

20. Roudier F et al. (2011) Integrative epigenomic mapping defines four main chromatin states in Arabidopsis. *The EMBO journal* 30:1928-38.

21. Bapat S a. et al. (2010) Multivalent epigenetic marks confer microenvironment-responsive epigenetic plasticity to ovarian cancer cells. *Epigenetics* 5:716-729.

22. Riddle NC et al. (2011) Plasticity in patterns of histone modifications and chromosomal proteins in Drosophila heterochromatin. *Genome research* 21:147-63.

23. Wen B, Wu H, Shinkai Y, Irizarry R a, Feinberg AP (2009) Large histone H3 lysine 9 dimethylated chromatin blocks distinguish differentiated from embryonic stem cells. *Nature genetics* 41:246-50.

24. Lanzuolo C, Sardo FL, Orlando V, Drosophila I (2012) Concerted epigenetic signatures inheritance at PcG targets through replication. *cell cycle* 11:1296-1300.

25. Vartanian M, Desclés J, Quinet M, Douady S, Lopez PJ (2009) Plasticity and robustness of pattern formation in the model diatom Phaeodactylum tricornutum. *The New phytologist* 182:429-42.

Table 1. Primer sequences used for QPCR analysis

Primer name	Sequence
Tp H4 Promoter fw	AGCCTGATGGAGAGAGTGGA
Tp H4 Promoter rev	TACATCCAGGACCTCCGTTC
Tp H4 body fw	TCGTAGAAAACGGTCCCATC
Tp H4 body rev	CCTCCACTTGGAAGAAGCAG
Tp Phyto promoter fw	CGATGTTGGTTGAGTGTTGG
Tp Phyto promoter rev	GCGATGTGCTCTTTTTGACA
Tp Phyto body fwd	TTTGGATTCCGTGAGAAAGG
Tp Phyto body rev	GTCTCGTGCACTCATCTCCA
Tp EU432485 fw	CCAGAGCTCGACAAACATGA
Tp EU432485 rev	TCGTTTTCCCTACGTGGAAC
Tp EU432490 fw	AGGAACTCGGAGACAAAGCA
Tp EU432490 rev	ATGTGCCCTCTTCAACAACC
Tp EU432492 fw	GCTCTGTCGTCGGAAAACTC
Tp EU432492 rev	AGGACAGCCTGCGTAGAAAA
Tp EU432500 fw	TGATGCAACAGGACGAAGAG
Tp EU432500 rev	GCATTGTTGGCCTTGTACCT
Pt H4 promoter fw	GTTGGTCGTCCATCGTTAGC
Pt H4 promoter rev	CCGTGGACGTTCTTGGTAGT
Pt H4 body fw	AATTACCAAGCCCGCTATCC
Pt H4 body rev	GTTTCTGTGAAACGGCAGGT
Pt PHY promoter fw	CTTGCCATGTCTTTGCAGTG
Pt PHY promoter rev	GTCAACACGCAATCAAGCAC
Pt PHY body1 fw	CAGCGACGGAAATGGACTAC
Pt PHY body1 rev	TTAGCAAGCAAGTGCGTCAG
Pt BKB4022 fw	CGAAGCTACTATGCCGGAAG
Pt BKB4022 rev	AAGGACACGAGAGTCGAGGA
PTC30 fw	CGGACTTCACCGAAGACAAT
PTC30 rev	GAATGGCTTTGGCATCATCT
PTC66 fw	AGCGATGGAACATTGGTTTATC
PTC66 rev	AACGTATCGTGAGCCTGACC
PTC25 fw	GCCTACCCCATGAAAACTGA
PTC25 rev	AGGCTCACTCTGCCACTGAT
SCF Fw	CAGCCTGAGGCGAAAGATAC
SCF rev	TAGTTCTGACATGCGCCAAG

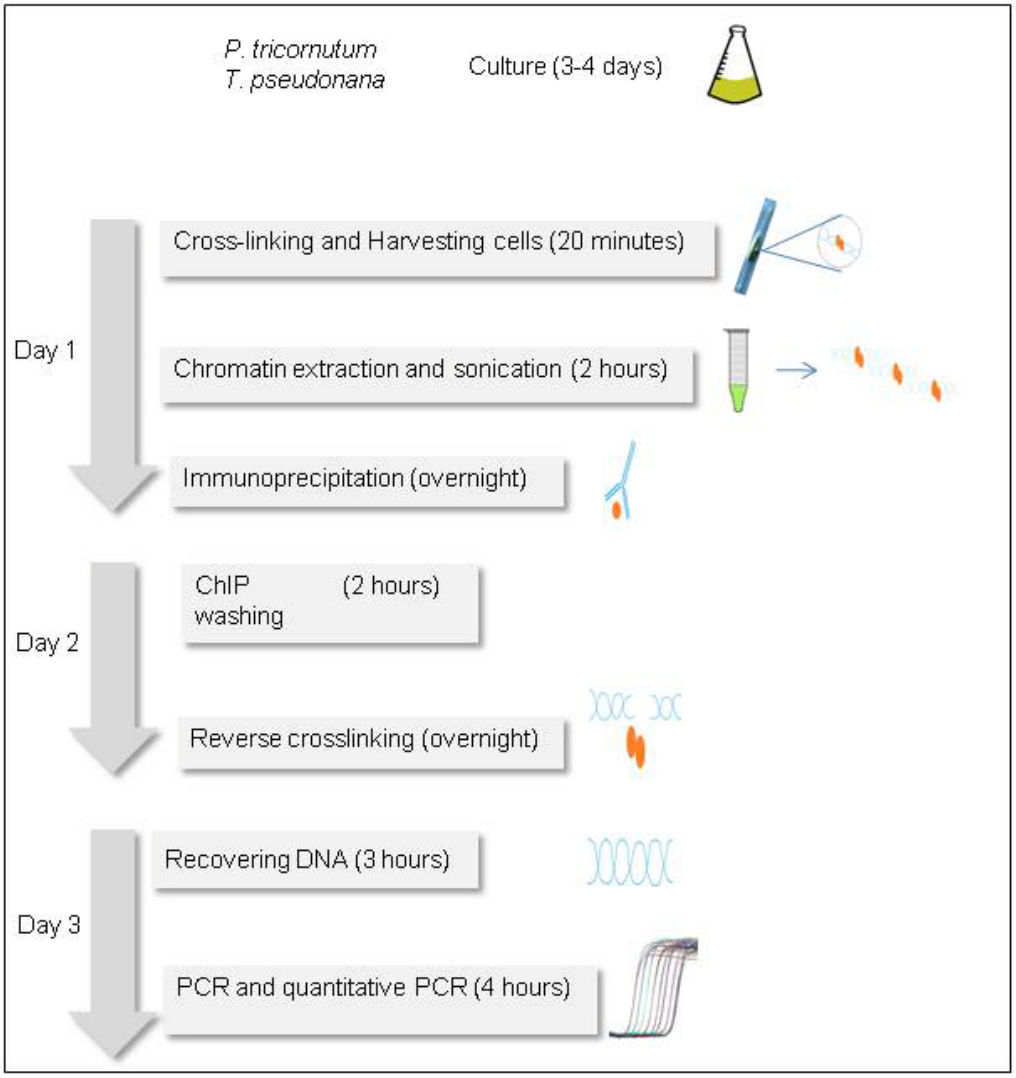


Figure 1 Outline of the ChIP QPCR protocol. Timing for each step is indicated between parentheses.

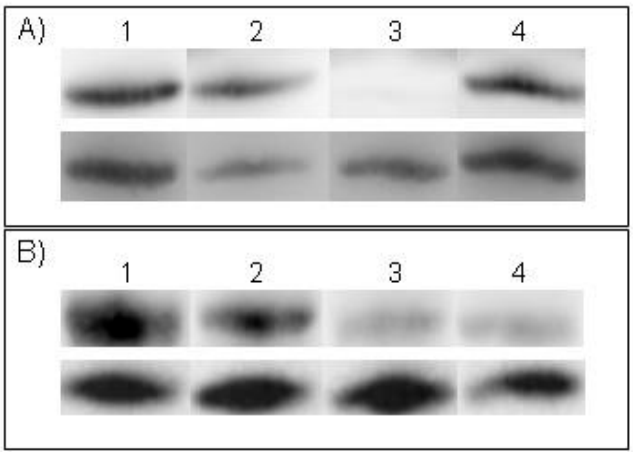


Figure 2 Analysis of antibody specificity using peptide competition assays on western blots of *P. tricornutum* nuclear extracts. (A**)** Upper lane contains from left to right 1 µg of H3K4me2 antibody alone or with 0.25 µg of one of the different modified peptides, H3K4me1, H3K4me2 and H3K4me3. Lower lane contains H4 antibody as internal loading control. (B) Upper lane contains 1 µg of H3K9me2 antibody alone or with 0.25 µg of one of the different modified peptides, H3K9me1, H3K9me2 and H3K9me3. Lower lane contains H4 antibody as internal loading control.

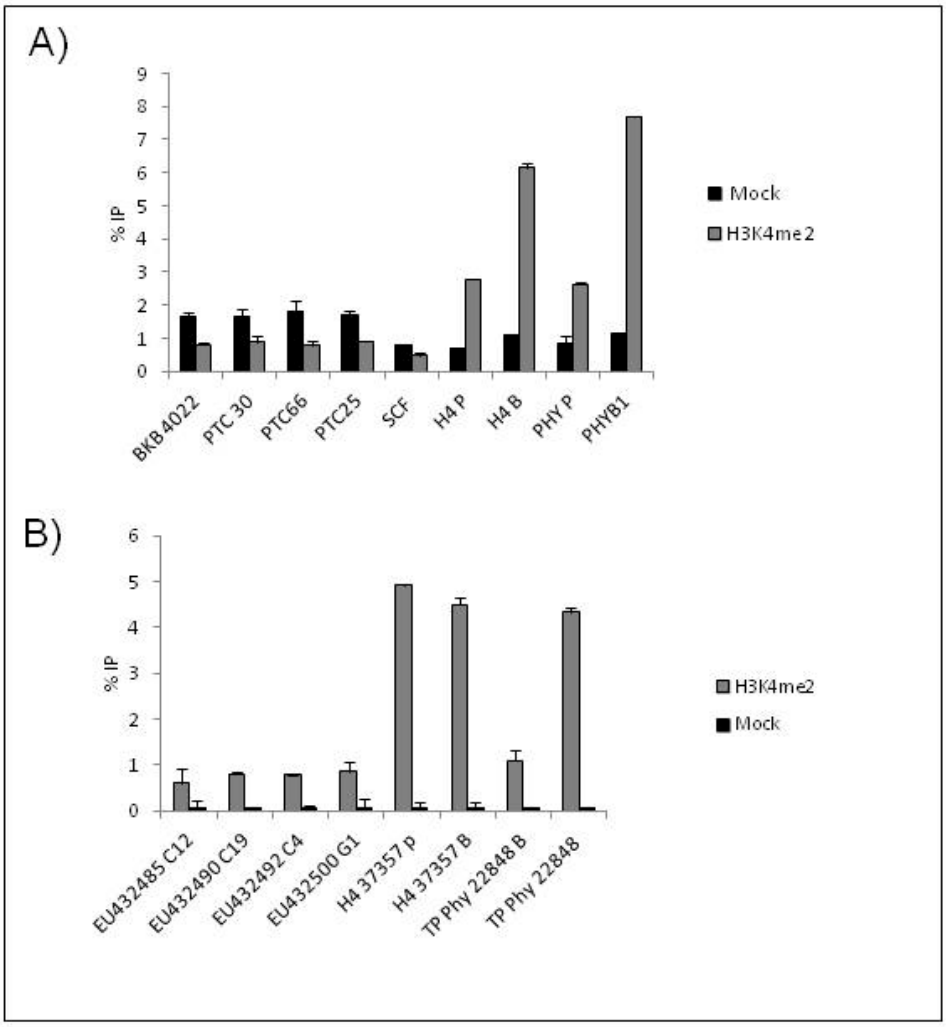


Figure 3 H3K4me2 histone modification on different regions of genes and TEs in *P. tricornutum* (A) and *T. pseudonana* (B). % IP indicates the enrichment. H4 P: promoter region of H4 histone gene. H4 B: body region of H4 histone gene. PHY P: promoter region of Dph gene. PHY B: body region of Dph.

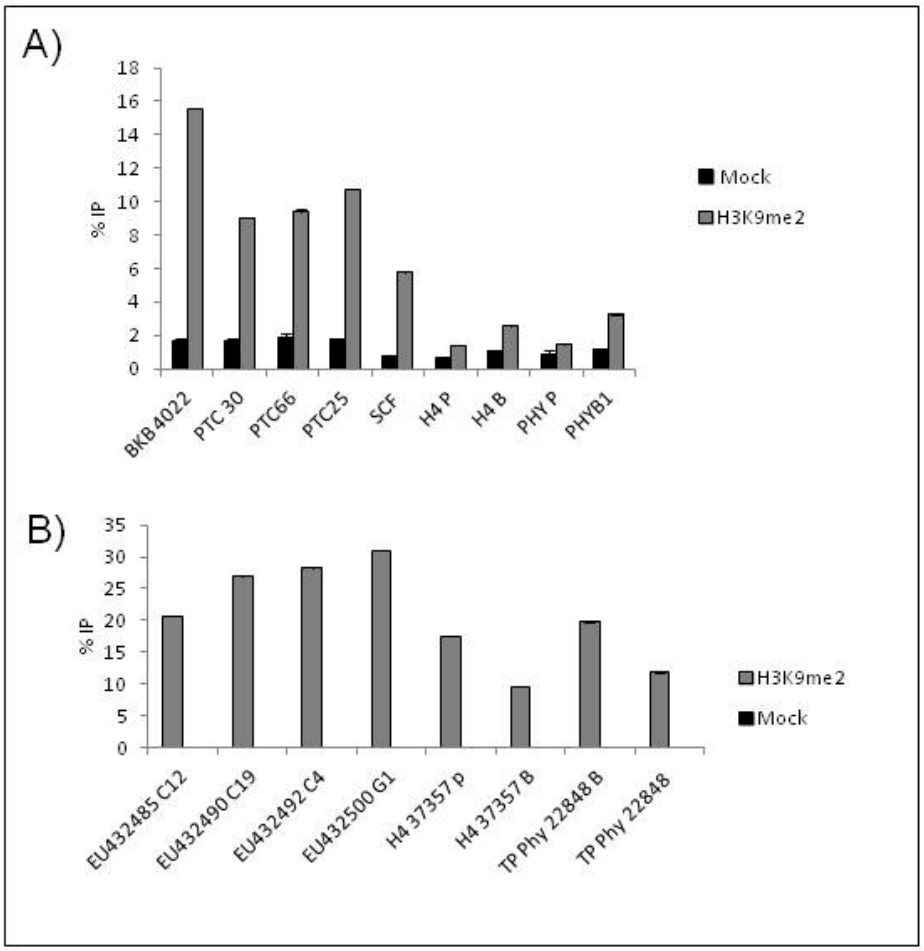


Figure 4 H3K9me2 histone modification on different regions of genes and TEs in *P. tricornutum*. (A) and *T. pseudonana* (B). % IP indicates the enrichment. H4 P: promoter region of H4 histone gene. H4 B: body region of H4 histone gene. PHY P: promoter region of Dph gene. PHY B: body region of Dph.

Chapter IV

Putative epigenetic components and reverse genetic approach for dissecting epigenetic machineries in *Phaeodactylum tricornutum*.

Tables of contents

Chapter IV 187

Putative epigenetic components and reverse genetic approach for dissecting epigenetic machineries in *Phaeodactylum tricornutum* 187

4.1 Introduction 191

4.1.1 C5-DNA methylation and its machinery 191

4.1.2 N6- adenine methylation 197

4.1.3 C5-DNA demethylation machinery 198

4.1.4 SET domain containing proteins for histone lysine methylation 199

4.1.4.1 Polycomb group proteins 199

4.2 Results 204

4.2.1 Putative C5-Methyl transferases in diatoms 204

4.2.2 Genome wide DNA methylation quantification of wild type and C5-MTases mutants 206

4.2.3 DNA demethylation machinery in diatoms 212

4.2.4 Identification of proteins containing SET domains for histone methylation in *P. tricornutum* 215

4.2.5 Distribution of PRC2 core subunits in diatoms and other eukaryotic algae 217

4.2.6 Phylogentic analyses of PRC2 component E(z) 219

4.2.7 Phylogentic analyses of PRC2 components ESC 219

4.2.8 Phylogentic analyses of PRC2 components Suz(12) 220

4.2.9 Distribution of PRC1 core subunits in diatoms and other eukaryotic algae 223

4.2.10 Knockdown of E(z) homolog 32817 in*P. tricornutum* 226

4.2.11 Effect of EZH down regulation on global post-translational histones modifications 226

4.2.12 Light microscopy observations of Ez antisencse strains 228

4.3 Discussion 230

4.3.1 C5-MTase knockdown in *P. tricornutum* 230

4.3.2 Distribution of PcG components in algae and E(z) knockdown in *P. tricornutum* 230

4.4 Materials and methods 233

4.4.1 Cell culture 233

4.4.2 Antisense vector construction 233

4.4.3 Genetic transformation of *P. tricornutum* 234

4.4.4 DNA methylation quantification 234

4.4.5 Analysis of DNA methylation by McrBC-qPCR 235

4.4.6 Western blotting 238

4.4.7 Phylogenetic analyse 238

4.5 References 239

4.1 Introduction

DNA methylation and histone modifications are two major epigenetic components and the machineries involved in their regulation in different organisms are extensively being studied (Cedar & Bergman, 2009). The genome wide distributions of DNA methylation and several histone modifications in *P.tricornutum* are described in Chapter II and Chapter III. In this chapter, I focus on the epigenetic machineries of *P. tricornutum*, in particular the mainly include C5-DNA methylation machinery, N6 adenine methylation machinery, C5-demethylation machinery and histone lysine methyltransferases, with a focus on Polycomb group proteins which play crucial roles in H3K27me3 deposition. I also utilize a reverse genetic approach with several C5-MTase genes and the Ez (Enhancer of Zester) gene for exploring their roles in *P. tricornutum* which is a useful start point for dissecting the epigenetic machinery of diatoms.

4.1.1 C5-DNA methylation and its machinery

DNA methylation is the most extensively studied epigenetic mark. DNA methylation was first found to have a major role in bacteria, defending against phage DNA by DNA methylation of restriction sites that would otherwise be recognized and cleaved. DNA methylation was found in different lineages of eukaryotes including fungi, plants, and animals. In bacteria, DNA methylation includes N6-methyladenine (m6A) and N4-methylanine (m4A). However, only C5 and N6 methylation are found in eukaryotes. Based on retrospective research, it seems that C5 methylation has a more profound role than N6 in eukaryotes (Iyer, Abhiman, & Aravind, 2011).

The context of DNA methylation in plants and animals is different. In animals, initially only CG context methylation was found. However, recent studies showed that non CG methylation also exists in stem cells (Lister et al., 2009). In plants not only CG methylation but also CHG and CHH methylation are found. The propagation of DNA methylation in plants and animals is also different: in plants DNA methylation can propagate across generations, whereas in animals the DNA methylation mark is removed during zygote formation and re-established through cell division during development. Recently some studies showed that a “memory system” probably regulates the re-establishment of DNA methylation during development. However, how the “memory system” manipulates the whole process is still obscure. There are

two types of DNA methylation mechanisms: *de novo* DNA methylation and maintenance DNA methylation. *De novo* methylation mechanism acts by adding a methyl group to the unmethylated DNA residues. Maintenance methylation mechanism is responsible for maintaining the methylation status of the newly synthesized DNA strand during methylated DNA replication. Mechanisms of DNA methylation in plants and animals are also not the same. In plants, most of the *de novo*DNA methylation is guided by small RNAs in a process known as RNA-dependent DNA methylation (RdDM), while RdDM machinery has not been found in animals so far. The group of enzymes that can add methyl to the cytosine are called cytosine DNA methyltransferases (C5-MTase). The mechanisms of C5-MTases have been studied for a long time. Application of forward and reverse genetic approaches on C5-MTases in model plant and animal species has greatly improved our understanding of the role of C5-MTases in development and gene regulation.

C5-MTases involved in catalyzing DNA methylation vary from different kingdoms of life. The first C5-MTase was found in *Escherichia coli*(Hermann, Gowher, & Jeltsch, 2004) but a large number of eukaryotic DNA methyltransferase homologs have now been reported. Some of them, especially those from model species have been shown to methylate DNA *in vitro* or be involved in cytosine methylation. So far, in mammals, at least five C5-MTases: DNMT1, DNMT2, DNMT3a, DNMT3b, and DNMT3L have been identified. The first mammalian MTase discovered was Dnmt1, which is conserved among eukaryotes and is responsible for DNA maintenance methylation (Bestor, 2000). DNMT2 is highly conserved among all the eukaryotes including the organisms that lack DNA methylation such as *Schizosaccharomyces pombe*. DNMT2 has not been identified as an active C5-MTase though it has highly conserved catalytic motifs. This is due to the insertion of a serine residue into a critical prolinecysteine dipeptide that is essential for DNA methyltransferase activity(Schaefer et al., 2010).

DNMT3a, DNMT3b encode highly related but different C5-MTases. *de novo* DNA methylation catalytic activities of DNMT3a and DNMT3b *in vivo* have been detected (Hsieh, 1999; F Lyko et al., 1999). The DNMT3L gene encodes DNA (cytosine-5)-methyltransferase 3-like protein which does not contain the amino acid residues necessary for methyltransferase activity. However, DNMT3L as a stimulatory factor can modulate *de novo* methylation by DNMT3A and DNMT3B (Chedin et al., 2002).

In the model plant *A. thaliana*, C5-MTases include methyltransferases 1 (MET1), chromomethyltransferases3 (CMT3), domains rearranged methyltransferases (DRM) and DNMT2. MET1 is the homolog of DNMT1 in mammals and considered to be responsible for the CG maintenance DNA methylation required for most TEs and gene methylation. CMT3 and DRM are plant specific C5-MTases. CMT3 is presumed to maintain CHG DNA methylation, while DRM appears to be the principal *de novo* methyltransferase implicated in RdDM.

In fungus *Neurospo*ra, DIM-2 and RIP defective (RID) were found to have similar roles than C5-MTases (Freitag, Williams, Kothe, & Selker, 2002). Besides the C5-MTases mentioned above, two super families of C5-MTases, DNMT5 and DNMT6, were identified. DNMT5 was found in *Aspergillus fumigates* (Ascomycetes) and *Cryptococcus neoformans* (Basidomycetes) (Ponger & Li, 2005).

C5-MTases can be divided into an N-terminal regulatory part and a C-terminal catalytic domain (**Figure 4.1**). All these C5-MTases share the common structure of the catalytic domain in both prokaryotes and eukaryotes characterized by 10 conserved amino acid motifs involved in the catalytic function (Goll & Bestor, 2005). The classification of C5-MTases is based on the sequence homology within their C-terminal catalytic domains. Dnmt1 contains a large N-terminal regulatory part which comprises different motifs such as PCNA-NLS domain, DNA replication loci target sequence, a Cys-rich region and BAH (Bromo adjacent homology) domain (**Figure 4.1**). In plants, MET1 contains a large N-terminal regulatory part resembling its homolog DNMT1. All the DNMT2 homologues are much shorter because they lack the N-terminal domains. DNMT3A and DNMT3B have similar structures toDNMT1 but the protein length is shorter than DNMT1. DNMT3A and DNMT3B contain PWWP and Csy-rich domain in N-terminal regions. DRM is confined to flowering plants based on current research. The RNAi machinery involving DRM is also found in plants. DRM proteins show an average of 28% amino acid identity to mammalian Dnmt3A and Dnmt3B protein in their C-terminal part containing highly conserved catalytic motifs. CMT3, another plant specific C5-MTase,is a chromomethylase characterised by the presence of a chromodomain (Bartee, Malagnac, & Bender, 2001) (**Figure 4.1**). DIM-2 has a C-terminal domain which is the homolog of other C5-MTases and a novel N-terminal tail that bears a degenerate BAH domain and an ATP/GTP-binding motif (Kouzminova & Selker, 2001). The catalytic C-

terminus of DIM-2 shows distant similarity to MET1 and Dnmt1 proteins and is likely to be a highly diverged member of the Dnmt1 family. DNMT5 proteins share a long C-terminal part comprising a DEXDc domain and helicase domain (**Figure 4.1**).

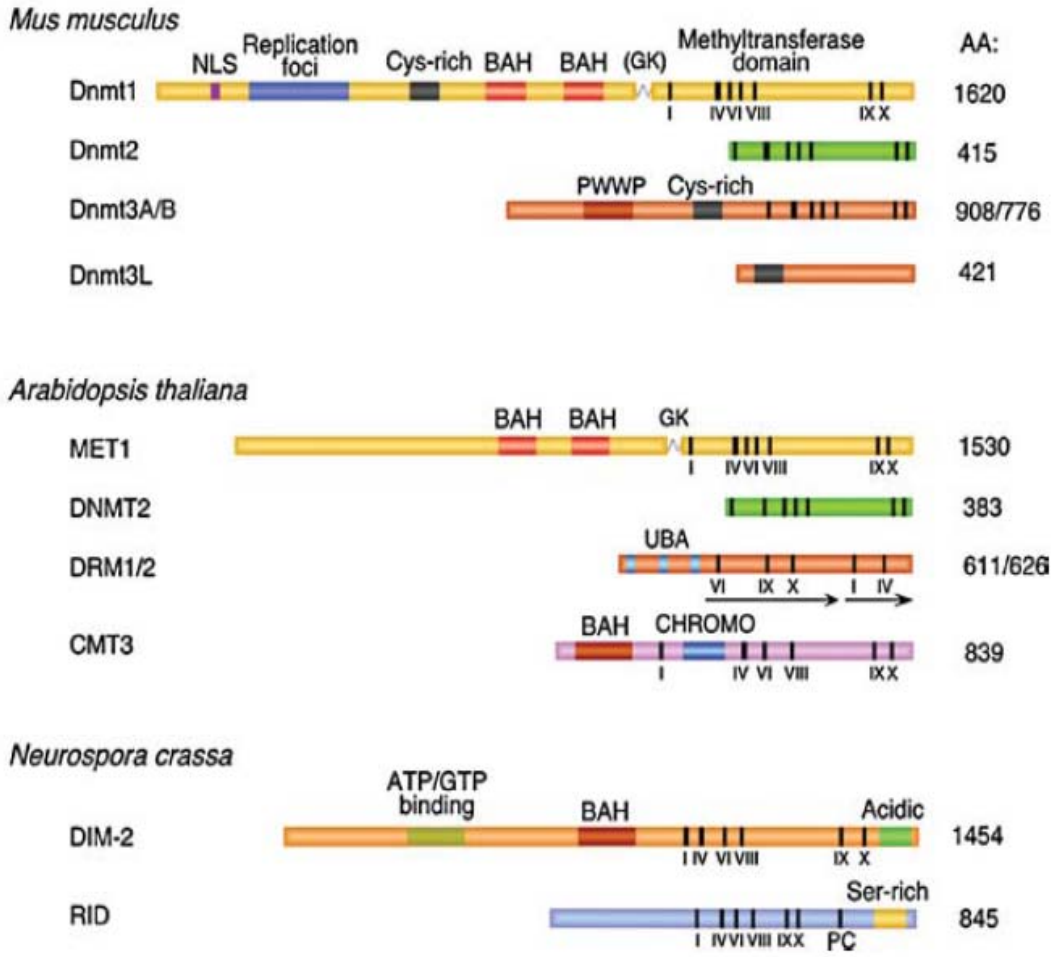


Figure 4.1 Organization and sequence relationships of DNA cytosine methyltransferases. Classes of cytosine methyltransferases and cytosine methyltransferase homologues in mouse, *A. thaliana*, and ascomycete fungus *N. crassa*. The positions of sequence motifs are indicated (Goll & Bestor, 2005).

The functions of C5-MTases in model animal and plant species have been extensively studied. Mounting evidence has showed that C5-MTases play vital roles in biological processes. DNMT1 deficient mice show a genome wide decrease of ~90% in DNA methylation and die early in embryogenesis stage(Y. Li et al., 1992). Similarly DNMT1 deficient mouse embryonic stem cells die when differentiation was induced. Inducible DNMT1 knockout mouse fibroblast cells die within a few cell divisions after induction of DNMT1 knockout. All these data indicate that DMNT1 is indispensable for cell division and proliferation(Dudley, Revill, Whitby, Clayton, & Farrell, 2008). DNMT1 has also been found to bind p53 tumor suppressor protein and to co-localize in the nucleus which indicates its role in some checkpoint signaling mechanisms (Chin, Pradhan, & Este, 2005). Inactivation of DNMT3a and DNMT3b lead to early embryonic lethality in mice which indicates the vital roles of DNMT3a and DNMT3b in development (Okano et al., 1999). The disruption of DNMT3a and DNMT3b in germ cells accomplished by conditional knockout technology showed that DNMT3a and DNMT3L are required for DNA methylation on imprinted loc i(Ooi et al., 2007). Human DNMT3L has been shown to cooperate with DNMT3a and enhance its activity, which suggests that Dnmt3L is an important cofactor for Dnmt3a (Chedin et al., 2002).

A large body of plant C5-MTase functional research has been carried out in high flowering plant *A. thaliana* by knocking down or knocking out the C5-MTases. MET1, which is responsible for CG maintenance methylation in *A. thaliana,* is the most extensively studied plant C5-MTase. Nearly 60% of the methylated regions became demethylated with transcriptional activation of TEs and pseudo-genes residing in heterochromatic regions in *A. thaliana MET1* mutants (Xiaoyu Zhang et al., 2006). These results showed that CG maintenance methylation mediated by MET1 has a curial role in maintaining the stability of heterochromatic regions. As for the plant specific C5-MTase CMT3, *CTM3* single mutant shows little reduction in DNA methylation at CHG sites wherase the *drm* mutant showed almost complete loss of asymmetric CHH DNA methylation and a partial loss of CHG methylation. However, in the *drm1 drm2 cmt3* triple mutant, CHG methylation was lost. These data demonstrated that CHH asymmetric DNA methylation is mediated by DRM (Cao

et al., 2003). Both DRM and CMT3 act redundantly on CHG DNA methylation(Cao et al., 2003). So far, knockdown and knockout of DNMT2 in different organisms did not lead to disruption of DNA methylation, which implies that DNMT2 has no or very low C5-MTase catalytic activity. Recently DNMT2 was found to be a RNA methyltransferase for tRNA (Jurkowski et al., 2008).

The genome wide distribution of DNA methylation in *P. tricornutum* combined with gene expression has been profiled, which revealed the characteristics of DNA methylation distribution in *P. tricornutum* and eventually some hints on its putative role. However, nothing is known about the mechanisms regulating DNA methylation in *P. tricornutum*. The questions are: how do the putative methylases act, what does each methylase do in terms of maintenance and *de novo* methylation and is DNA methylation function is conserved between *P. tricornutum* and other species? In an attempt to help answering some of these questions, I initiated a reverse genetic approach by knocking down some of the known putative C5-MTases in *P. tricornutum*. The generated knockdown mutants were characterized.

4.1.2 N6- adenine methylation

N6-methyladenine (m6A) is responsible for another DNA modification, which is the transfer of an adenine on the NH_2 group attached to the 6^{th} position of the purine ring. This is the most frequent DNA methylation after m5C and has been found in various eukaryotes and prokaryotes. m6A was first found in *Escherichia coli* (Dunn & Smith, 1958), andthen in lower eukaryotes such as *Penicillium chrysogenum* (Rogers, Rogers, Saunders, & Holt, 1986), the green alga *Chlamydomonas reinhardtii* (Hattman, Kenny, Berger, & Pratt, 1978) and several ciliates (Hattman, 2005). Interestingly, the nuclear genome of the ciliated protozoa *Tetrahymena thermophila* contains N6-methyladenine but no C5-methylation (Karrer & VanNuland, 2002). Unfortunately, the roles of m6A in lower eukaryotes are still not characterized.

In higher plants, m6A was found in nuclear DNA, plastid DNA and heavy mitochondrial DNA (one strand rich in guanine is referred to as the heavy strand). m6A in heavy mtDNA(the guanine-rich strand) is also detected in archegoniates (mosses, fern etc) (Vanyushin, 2005). The presence of m6A in heavy mtDNA which commonly exists in aging plant organs and apoptotic cells indicates that it may play a role in regulating mtDNA

replication and mt gene expression. Studies of m6A lag far behind those of mC5. The putative genes coding N6-methylation could be detected in *P. tricornutum*, *T. pseudonana* and *F.cylindrus* (Pt44651, Tp21517 and Fc216468), so N6-methylation might exist in diatoms.

4.1.3 C5-DNA demethylation machinery

Beside the m5C DNA methylatransferases, a system for removing methylation from cytosine called m5C demethylation is also vital for keeping m5C in a dynamic balance. DNA demethylation can be passive and active. Passive demethylation occurs both in plants and animals when cells fail to maintain DNA methylation in both strands during DNA replication. For active demethylation, it has been shown that plants use a family of 5Mc glycosylases including DEMETER (DME), DEMETER-LIKE 2 and 3 (DML2 and DML3), and REPRESSOR OF SILENCING 1 (ROS1) which remove m5C and create an abasic site. The gap is then refilled with an unmethylated cytosine through a base-excision-repair (BER) pathway (J.-K. Zhu, 2009; Ooi & Bestor, 2008). However, glycosylase homologs have not been found in mammals and the active demethylation system in animals is still mysterious. Active demethylation in mammals is proposed to result from mechanism different from the ones known for plants.

It has been proposed that methyltransferases DNMT3A and DNMT3B mediate oxidative deamination on 5Mc followed by repair of the guanine:thymine (G:T) mismatch by thymine DNA glycosylase (TDG) and methyl-CpG-binding domain protein 4 (MBD4). The BER machinery then returns the mismatch to an unmethylated guanine: cytosine (G:C) pair (Kangaspeska et al., 2008; Métivier et al., 2008). The mechanism requires that DNMT3A and DNMT3B are efficient m5C deaminases. However, the efficiency of deamination of m5C by DNMT3A and DNMT3B is not good enough as it is proposed to make the reported rapid cyclical methylation and demethylation (Ooi & Bestor, 2008).

Recently, studies on a small family of 5-methylcytosine hydroxylases (TET1, TET2, TET3; “TET” refers to Ten-Eleven-Translocation) in mammals have revealed an oxidative mechanism, implying that BER mechanism is preceded by oxidation and/or deamination for demethylation (Gong & Zhu, 2011). Obviously, this demethylation pathway is different from the glycosylase system in plants. The TET family enzymes, which are responsible for 5mC-OH, was also considered an important intermediate product of the demethylation system in

mammals (Ito et al., 2010). 5mC-OH can be a substrate for deamination enzymes, and the resultant hydroxymethyl uracil may be replaced by methyl cytosine through a threonine dehydrogenase (TDG) and BER pathways (Gong & Zhu, 2011). Furthermore, 5mC-OH was found to be involved in different biological processes such as DNA methylation fidelity, embryonic stem cell maintenance and probably transcriptional regulation (Ficz et al., 2011; Williams, Christensen, Pedersen, et al., 2011; Wu et al., 2011).

The Cytosine deaminases AID (activation-induced cytidinedeaminase) was found to be involved in demethylation in lower vertebrates and heterocharyon (Fritz & Papavasiliou, 2010). Primordial germ cells of AID deficient mouse mice display hypermethylation which showed that AID plays a role in erasing DNA methylation in mammals (Popp et al., 2010). Cytosine deaminases AID and Apobec1 (apolipoprotein B mRNA editing enzyme, catalytic polypeptide 1) can deaminate 5meC both *in vitro* and in *E. coli*, suggesting deamination of 5meC followed by T:G base excision repair by glycosylases such as Tdg or Mbd4 as an equivalent pathway for demethylation of DNA (Morgan, Dean, Coker, Reik, & Petersen-Mahrt, 2004).

4.1.4 SET domain containing proteins for histone lysine methylation

SET domain containing proteins are responsible for histone lysine methylation. The SET domain is folded in all the solved structures into several small β sheets surrounding a knot-like structure by threading of the carboxyl terminus through an opening of a short loop formed by a preceding stretch of the sequence. The two most-conserved sequence motifs in SET domains are RFINHXCXPN and ELXFDY(Dillon, Zhang, Trievel, & Cheng, 2005). MLL and ASH can methylate lysine 4 of histone H3 (H3K4). Suv 39 can add methyl to lysine9 (H3K9). E(z) has the enzymatic activity for adding methyl on H3K27me1 and H3K27me2 turning to H3K27me2 and H3K27me3. The enzymes catalyzing H3K27 and H3K4 methylation belong to Polycomb-group (PcG) and trithorax-group (trxG) protein complexes, respectively, which can mediate gene expression patterns in multicellular organisms (Schuettengruber, Martinez, Iovino, & Cavalli, 2011).

4.1.4.1 Polycomb group proteins

Polycomb group proteins, which are related to H3K27me3 deposition, have been extensively studied in different organisms. H3K27me3 has novel distribution in *P. tricornutum* which raised my interests in exploring the functions of Polycomb group proteins in diatom. Polycomb group (PcG) proteins were initially discovered in *Drosophila melanogaster*. *PcG D. melanogaster* mutants display improper body segmentation which suggests PcG proteins regulate homeotic genes that are required for segmentation (Raphaël Margueron & Reinberg, 2011). Polycomb group proteins are mainly constituted by polycomb repressive complex 1 (PRC1) and 2 (PRC2). The core PRC2 complex, conserved from *Drosophila* to mammals, contains four components: Enhancer of Zeste (E(Z)), Extra sex combs (ESC), Suppressor of zeste-12 (SU(Z)12), and nucleosome-remodeling factor 55 (NURF-55) (Simon & Kingston, 2009). In plants, PCR2 complexes have evolved into more complicated conformations. In *A. thaliana,* there are three homologs of E(z) (CURLY LEAF (CLF), MEDEA (MEA) and SWINGER (SWN)), three homolog of Su(z)12 (EMBRYONIC FLOWER 2 (EMF2), VERNALIZATION 2 (VRN2), and FERTILIZATION INDEPENDENT SEED 2 (FIS2)), the ESC homology FERTILIZATION INDEPENDENT ENDOSPERM (FIE) and Nurf55 homolog MULTICOPY SUPPRESSOR OF IRA 1(MSI1) (Hennig & Derkacheva, 2009).

Compared to PCR2, PRC1 components are more variable. The components of PRC1 are not evolutionarily conserved from plants to animals. In mammals, the PRC1 complex includes two core components RING1A/B together with BMI1, MEL18 (PCGF2) or NSPC1 (PCGF1) (Whitcomb, Basu, Allis, & Bernstein, 2007). In plants, LHP1 (Like heterochromatin protein) and AtRING1a are considered as PRC1 components. LHP1 does not have a homolog in *Drosophila*. *Drosophila* Pc is considered as a functional analogs but is not homologs to plant LHP1 (Hennig & Derkacheva, 2009).

It has been proved that the PcG proteins play roles in deciphering chromatin mechanisms in multicellular development, stem cell biology and cancer (Jaenisch & Young, 2008; Jones & Baylin, 2007; Lechner et al., 2010; Rajasekhar & Begemann, 2007). The PRC 2 proteins are important for early mouse development control. Mice with mutations in genes encoding PRC2 proteins lead to early embryonic lethality (Raphaël Margueron & Reinberg, 2011). Combined deletion of *Ring1b* and *Ring1a* cause inhibition of ES cell proliferation (Endoh et

al., 2008). In mouse embryonic stem (ES) cells, PRC2 is involved in maintenance of pluripotency by regulating gene expression, possible by repressing numerous developmental regulators inducing cell–lineage commitment or sustaining the expression of pluripotency factors. The PRC2 complex may directly bind to H3K27me3, most likely through the WD40 domain of ESC and its homolog FIE (Xu et al., 2010), thus ensuring a semi-conservative mode of maintaining the PRC2 mark.

Recent studies suggest that not only PRC2 but also PRC1 is implicated in pluripotency maintenance (Raphaël Margueron & Reinberg, 2011). In animals, the PRC1 complex monoubiquitinates H2A after H3K27 trimethylation deposition mediated by PRC2 complex and triggers compaction of the chromatin into a heterochromatin state that stably represses expression. However, the mechanism of PRC1 complex monoubiquitination in plants is still not clear (Holec & Berger, 2012).

The PcG machinery is also deployed in X-chromosome inactivation and parent-of-origin imprinting of epigenetic silencing in mammals. Both PRC1 and PRC2 proteins are recruited to the inactive X in a manner that depends on the non coding Xist RNA that is transcribed from the X chromosome locus that initiates X inactivation (Kerppola, 2009). The interaction and cooperation between non coding RNA and PcG proteins has been found in both plants and animals. In *A. thaliana*, a long intronic noncoding RNA termed COLD ASSISTED INTRONIC NONCODING RNA (COLDAIR) was found to corporate with PRC2 for triggering enrichment of tri-methylated histone H3 Lys27 at chromatin of the floral repressor, FLOWERING LOCUS C (FLC), and to result in epigenetically stable repression of FLC after a cold period (Heo & Sung, 2011). In humans, a 2.2 kilobase ncRNA residing in the HOXC locus termed HOTAIR, which represses transcription in trans-across 40 kilobases of the HOXD locus. HOTAIR interacts with PRC2 and is required for PRC2 occupancy and histone H3 lysine-27 trimethylation of the HOXD locus (Rinn et al., 2007).

PcG proteins are appreciated as a crucial set of global chromatin regulators because of the advances in understanding the molecular mechanisms and biological functions of PcG proteins in silencing. In *A. thaliana*, PcG proteins control the expression of the main transcrip- tional regulators of the central developmental switch: shoot apical meristems can be vegetative and generate more leaves or become reproductive and generate flowers. (Köhler &

Hennig, 2010). PcG proteins also regulate flowering by repressing the flowering inhibitor FLOWERING LOCUS C (FLC) after long periods of cold (vernalization) (Bastow et al., 2004). The complete loss of PcG function in *clf swn* double mutants leads to loss of cell differentiation and activation of the embryonic program (Chanvivattana et al., 2004). In moss *Physcomitrella patens*, PcG proteins (FIS and CLF) were found to have an impact on the between gametophyte-to-sporophyte transition (Chopra et al., 2011). Mosses have a predominant vegetative gametophytic phase and short sporophyte phase. In the gametophyte, FIS and CLF in *P. patens* have vital roles in repressing initiation of sporophytic pluripotent stem cell development.

The SET domain of E(z) (Enhancer of zeste), the component of PRC2, is the key domain for methylating H3K27. In metazoans, H3K27 is primarily mono- and di- methylated by PRC2 *in vivo* and H3K27 tri-methylation is catalyzed by a related complex (Pcl-PRC2), which contains additional subunits such as Polycomb-like (Pcl. In *A. thaliana*, monomethylation of H3K27 is carried out by ATXR5 and ATXR6 which are distinct from PRC2 and their homologues are absent in mammals (Jacob et al., 2009). VEL1 in *Arabidopsis* is also considered as the functional analog of *Drosophila* Pcl but not the homolog (Hennig & Derkacheva, 2009). It is still not clear which enzyme deposits monomethylation on H3K27 in mammals.

PRC2 complex is responsible for methylation of (di or tri) of H3K27 (H3K27me2/3) through EZH1 AND EZH2. In *Arabidopsis* seedlings, 15–24% of all genes are marked by H3K27me3 and most of them are repressed genes which implies the role of PcG proteins in gene silencing (Xiaoyu Zhang et al., 2007b). In *Xenopus*, H3K27me3 is associated with spatial gene regulation because H3K27me3 was found mainly deposited at promoters of genes that are preferentially expressed at vegetal poles. This suggests that H3K27me3 deposition and PcG proteins have crucial roles in gene regulation during development (Akkers et al., 2009).

It was initially believed that PcG proteins and Polycomb epigenetic mechanisms only exist in multicellular organisms because PcG proteins are absent in unicellular fungi *Schizosaccharomyces pombe* and *Saccharolyces cerevisiae*. However, the homology of PcG were found in filamentous fungus *Neurospora crassa* which led to speculation that PcG proteins may contribute to multicelluar developmental stages of this organism given that PcG proteins are critical for cellular differentiation (Whitcomb et al., 2007). Later Shaver et al

systematically demonstrated that PcG gene homologs exist in some unicellular organisms belonging to Opisthokonta, Chromalveolata and Archaeplastida. This identification of PRC2 subunits and H3K27methylation in extant unicellular organisms strengthened the hypothesis that PcG proteins were present in their last unicellular common ancestor and were subsequently lost in certain single-celled lineages (Shaver et al., 2010).

It is very intriguing to study the functions of PcG proteins in single celled eukaryotes. So far, to our knowledge, only few reports related to PcG protein in unicellular organisms have been published. In the unicellular ciliate *Tetrathymena thermophila*, the deposition of H3K27me3 is correlated with heterochromatin (Y. Liu et al., 2007). This suggests that PcG epigenetic system can be traced back at least to this ancient unicellular ciliate. Shaver et al knocked down the E(z) gene in *Chlamydomonas reinhardtii* by RNAi. They observed that E(z) RNAi resulted in the release of transcriptional silencing of tandemly repeated genes and retrotransposons. It was concluded that E(z) may be involved in a genomic defense response against invading foreign sequences. H3K4me3 and H4 acetylation were increased both of, which are mediated by depletion of E(z). Unfortunately, histone H 3 sequences around lysine 27 are not conserved in *C. reinhardtii* compared to plants and animals, so the commercial H3K27 methyl antibody could not be used for further investigation. In diatoms, PcG protein functions have never been explored. Here I investigate the putative PcG proteins in eukaryotic unicellular algae species with focus on diatom species especially *P. tricornutum* in which genome wide H3K27me3 distribution has been profiled.The preliminary results demonstrate that PcG proteins are widely distributed in unicellular organisms.

P. tricornutum as a single celled diatom model species with sequenced genome and molecular tools is a suitable model to explore the role of PcG proteins in unicellular organisms. Furthermore, *P. tricornutum* has the unique feature of possessing several ecotypes with different morphotypes which makes it an attractive system for studying the underlying mechanisms that govern morphological transition and whether PcG proteins have a potential role in *P. tricornutum* morphogenesis.

I therefore carried out a reverse genetic approach by knocking down the putative E(z) in *P. tricornutum* Pt32817. Several knockdown transgenic lines have been generated and their characterization is still under way. The amino acid sequence of H3 in *P. tricornutum* is highly

conserved compared to human and plants, so the use of the commercial antibody H3K27me3 is applicable to *P. tricornutum*. The preliminary results using western blot show that E(z) knock down strain showed decreased levels of H3K27me2 and H3K27me3 while H4 acetylation levels were increased. The most exciting and interesting phenotype of the E(z) knockdown mutants is that the cell shape change from fusiform to oval or oval like shape which suggests a putative role of E(z) in cell differentiation or morphtype transition.

4.2 Results

4.2.1 Putative C5-Methyl transferases in diatoms

Genes encoding putative MTases can be detected in three fully sequenced diatoms *T. pseudonana*, *P. tricornutum*, and *F. Cylindrus*. The highly conserved putative DNMT2 (Tp 22139, Fc139137 andPt 16674) were found in these three diatoms. Other putative DNA MTases in diatoms are not easily classified into the well known DNA MTase classes. Generally the DNA MTases in diatoms have shorter N- terminal regulatory domains. Pt44453 does not contain a C5-MTase catalytic domain but bromo-adjacent homology (BAH) domain and a cysteine rich region (ZF_CXXX) two motifs characteristic of Dnmt1 are present. Pt46156 and Tp9575 belong to the Dnmt3 group but the similarity is not very high. In the three diatom genomes, putative DNMT5 genes were also detected (Pt45072, Pt45071, Fc250902, and Tp3158). DNMT5 is commonly found in Ascomycetes and Basidiomycetes and later on detected in the *Pelagophyte Aureococcusano phagefferens* and the *Prasinophycte* (green algae) *Ostreococcus tauri*, *Ostreococcus lucimarinus,* and *Micromonas pusilla* (Maumus et al., 2011). In contrast to the long N-terminal regulatory domain found in DNMT1 and DNMT3, DNMT5 does not have a long N-terminal domain, but contains instead an SNF2-type DEXDc/HELICc helicase domain at the C-terminus. Interestingly, it seems that putative MTases Pt47357, Tp2094 and Fc148014 with bacterial origin belong to a diatom specific MTase group (**Figure 4.2**).

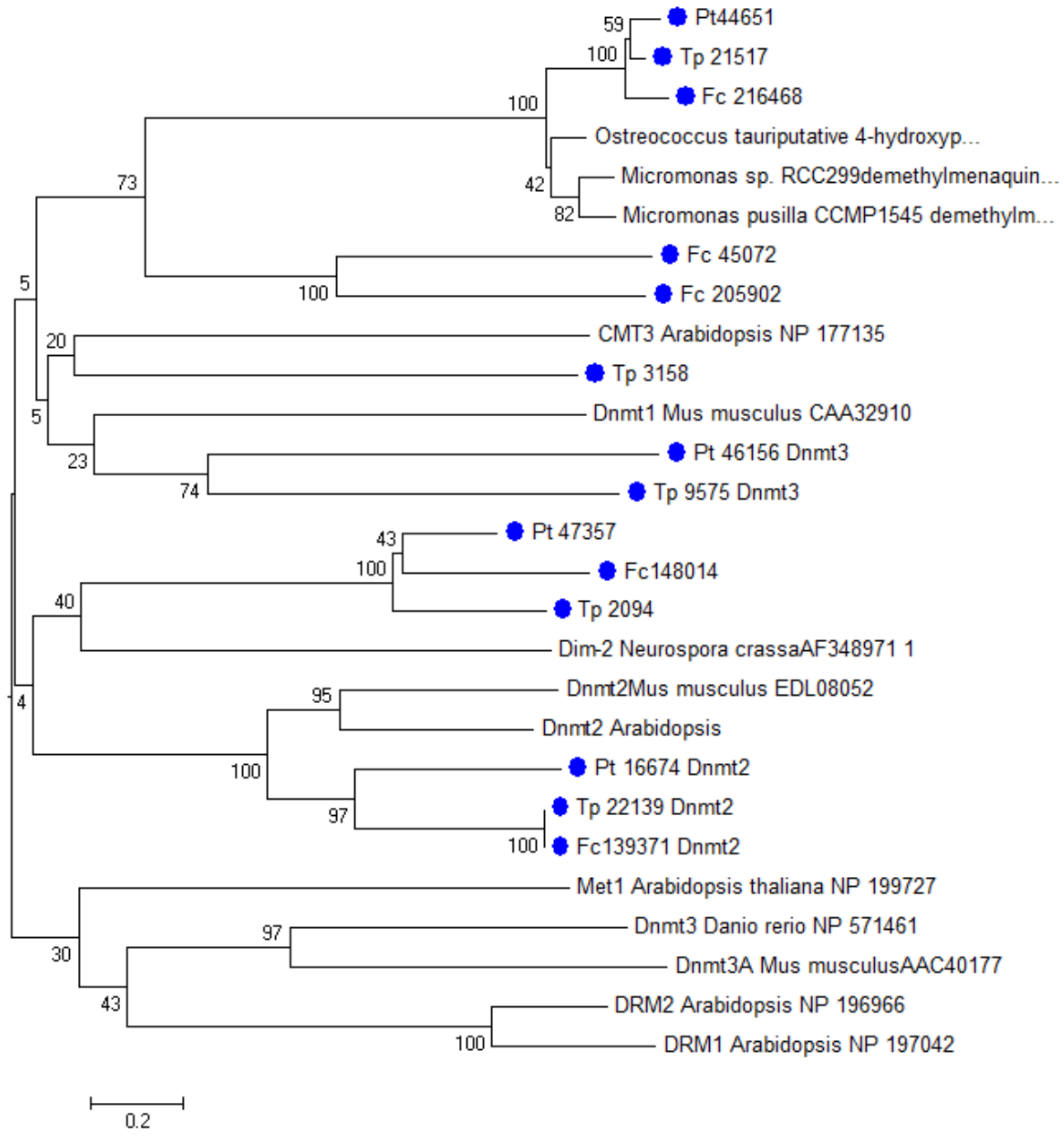


Figure4.2 Neighbor joining phylogenetic tree of C5-MTases from three diatom species and several other model species (bootstrap is 1000 replicates).

In order to begin to investigate the functions of putative DNA MTases in *P. tricornutum*, we therefore carried out a reverse genetic approach by knocking down three putative DNA MTases (46156, 47357 and 45072) in *P. tricornutum*. Antisense constructs were made for each gene and transformed into Pt1 cells (see Materials and Methods section for details). Putative transformed lines were checked by PCR for the presence of the transgene. For each gene, 10 transformed lines were retained for further analysis.

4.2.2 Genome wide DNA methylation quantification of wild type and C5-MTases mutants

To characterize the resulting knockdown C5-MTasesmutants, the relative levels of mRNA of C5-MTases in wild type and C5-MTases were analyzed by performing RT-qPCR. The reductions of mRNA levels in C5-MTases were not observed. However, for some genes of RNAi and antisense knockdown mutants, the reductions can be only observed in protein level not in mRNA level in *P. tricornutum* (De Riso et al., 2009). This is might be due to the absence of silencing at the RNA level and its occurrence only at the post transcriptional level. The best way to characterize the resulting knockdown DNA is analyze the protein levels of C5-MTases in knockdown lines compared to wild type. Currently I do not have the antibodies specific to these C5-MTases, so I used McrBC-qPCR to analyze the DNA methylation level in different locus and I also performed a global quantification of DNA methylation using a kit.

McrBC-Qpcr was performed after digestion of extracted DNA. McrBC can specifically recognize and cut methylated DNA from Pt47357, Pt45072 and Pt 43156 antisense lines. The extent of DNA methylation loss through digestion can be demonstrated through quantitative PCR on McrBC digested and undigested genomic DNA. In this way, McrBC- qPCR r can indicate the status and extent of DNA methylation of specific regions. Six sequences (three genes and three TE regions) were chosen for McrBC- qPCR because they were shown to be methylated in the wild type *P. tricornutum* based on McrBC-ChIP data (Chapter II and http://ptepi.biologie.ens.fr/cgi-bin/gbrowse/Pt_Epigenome/). The Unmethylated H4 gene in wild type *P. tricornutum* was used as the internal control for each digestion and qPCR.In 2-45072-2, all the selected genes and TEs showed decreased DNA methylation level compared to wild type. 47357-2 also showed decreased DNA methylation levels of all the selected loci.

For 47357-5, not all the selected loci showed decreased DNA methylation levels (**Figure 4.3 Figure 4.4**).

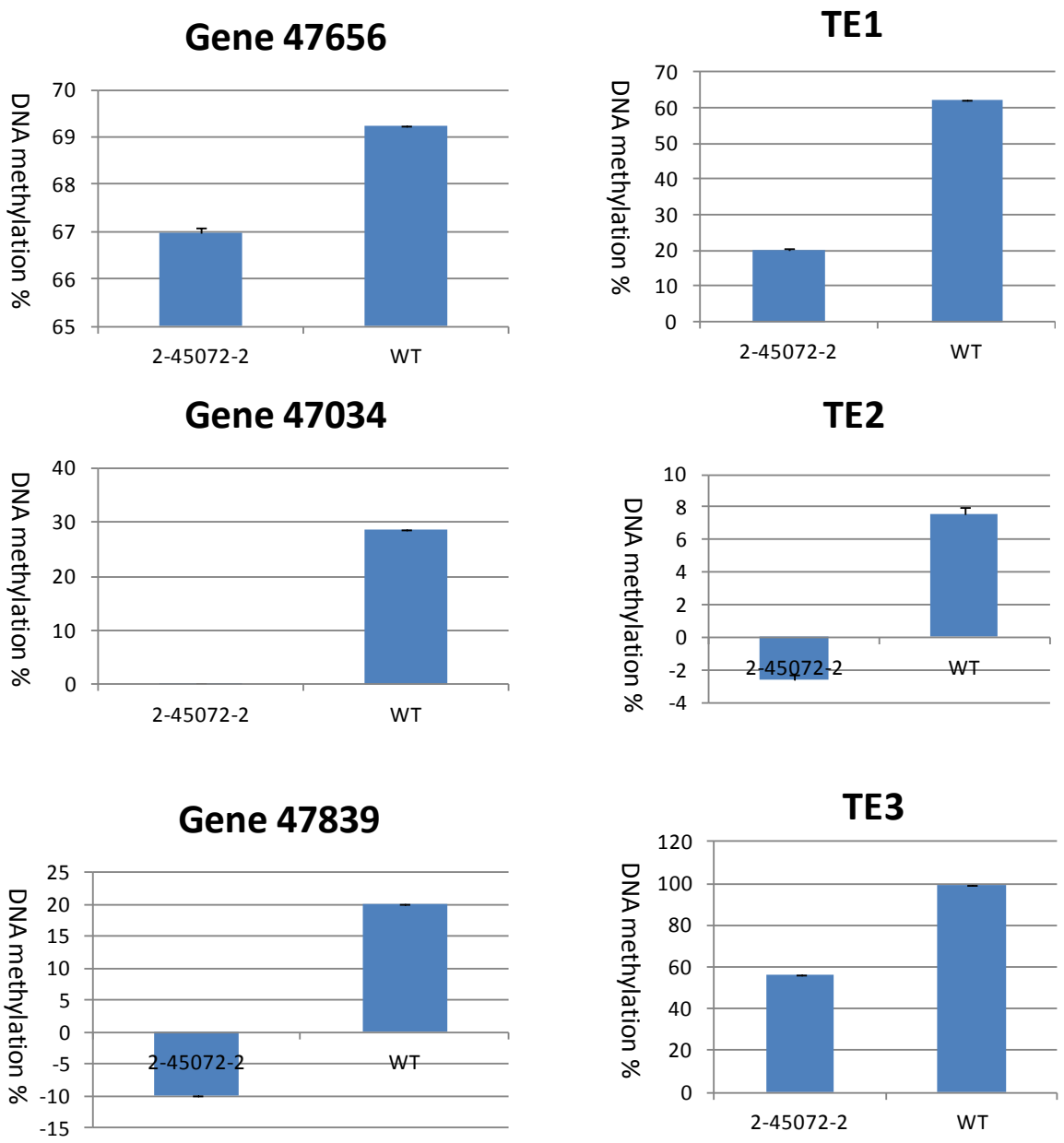


Figure 4.3 McrBC-QPCR analyses of DNA methylationon genes and transposons. 2-45072-2:45072 antisense strain, WT: wild type *P. tricornutum*.

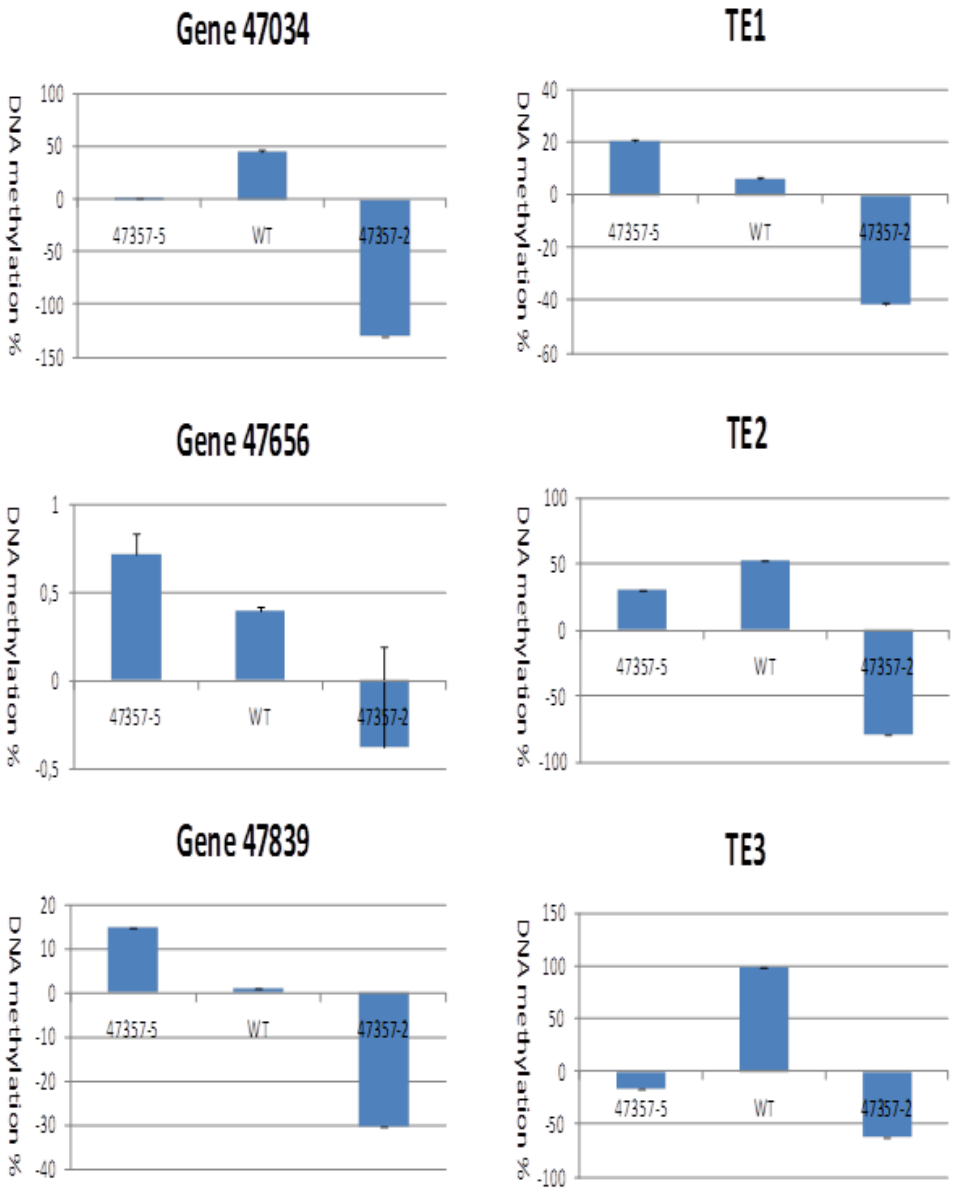


Figure 4.4 McrBC-QPCR analyses of DNA methylationon genes and transposons. 47357-5-2, 47357-5: 47357 antisense strains, WT: wild type *P. tricornutum*.

In summary, the current experimental results are consistent with the hypothesis that putative DNA MTase Pt45072 and Pt47357 seem to have DNA methylation catalytic activity in *P. tricornutum*. However, it is still too early to conclude that putative DNA MTase Pt46156 has no DNA methylation catalytic activity because the experiments are still in progress. Additional transgenic lines of Pt46156 need to be screened and analyzed. Overexpression of the three putative DNA MTases is also in progress.

The further global DNA methylation analysis revealed that DNA methylation levels in 45072 (AS2-45072-2) and 47357 anti sense mutant strains (AS-47357-2 and AS-47357-5) were lower than in the wild type *P. tricornutum*. However the 46156 antisense mutants did not show significant differences in the level of DNA methylation with the wild type (**Figure 4.5**).

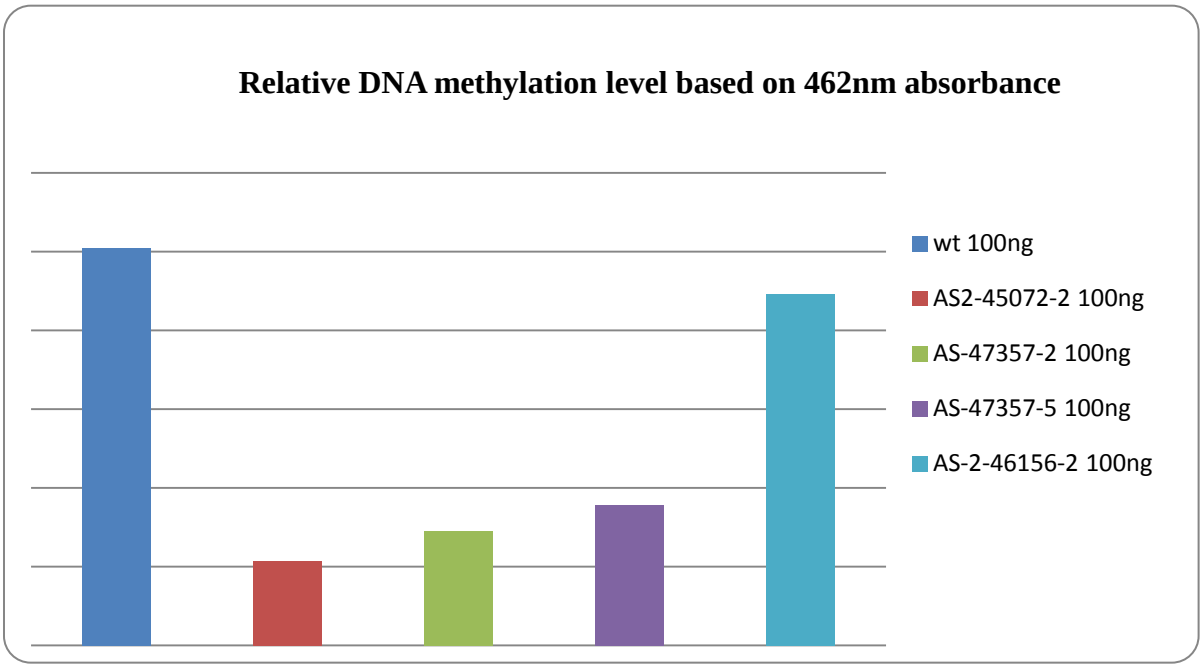


Figure 4.5 Relative DNA methylation level (462nm absorbance) detection by Imprint® Methylated DNA Quantification kit.

4.2.3 DNA demethylation machinery in diatoms

In *P. tricornutum*, genes encoding TET family homologs cannot be found while homologs containing DNA glycosylase domains can be detected. This indicates that the demethylation system in *P. tricornutum* is probably more similar to that of higher plants such as *A. thaliana* (**Figure 4.6**). It is very interesting that the putative demethylases, Pt55285 or Pt53989 and Pt51398 in *P. tricornutum* contain not only a DNA glycosylase domain and "heliex turn heliex base-exersion DNA repair C-terminal" domain, as in *A. thaliana*, but also another domain called "8-oxogunine DNA glycosylase, N terminal" which usually exists in 8-oxoguanine DNA-glycosylases (OGG1). OGG1 repairs oxidatively damaged DNA at sites of 7,8-dihydro-8-oxoguanine (8-oxoguanine) in metazoans and yeast (Mabley et al., 2004; Sandigursky et al., 1997; Sheehan et al., 2005; Youn et al., 2007). The same situation is also found in *F. cylindrus*. In *T. pseudonana*, the situation is different: putative demethylase Tp2836 and Tp31979, like ROS1, DML2 and DML3 *in A. thaliana*, do not contain 8-oxogunine DNA glycosylase while putative demethylase Tp17626 contains the 8-oxogunine DNA glycosylase domains as well as the N terminal domain as in the putative demethylases in *P. tricornutum* and *F. cylindrus* (**Figure 4.7**).

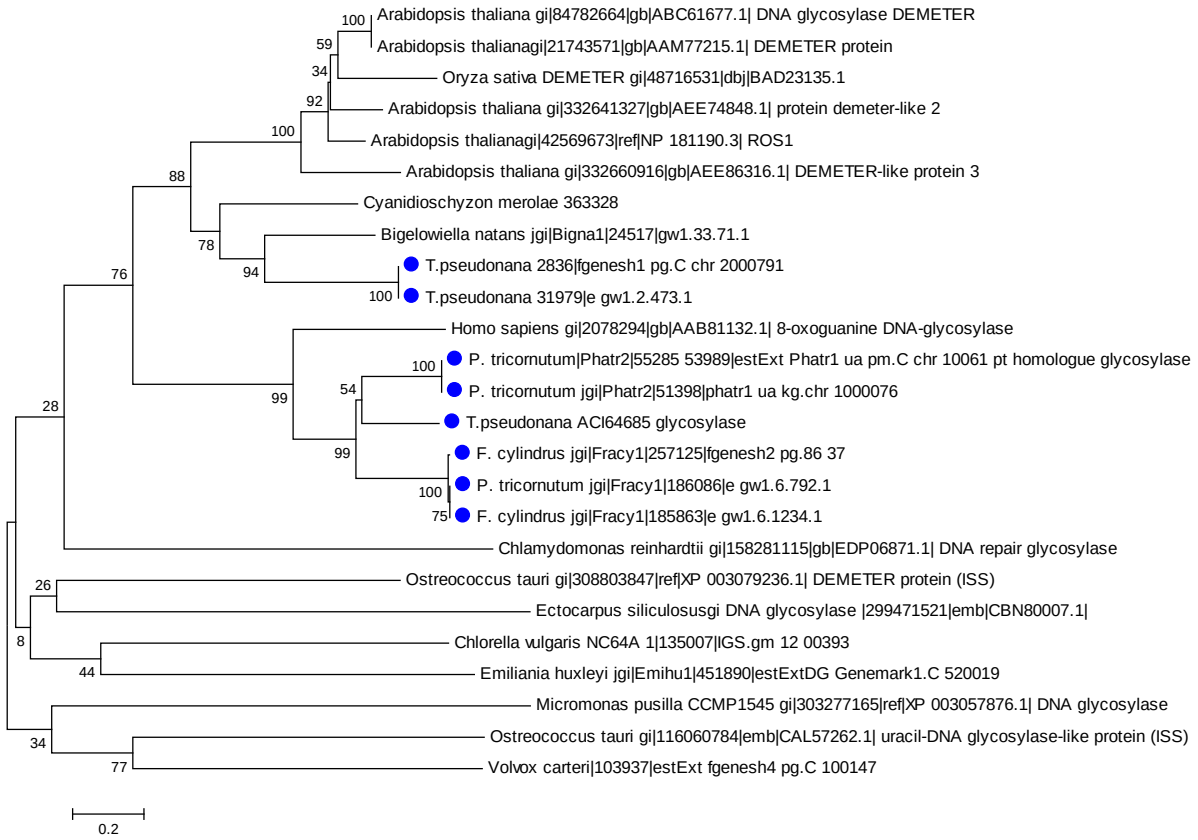


Figure 4.6 Neighbor joining phylogenetic tree of demethylases in plants and algae (bootstrap is 1000 replicates).

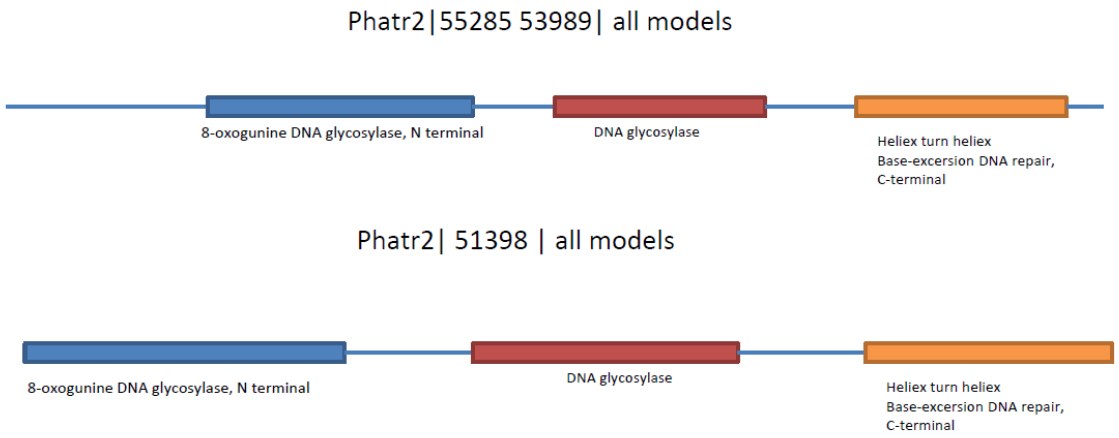


Figure 4.7 Schematic representation of putative demethylases in *P. tricornutum*.

4.2.4 Identification of proteins containing SET domains for histone methylation in *P. tricornutum*

SET domain-containing proteins are responsible for histone methylation. A wide range of putative SET domain-containing proteins can be identified in *P. tricornutum* (**Figure 4.8**). However, in most cases, it is impossible to analyze specific putative functions because of their sequence divergence with respect to known SET domain containing proteins involved in histon methylation. In the future, it will be worth to manipulate by forward and reverse genetic some of these putative genes to explore their functions and cross-talks between different putative genes responsible for different histone modifications.

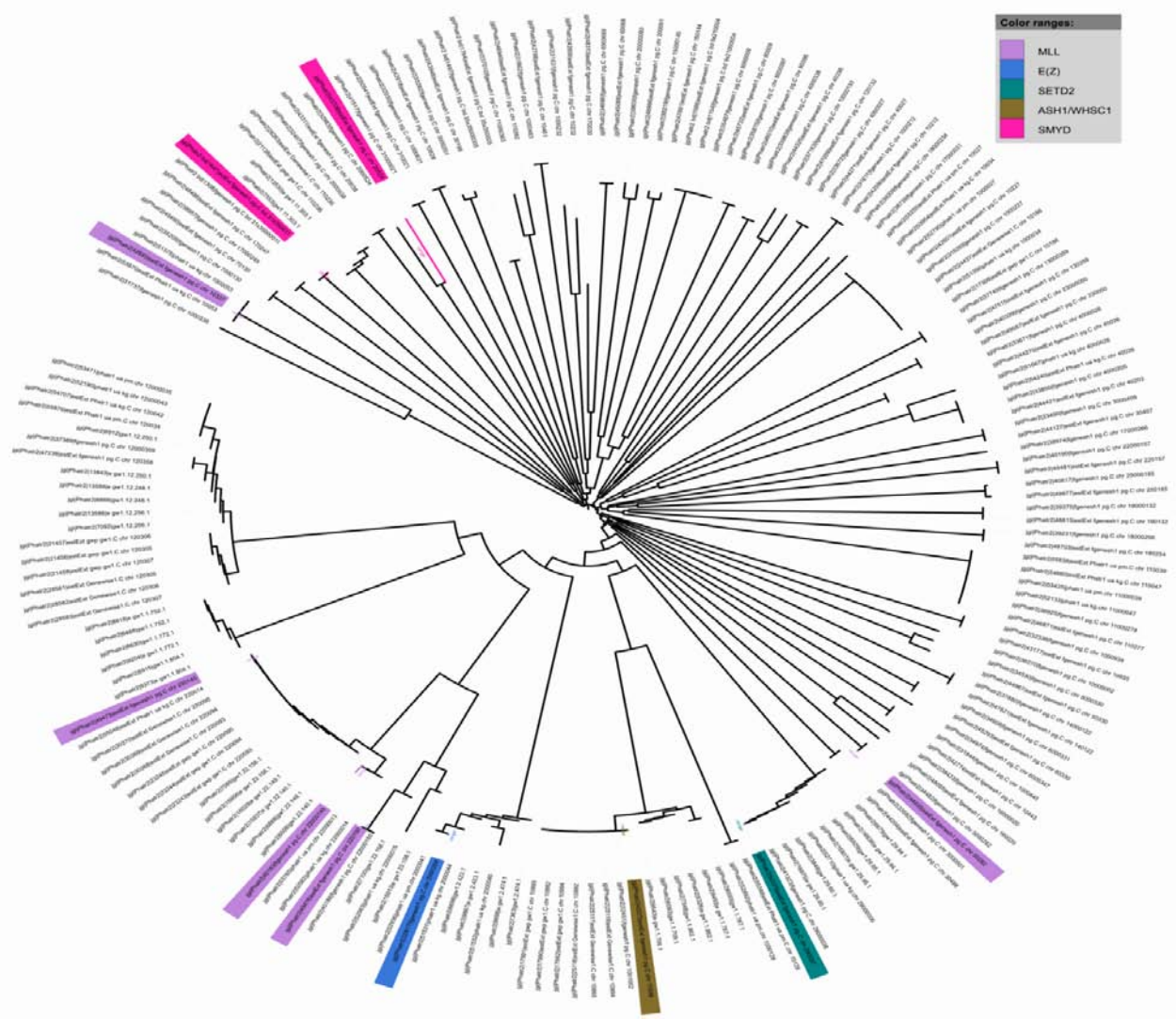


Figure 4.8 SET domain containing proteins in *P. tricornutum*.

MLL: Lysine Methyltransferase on H3K4

E(z): Lysine Methyltransferases on H3K27

SETD2: Lysine Methyltransferases on H3K36

ASH1/WH5C1: Lysine Methyltransferases on H3K4

SMYD: Lysine Methyltransferases on H3K4

4.2.5 Distribution of PRC2 core subunits in diatoms and other eukaryotic algae

Genome comparison analyses have revealed that not only multicellular organisms but also unicellular organisms contain the homologs of PRC2 core subunits (Shaver et al., 2010). A sequence similarity-based method between domains/motifs and full-length sequences of gene models was used to identify PcG homologs. Here, I explored homologs of the PRC2 components in eukaryotic organisms with focus on unicellular algae. The amino acid sequences of *D. melanogaster* were used to for BLAST analysis in different genomes. The putative proteins with 50 amino acid similarity were considered as homologs. In order to evaluate the taxonomic distribution of the PRC2 core components, E(z), ESC, Su(z)12 and Nurf55, I surveyed the genomes from unicellular algae together with some model species. Nurf55 is widely distributed in all the organisms I surveyed (**Figure 4.9**). This is not surprising because this protein is shared not only by PcG complex, but also by several other distinct machineries such as HAT1 (histone acetyltransferase1) and histone deacetylase complexessuch as CAF-1 (chromatin assembly factor-1), and ATP-dependent nucleosome-remodeling complexes NURF and NuRD (Nowak et al., 2011; Suganuma, Pattenden, & Workman, 2008). So in order to avoid confusing interpretations, only the taxonomic distribution of E(z), ESC and Su(z) was considered when making inferences about the origin of PRC2 complexes.

Shaver et al demonstrated that PRC2 appeared early in eukaryotic evolution and that the advent of the PRC2 complex is unlikely to have been associated with the emergence of multicellularity (Shaver et al., 2010). Here we examined the eukaryotic unicellular algae from Stramenopiles, Chlorophyta (green algae), Cryptomonad, Chromalveolata and Rhodophyta (**Figure 4.9**). The results demonstrate that the eukaryotic unicellular algae *Bigelowiella natans* from Cryptomonad, *Guillardia theta* from Chromalveolata, the unicellular red alga*Cyanidioschywon merolae*, two diatom species *P. tricornutum* and *F. cylindrus*, two picoplankton Prasinophyceae *Ostreococcus tauri*, and *Bathycoccus* also contain PRC2 components. It therefore seems evident that PCR2 components were lost in Excata, Amoebozoa and budding yeast *S. cerevisiae* and *S. pombe* but retained in other lineages of life. Our results further confirmed the irrelevance between mutlicellularity and PRC2 components.

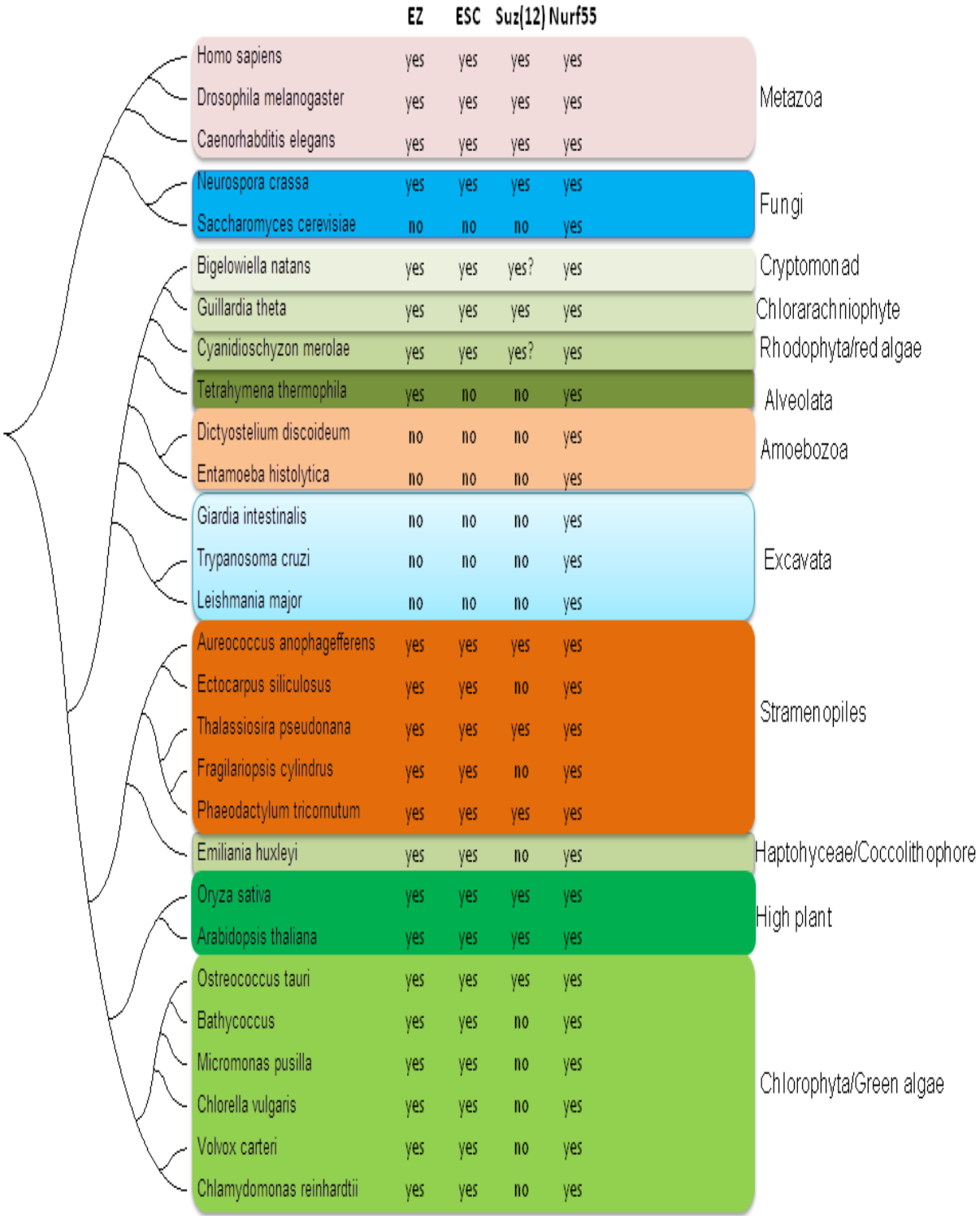


Figure 4.9 Distribution of PRC2 core subunits among eukaryotic organisms. The phylogenetic tree was constructed based on NCBI taxonomy data.

4.2.6 Phylogentic analyses of PRC2 component E(z)

E(z) is the key component of PRC2 complex because it is the catalytic unit for adding the methyl group to H3K27me1 and H3K27me2 for H3K27me2 and H3K27me3 formation. The polypeptides corresponding to E(z) proteins are quite conserved in species where H3K27 methylation has been experimentally demonstrated in metazoans and higher plants. Mammalian organisms have two E(z) homologs: Ezh1 and Ezh2 whereas E(z) in plants (*A. thaliana*) has expanded into three homologs. In mammals Ezh1 and Ezh2 maintain repressive chromatin through different mechanisms (Raphael Margueron et al., 2008). Single E(z) homologs were detected in the unicellular algae with fully sequenced genomes. This includes the following species: green algae (*Ostreococcus tauri, Bathycoccus*, *Micromonas*, *Chlorella vulgaris*, *Volvox carteri*, *Chlamydomonas reinhardtii*), red algae (Cyanidioschyzon merolae), brown algae (*Ecocarpus siliculosus*), diatoms (*Phaeodactylum tricornutum*, *Thalassiosira pseudonana*, *Fragilariopsis cylindrus*), haptophyceae (*Emiliana huxleyi*), Cryptomonad (*Bigelowiella natans*), and Cryptophyta (*Guillardia theta*). The wide distribution of E(z) in unicellular algae suggests important potential roles of H3K27me3 in these species. The putative E(z) proteins detected in different organisms are too diverse so the phylogenetic tree cannot be well resolved. Therefore a phylogenetic tree was constructed based only on the SET domains **(Figure 4.10)**. However, this phylogenetic tree also did not allow me toresolve the relationship among most of these proteins unequivocally. This implies that E(z) has a very complicated evolutionary history. However, the usual problems of weakness of phylogenetic signal, later gene transfers, hidden paralogy, tree reconstruction artifacts and incorrect predicted genes should also be considered. Notwithstanding, comparative genome analyses have revealed that putative E(z) homologs exist in the diatoms *P. tricornutum* (Pt32817), *T. pseudonana* (Tp268872) and *F. cydinrus* (Fc181541). To our surprise, the putative E(z) SET domain in the centric diatom *T. pseudonana* did not form a single clade with the pennate diatoms *P. tricornutum* and *F. cydinrus* which further complicate interpretation of evolutionary history

4.2.7 Phylogentic analyses of PRC2 components ESC

ESC is another key component of the Polycomb repressive complex 2 (PRC2).A phylogenetic tree of ESC-like proteins, constructed by aligning the full length sequences, suggests a

monophyletic origin for the polypetides found in all the organisms containing putative E(z) homologs expect the unicellular organism *Tetrahymena thermophila*. The existence of different forms of H3K27 methylation (H3K27me1, H3K27me2 and H3K27me3) has been demonstrated by mass spectrometry in *T. thermophila* (Garcia, Shabanowitz, et al., 2007). The absence of ESC in *T. thermophila* and the existence of different forms of H3K27 methylation imply therefore that ESC is not correlated with methylation deposition on H3K27. All the unicellular algae examined here contain ESC components. The functions of ESC in unicellular algae have not been explored yet. Compared to the SET domains of E(z), aphylogenetic tree of ESC sequence treefor unicellular algae has better phylogentic relationship resolution: three diatom species are clustered together and three prasinophytes are also clustered together (Figure 4.11). However, the putative ESC in *B. natans*, *G. theta* and *C. merolae* do not show close relationships as they do in the taxonomic phylogenetic tree.

4.2.8 Phylogentic analyses of PRC2 components Suz(12)

Su(z)12 proteins are not widely distributed like E(z) and ESC. They have been lost in most green algae species, one diatom species (*F. cylindrus*), the brown algae *E. siliculosus*, and the haptophyte *E. huxleyi* among the unicellular algae species we have examined here. In green algae/Chlorophyta group, except *Ostreococcus tauri* which has Suz(12), *Bathycoccus*, *Micromonas pusilla*, *Chlorella vulgaris*, *Volvox carteri* and *Chamydomonas* do not contain Suz(12) which implies that the PRC2 component Suz(12) was lost in most green algae species during evolution. In another important marine algae *E. huxleyi*, Suz(12) is absent. Among the three diatom species *P. tricornutum*, *T. pseudonana* and *F. cylindrus*, *P. tricornutum* and *T. pseudonana* contain Su(12) while *F. cylindrus* seems to have lost it. It is interesting to note that the brown alga *E. siliculosus* closely relative group of diatoms also lacks Suz(12). As we can see from the Suz(12) phylogentic tree, the metazoan sequences are distinct from others, but within the algae and plant the relationship cannot be resolved very clearly (**Figure 4.12**).

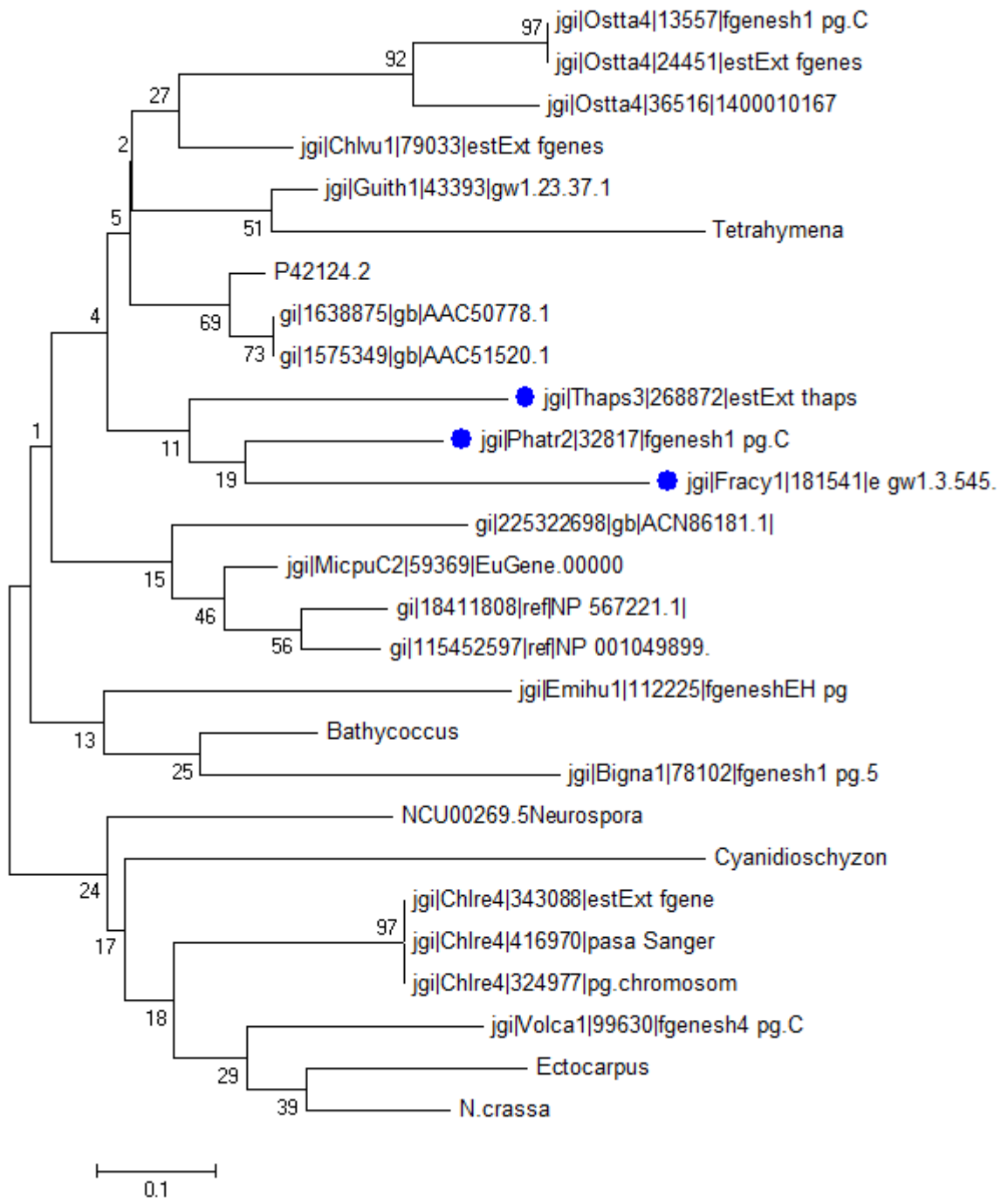


Figure 4.10 Phylogenetic NJ tree of SET domains of EZ in different organisms (bootstrap is 1000 replicates).

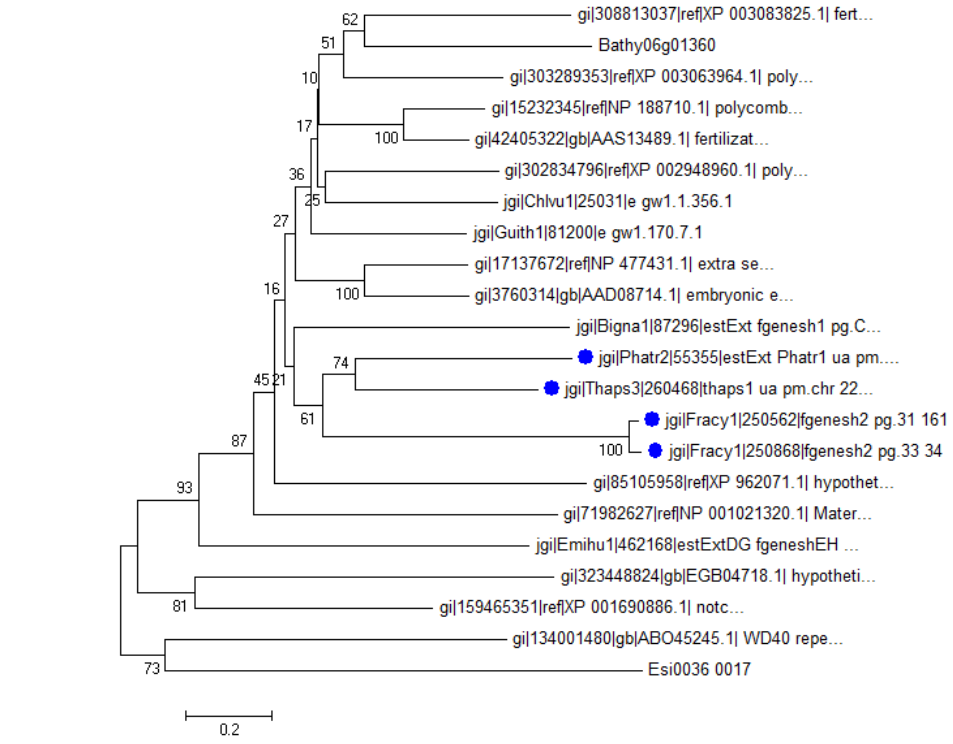


Figure 4.11 The phylogenetic NJ tree of putative ESC in different organisms (bootstrap is 1000 replicates).

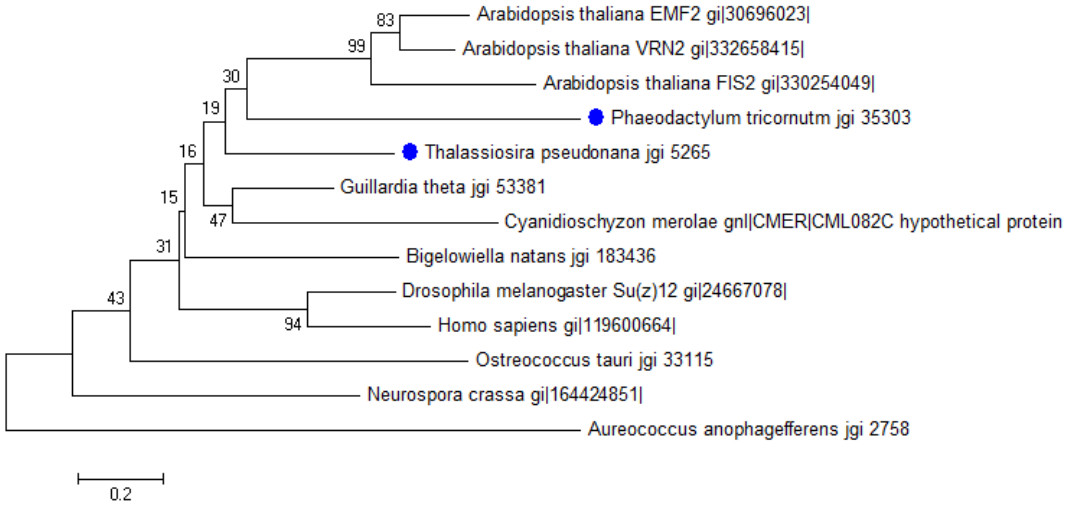


Figure 4.12 The phylogenetic NJ tree of putative Suz(12) in different organisms (bootstrap is 1000 replicates).

4.2.9 Distribution of PRC1 core subunits in diatoms and other eukaryotic algae

PRC1 components are not very conserved in plants and metazoans. RING1A/B together with BMI1, MEL18 (PCGF2) or NSPC1(PCGF1) or Pc (in Drosophila) are considered as the principal PRC1 components in metazoans (Whitcomb et al., 2007) where LHP1 and AtRING1a are considered as the major PRC1 components in plants. Among them, RING1A and AtRING1a are homologs.

Firstly, I examined the presence of putative RING1A or AtRING1a in eukaryotic unicellular algae anda constructed phylogenetic tree (**Figure 4.13**). The RING1A or AtRING1a homologs were not widely distributed in eukaryotic unicellular algae we have examined. It is missing in the green algae *O. tauri* and Bathycoccus, brown algal *E. siliculosus*, brown tide algal *A. anophagefferens*, the diatom *F. cylindrus* and the red alga *C. merolae*. The homologs of animal PRC1 components Pc BMI1 protein are not detected in the eukaryotic unicellular algae we have examined here. As for the homomlogs of LHP1-like protein, which contains a chromo domain, only one homolog was detected in *E. huxleyi* wherase putative chimeric proteins containing not only a chromo domain but also a Jmjc domain were found in *P. tricornutum* and *T. peudonana* (Pt 34913, Tp1863) (**Figure 4.14**). It is very interesting that the Jmjc domain which is responsible for histone demethylation while chromodomains are commonly found in proteins associated with remodeling and manipulation of chromatin (Flanagan et al., 2005; Klose, Kallin, & Zhang, 2006; Xhemalce & Kouzarides, 2010). In *A. thaliana*, the chromodomain-containing protein LHP1 colocalizes with H3K27me3 genome-wide. The LHP1 chromodomain also binds H3K27me3 with high affinity, suggesting that LHP1 has functions like the Polycomb protein (Exner et al., 2009). In *Drosophila*, the chromodomain containing protein heterochromatin protein 1 (HP1) is associated with H3K9me2 modification for heterochromatin formation (Elgin & Grewal, 2003; LeRoy et al., 2009). In diatom genomes, the chimera proteins which have putative dual functions are very common (Bowler et al., 2008). However, the actual function of this chimeric diatom protein containing Jmjc and chromo domains is yet to be determined.

The homologs of *Drosophila* Pcl and *Arabidopsis* VEL1 protein could not be detected in any of the eukaryotic algae examined here. Since *D. melanogaster* Pcl and *A. thaliana* VEL1 proteins are analogs not homologs, it is possible that these algae have another protein with

distinct amino acid sequences which has similar functions as Pcl and VEL1 for H3K27me3 deposition.

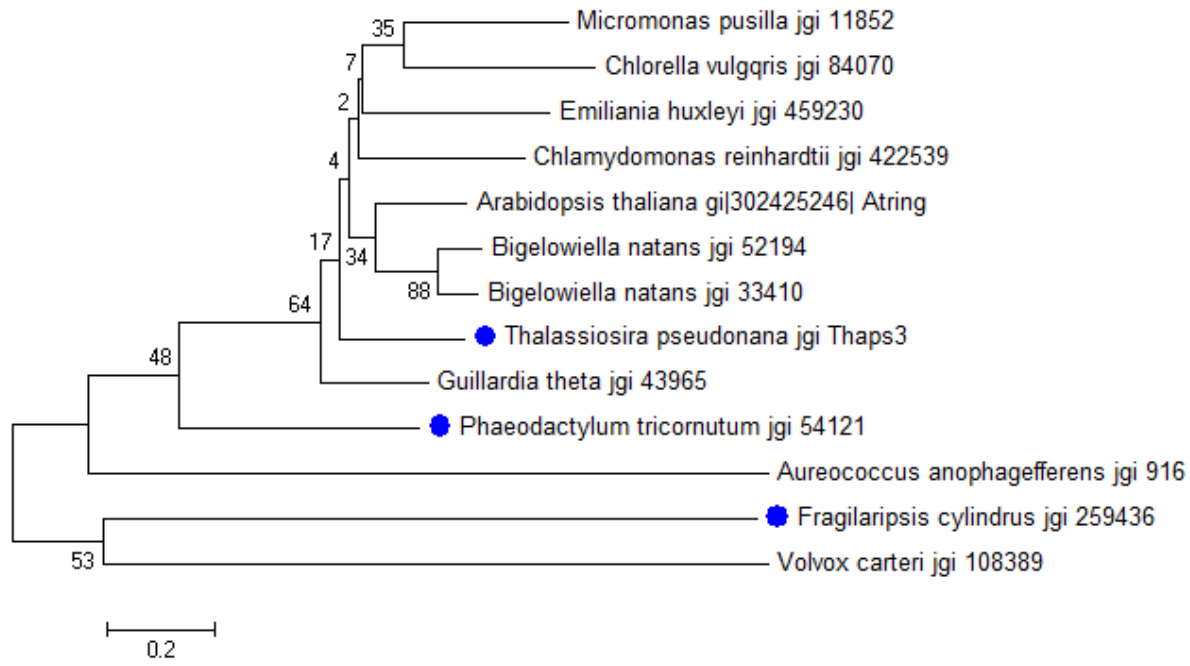


Figure 4.13 Phylogenetic NJ tree of AtRING1A in different organisms (bootstrap is 1000 replicates).

Figure 4.14 Schematic representation of Pt34913 LHP homolog.

4.2.10 Knockdown of E(z) homolog 32817 in*P. tricornutum*

According to RNA-seq data (data not shown), the putative E(z) gene Pt32817 gene expressed in normal *P. tricornutum* cultures of the Pt1.8.6 which is composed of 90% fusiform cells *P. tricornutum* has four morphytypes: fusiform, triradiate, oval and roundish. Under normal lab

conditions, most cells are fusiform while others are triadiate, oval and round cells. Under stresses, the cells tend to be oval and roundish. I also utilized quantitative reverse transcription (RT)-PCR to detect the expression level of EZ homolog 32817 wild type and mutants in *P. tricornutum* under normal conditions. I did not observe a decrease in the expression level of 32817 mRNA in mutant lines (data not shown, the mutant lines were shown in **Figure 4.15**). This is might be due to the absence of silencing at the RNA level and its occurrence only at the post transcriptional level. To further investigate the silencing of the mutants, we need to perform a western blot using an antibody against the protein itself, but such an antibody is not yet available.

4.2.11 Effect of EZH down regulation on global post-translational histones modifications

In order to gain insight on the chromatin state alterations caused by down regulation of E(z), we carried out western boltting using antibodies specific to certain histone modification. As we know, E(z) mediates H3K27me2 and H3K27me3 in plants and animals, so firstly we examined whether down regulation of Ez decreases global H3K27me3 and H3K27me2. I observed that the levels of H3K27me3 and H3K27me2 were indeed decreased in Pt 32817 knockdown mutants compared to wild type. In contrast global H4 acethyation level was increased in As32817 (**Figure 4.15**).

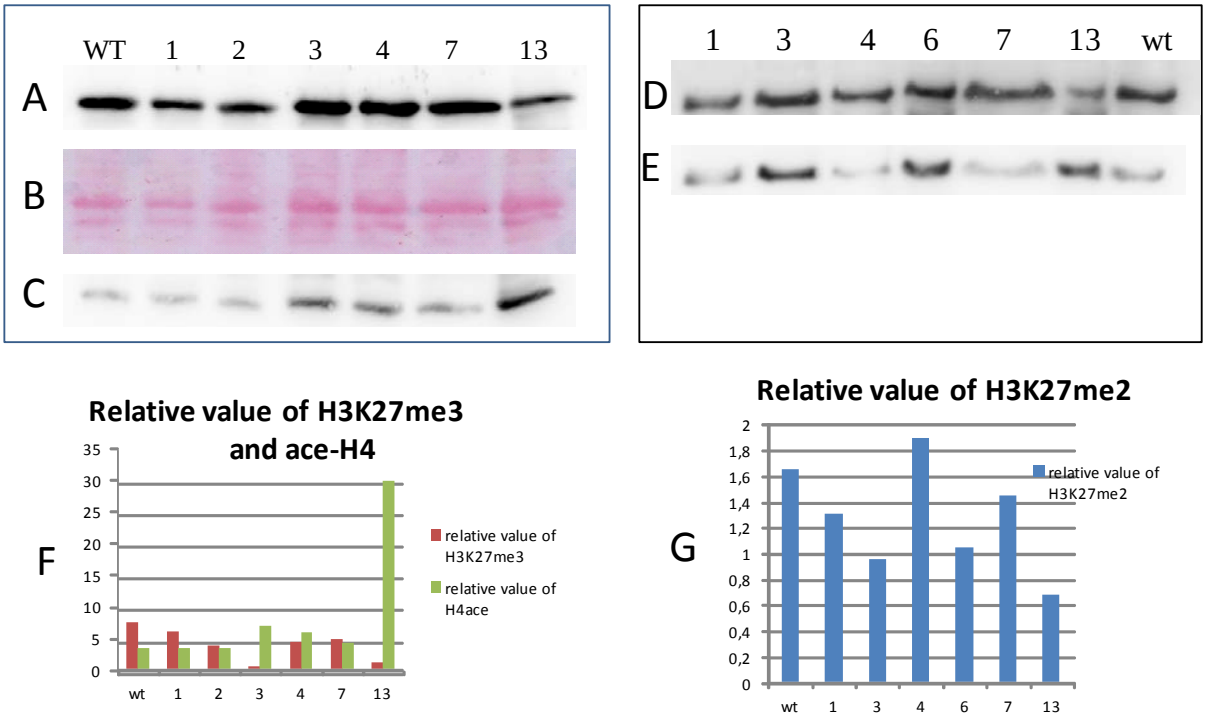


Figure 4.15 Western blotting detection of H3K27me3, H3K27me2 and ace H4 in wild type (wt) and antisense lines1, 2, 3, 4, 6, 7 and 13. A: western blotting detection of results H3K27me3, B: red ponceau loading control for H3K27me3 and acethyl-H4 which recognizes acetylated histone H4, C: western blotting detection results of acethyl-H4, D: western blotting detection results H3K27me2, D: western blotting detection results of H4 (loading control for H3K27me2). In F and G the relative value is calculated by Imagine J.

4.2.12 Light microscopy observations of Ez antisencse strains

Interestingly, I found that the cell shape of *P. tricornutum* AS3217 strains was different from wild type *P. tricornutum* cells. The wild type *P.tricornutum* Pt1 grown in the lab usually sustains the fusiform morphotype and the cells used for transformation are also fusiform. However, the E(z) mutant strains showed a tendency to be oval instead of fusiform In particular, the freshly inoculated cells become roundish and even form aggregates (**Figure 4.16**).

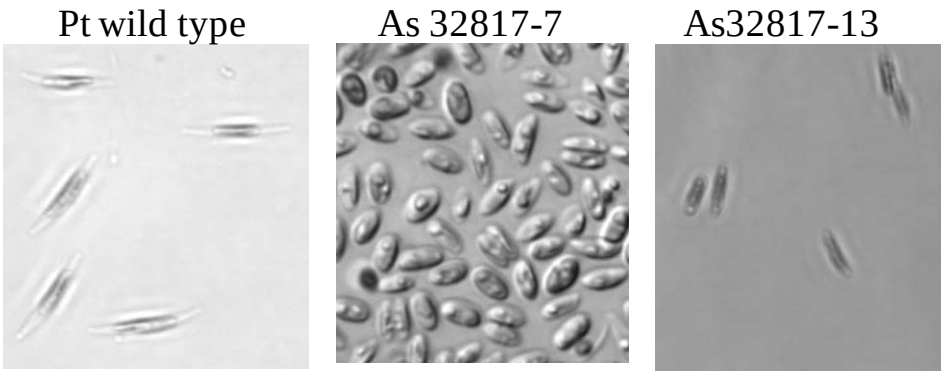


Figure 4.16 Light microscopy pictures of wild type *P. tricornutum* and As32817 mutants.

4.3 Discussion

4.3.1 C5-MTase knockdown in *P. tricornutum*

These preliminary results of C5-MTase knockdown in *P. tricornutum* are encouraging but need further analysis before envisaging sophisticated studies genome wide. It will firt be necessary to design antibodies specific for proteins 46156, 45072 and 47357 for western blotting experiments to confirm decreased putative MTase protein levels in the mutants. Secondly, as we know, budding yeast does not have DNA methylation which makes it a good model for studying DNA methylation *in vitro*. Therefore, overexpression of the three putative DNA MTases in budding yeasts may help us figure out which putative MTase has DNA methylation catalytic activity. A next step will be genome wide profiling of DNA methylation in selected lines by bisulfite sequencing. This approach will help figure out whether any of the different contexts of methylation in *P. tricornutum* C5-MTase. Furthermore RNA-seq analysis of the mutants will help understand how DNA methylation affects gene expression. Small RNA profiling combined with other epigenetic key components such as histone modifications will ultimately draw a clearer picture on the mechanisms underlying DNA methylation in diatoms as this is likely not to be an isolated phenomenon but acts under tight regulation and is correlated to dynamic chromatin states.

4.3.2 Distribution of PcG components in algae and E(z) knockdown in *P. tricornutum*

Previous research showed evidence for the widespread existence of PRC2 components in unicellular eukaryotes from the Opisthokonta, Archaeplastida and Chromalveolata supergroups, which contradicts the idea of their presence only in multicellular organisms (Shaver et al., 2010). Thus, it was proposed that the PRC2 components likely evolved in the last unicellular common ancestor of all eukaryotes and were subsequently lost from certain lineages. H3K27me2 and H3K27me3, the outcome of E(z) enzymatic activity, have been extensively studied in multicellular organisms such as *Arabidopsis* and human. However, the study on PcG protein and H3K27methylation in unicellular organsims is very limited with the exception of *Chlamydomonas* and *Tetrahymena thermophila.* The existence of H3K27 methylation in *Chlamydomonas* and *Tetrahymena thermophila* was identified by mass spectrometry analyses (Garcia, Hake, et al., 2007). Moreover, investigation of PcG proteins was carried out by knocking down the putative E(z) in *Tetrahymena thermophila* and

Chlamydomonas. In *T. thermophila*, H3K27me3 was found associated with heterochromatin formation and DNA elimination while in *Chlamydomonas* it has been revealed that E(z) is involved in transcriptional repression of repetitive transposons and retrotransposons (Y. Liu et al., 2007; Shaver et al., 2010). Therefore, these studies confirmed that PRC2 indeed provide a function in unicellular organisms. However, it seems that the functions of E(z) in unicellular organisms are different from what have been found in multicelluar organisms. Thus, it is very tempting to explore PcG protein functions in unicellular organisms. Here I focused on the distribution of PcG proteins in unicellular algae. Thanks to the whole genome sequencing of unicellular algae from different lineages, here I explored the putative PcG proteins in these species. The results show that PRC2 proteins are widely distributed in unicellular algae. Especially E(z), which has H3K27me3 methyltransferase enzymatic activity, appears in all the unicellular algae examined here.

I further explored the distribution of PRC1 proteins in these unicellular algae. The AtRING1A or RING homologs were detected in some species. It seems alga PRC1 components are more plant like. However, LHP homologs were detected in a few species. Because the components of PRC1 complexes in metazoans and plants are not conserved, it is possible that unicellular algae from different lineages have developed distinct system responsible for the job of PRC1. *T. thermophila* contains only E(z) but not ESC. Different statuses of H3K27 methylation were detected in *T. thermophila* (Garcia, Hake, et al., 2007) implying that E(z) is the core component while other components of PcG proteins probably are not so indispensible in unicellular organisms for H3K27 methylation deposition.

It is not surprising that *V. carteri*, the primitive multicellular green algae, contains all the PRC2 components. However, its relatives, the unicellular green algae *C. vulgaris* and *C. reinhardtii* also have PRC2 components which indicate that the appearance of PRC2 is not related to multicelluarity.

Stramenopiles and Prasinophytes are two out of four major eukaryotic phytoplankton lineages (prymnesiophytes, alveolates, stra-menopiles, and prasinophytes) which are responsible for close to half of marine photosynthesis (Cuvelier et al., 2010). Until now only Stramenopiles (*P. tricornutum, T. pseudonana, F. cylindrus, A. anophagefferenens, E. siliculosus*) and Prasinophytes (*O. tauri*, *Bathycoccus*, *M. pusilla*) have been sequenced. The wide distribution

of PRC2 components in Stramenopiles and Prasinophytes emphasize the potential roles of PRC2 in the context of the marine environment. Species from Parinophytes are very tiny and called pico-phytoplankton (0.2 and 2 µm). *O. tauri* is considered to be the smallest free living eukaryote (Claude et al, 1994). In such small cells, the genomes are also very small (about 12Mb). This tiny pico-phytoplankton is an important group of eukaryotic phytoplankton which contributes to the marine photosynthesis and carbon cycling. The PRC2 proteins were also detected in all three Prasinophytes.

Diatoms are the most diverse and abundant eukaryotic organisms in diverse oceanic environments and also play critical role in carbon cycling. PRC2 proteins E(z), ESC and Su(12) were detected in two diatom species *P. tricornutum* and *T. pseudonana*. It is known that H3K27me3 and E(z) are involved in cell identity. *P. tricornutum* has four morphotypes: fusiform, triradiate, oval and roundcells. The mechanism of cell morphotype transition is not clear. It is tempting to speculate that E(z) and H3K27me3 might be involved in this process. We found that the genome wide distribution pattern of H3K27me3 in *P. tricornutum* is unorthodox: H3K27me3 mainly marks TEs not genes which is distinct from the organisms where H3K27me3 distribution has been profiled so far (see Chapter III). This has raised our interests in the functions of E(z) and H3K27me3 in *P. tricornutum* and their putative importance for living in a fluctuating environment such as the ocean. Forward and reverse genetic manipulation systems have been well established in the Prasinophytes representative species *O. tauri* and in the diatom *P. tricornutum.* It is therefore very interesting to see the outcome of functional studies of PcG genes in these species.

The cells of E(z) 32817 *P. tricornutum* knockdown mutants tend to be roundish and oval in contrast to what is found in cultures of which are different from wild type cells. It is therefore possible that E(z) in *P. tricornutum* might be involved in the morphotype fate based on here preliminary experimental results. In order to confirm this hypothesis, several experiments need to be carried out: 1. western boltting experiment on mutants and wild type *P. tricornutum* using a antibody specific to 32817; 2. Global profiling of H3K27me3 suchmutants together with RNA seq to uncover the pathway of morphotype transition in *P. tricornutum* and to identify which genes are regulated by polycomb complex in diatoms. These mutants need also to be characterized for other histone marks and DNA methylation because Polycomb components act in specific chromatin contexts which need to be defined and studied to dissect

the existing cross-talk between different chromatin states. Furthermore, reverse genetic analysis of the remaining *P. tricornutum* PRC2 and PRC1 components is envisaged as these different components interact together for establishing silencing, differentiation and cell fate, as shown by previous work.

4.4 Materials and methods

4.4.1 Cell culture

P. tricornutum Bohlin (CCMP632) was obtained from the Provasoli-Guillard National Center for Culture of Marine Phytoplankton. Cells were grown in f/2 medium at 20 degrees under white fluorescent lights (70 $mmolm^{-1}$ s^{-1}), 12 h: 12 h dark–light cycle. Analyses of the wild-type and knockdown mutants have been performed on cells in exponential phase of growth and collected simultaneously.

4.4.2 Antisense vector construction

The original FcpBp GUS –antisense vector (De Riso et al.) was digested by Xbal and EcoRI enzyme. The original FcpBp GUS –antisense vector contains *Shble* gene which is for zeocin or phleomycin selection.The digestion was done as follow: 2µl 10 X tango Buffer, 2ulXbal, 1µl EcoRI, Vector 5µl (50ng/µl) and 10µl H_2O to reach a volume of 20µl in total. The mixture was incubated at 37 degrees for 2 hours. PCR fragments were amplified from gene Pt47357, Pt46156, Pt45072. The primers are:

Pt45072antisenseFW: AACGGAATTCGTACAACGGTT

Pt45072antisenseREV: CTGGCACGTCTCTAGATCTCC

Pt46156antisense FW: GAATTCGGACACCTTATTCGT

Pt46156antisense REV: GTTTCTCGTTGCCAGCCC

Pt47357antisenseFW: TACTAGAATTCTGCGATA

Pt47357antisenseREV: TCCATCCTGTCTAGAAGT

32817 EcoRFw1: TTACTGAATTCCAAAAGATCCTTTG

32817 Xbalrev1: CAGCATCTAGAACGCTGGGTGGGGAT

32817 EcoRFw: AGCGGGAATTCGATCCGGACCTTTG

32817 Xbalrev: GTTCAATCTAGATTCGGAGACAGTTAT

The fragment length is around 250bp. The forward primer contains an *EcoRI* site and the reverse primer contains a *XbaI* site). The PCR bands were cut and purified by GenElute PCR clean-up Kit. In order to get the fragments in the antisense orientation, the PCR fragments were then digested with EcoRI and XbaI, the digestion was done as follows: 15µl purified PCR fragment, 2µl 10X tango buffer, 1µlEcoRI, 2µlXbaI in total 20µl. The mixture was incubated at 37 degrees for 2 hours. In order to introduce the fragment of interest into the vector containing Shble, a ligation reaction was done as follow: 7µl PCR fragment digested by EcoRI and XbaI, 1µl vector digested by EcoRI and XbaI (25ng) and 1µl T4 ligase in total of 10µl volume. The mixture was incubated at 14 °C overnight.

4.4.3 Genetic transformation of *P. tricornutum*

The vectors containing antisense fragments of 47357, 46156, 45072 three putative C5-MTase were introduced into wild type *P. tricornutum*strain by microparticle bombardment using a Biolistic PDS-1000/He Particle Delivery System (Bio-Rad). Overexpressed vector constructs were cotransformed with the *pFCPFp-Shble*vector, as previously described (Falciatore, Casotti, Leblanc, Abrescia, & Bowler, 1999).Transformed colonies were first selected on 1% agar plates (50% f/2 medium) containing 50 µg/mL phleomycin (Invitrogen). For the antisense transformants, the presence of the fragment of interest was verified by both performing PCR using the *Sh ble1fw* and reverse primers and sequencing.

4.4.4 DNA methylation quantification

DNA was extracted from wild type *P. tricornutum* and mutant strains in exponential phase of growth and collected simultaneously. Imprint® Methylated DNA Quantification (Sigma) was used to detect DNA methylation. 2µl of 50ng/µl of DNA from wildtype and mutant (two replicates for each sample) were added into 28µl dilution buffer provided by the kit. The following procedure was performed according to the manusfracture's instruction. For each sample, the absorbance value represents the relative value of DNA methylation. The final

DNA methylation quantification analysis was based on the absorbance value of samples. Up to 200 ng of purified DNA is bound to the wells of the assay strip. The methylated DNA is detected using the Capture and Detection antibodies, then quantified colorimetrically. The amount of methylated DNA present in the sample is proportional to the absorbance measured. A standard curve can be performed (range: 10-100 ng) or a single quantity of DNA can be used as a positive control.

4.4.5 Analysis of DNA methylation by McrBC-qPCR

DNA was extracted from wild type *P. tricornutum* and mutant strains in exponential phase of growth and collected simultaneously. The Rnase (Invitrogen) treatment was performed after DNA extraction. McrBC (New England Biolabs) digestion was followed by quantitative PCR (McrBC-qPCR). 500ng of genomic DNA is mixed with 2µl of McrBC 10U/µL, 2 µL Buffer 10X, 2 µL GTP 100mM, 2 µL BSA 100X and H2O for a final volume of 20µL. In parallel, the same mixture is prepared by replacing 2µL GTP which is the cofactor for McrBC enzyme by H2O (known as undigested sample). The digestion reaction has been tested in methylated DNA provided with the kit. Both samples are incubated at 37°C for two hours, followed by incubation at 65°C during 15 min to stop the reaction. Quantitative PCR is performed using a Roche LightCycler® 480 machine on equal amounts of digested and undigested DNA samples. 1 µl sample (2,5 ng of genomic DNA) is mixed to 5 µlLightCycler® DNA Master SYBR Green I 2X, 3 µl forward/reverse primers 1µM, and 1 µL H2O. The PCR program is performed as follow: 10 minutes at 95°C; 45 cycles of 95°C for 15 seconds and 60°C for 1 minute. Dissociation curves generated through a thermal denaturing step are used to verify amplification specificity. Two technical replicates are performed for each sample. The list of primers is shown in Table1. Efficiency and specificity of primers have been tested by quantitative PCR using serial dilutions of genomic DNA (5ng; 0,5ng; 0,1ng; and 0,05ng).

DNA methylation results are expressed as percentage of molecules lost through McrBC digestion. The ΔCts (Cycle threshold) is calculated by “Ct of digested sample normalized by internal control – Ct of undigested sample normalized by internal control”. The ΔCt is then transformed into percentage of DNA methylation as follow: $(1\text{-}2^{-(\Delta Ct)}) \times 100$. The analysis of DNA methylation level of unmethylated regions is used as an internal control for each digestion and for each qPCR experiment. Here H4 was also used as the housekeeping gene for

normalization. Any deviation from the expected result for these sequences (i.e. $\Delta Ct = 0$) invalids further analysis for other regions. It is nonetheless important to notice that pipetting errors are an important issue when considering qPCR experiments. In this particular example, lower level of DNA methylation (up to 30%) is produced by a $\Delta Ct= 0,5$, which is in the limit of hand pipetting reproducibility. Therefore, DNA methylation levels inferior to 30% must be considered as not significant.

Table 4.1 Primer pairs used for RT-PCR and McrBC-qPCR analysis

	chr	Position and size	Primer sequence
47034	12	238135-238302 168bp	5'ATAGTTATGTGCCCCGTTGG3' 5'CGTGGACTGAAGAGCAACAA3'
47839	15	203290-203531 242bp	5'CTCGTCAATTCATCGGTCCT3' 5'GCCGGAAGTGTTAGTGTGGT3'
47656	14	426617-426767 151bp	5'TCATTTCTGTCCGGAACCTC3' 5'GTCGATGCCGAGAAGGAATA3'
Codi2.4 TE1	1	2506134-2506291 159bp	5'CGACGTTGTTCAACTCGATG3' 5'TGCAATAAGGCCGACATAAA3'
Codi2.4 TE2	1	2503705-2523863 160bp	5'TTGTCGACACAGCGTTTTTC3' 5'GGCATGGAGAATCACATTCA3'
Codi4.1 TE3	8	20653-20830 179bp	5'AGACAAATGACCCGAGAGGA3' 5'CTTTCTGCATTGTTGCTTGC3'

4.4.6 Western blotting

Nuclear enriched protein was extracted by the protocol modified from chromatin extraction protocol. The protein was quantified by BCA kit. 10 μg protein of each sample was used for western blotting. 14% acrylamide Tris-Tricine gel was used for western blotting. The references of antibodies used for western blot are as follows: H3K27me3 (Millipore 07-449), H3K27me2 (Millipore: 07-452), H4 ace (Upstate: 06-598), H4 (Millipore: 07-108).

4.4.7 Phylogenetic analyse

The SMART database and the InterPro data were employed to identify conserved domains present in Ez, ESC, Su(z)12 and Nurf55 from different organisms. Amino acid sequences alignment and phylogenetic trees construction were made by MEGA program version 5. All the phylogenetic trees were constructed by neighbor-joining (NJ) method and bootstrap is support 1,000 replicates.

4.5 References

Akkers, R. C., van Heeringen, S. J., Jacobi, U. G., Janssen-Megens, E. M., Françoijs, K.-J., Stunnenberg, H. G., & Veenstra, G. J. C. (2009). A hierarchy of H3K4me3 and H3K27me3 acquisition in spatial gene regulation in Xenopus embryos. *Developmental cell*, *17*(3), 425–34. doi:10.1016/j.devcel.2009.08.005

Bartee, L., Malagnac, F., & Bender, J. (2001). Arabidopsis cmt3 chromomethylase mutations block non-CG methylation and silencing of an endogenous gene. *Genes & development*, *15*(14), 1753–8. doi:10.1101/gad.905701

Bastow, R., Mylne, J. S., Lister, C., Lippman, Z., Martienssen, R. a, & Dean, C. (2004). Vernalization requires epigenetic silencing of FLC by histone methylation. *Nature*, *427*(6970), 164–7. doi:10.1038/nature02269

Bestor, T. H. (2000). The DNA methyltransferases of mammals. *Human molecular genetics*, *9*(16), 2395–402. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/11005794

Bowler, C., Allen, A. E., Badger, J. H., Grimwood, J., Jabbari, K., Kuo, A., Maheswari, U., et al. (2008). The Phaeodactylum genome reveals the evolutionary history of diatom genomes. *Nature*, *456*(7219), 239–44. doi:10.1038/nature07410

Cao, X., Aufsatz, W., Zilberman, D., Mette, M. F., Huang, M. S., Matzke, M., & Jacobsen, S. E. (2003). Role of the DRM and CMT3 Methyltransferases in RNA-Directed DNA Methylation. *Current Biology*, *13*(24), 2212–2217. doi:10.1016/j.cub.2003.11.052

Cedar, H., & Bergman, Y. (2009). Linking DNA methylation and histone modification: patterns and paradigms. *Nature reviews. Genetics*, *10*(5), 295–304. doi:10.1038/nrg2540

Chanvivattana, Y., Bishopp, A., Schubert, D., Stock, C., Moon, Y.-H., Sung, Z. R., & Goodrich, J. (2004). Interaction of Polycomb-group proteins controlling flowering in Arabidopsis. *Development Cambridge England*, *131*(21), 5263–5276. Retrieved from http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=15456723

Chedin, F., Lieber, M. R., & Hsieh, C.-L. (2002). The DNA methyltransferase-like protein DNMT3L stimulates de novo methylation by Dnmt3a. *Proceedings of the National Academy of Sciences of the United States of America*, *99*(26), 16916–21. doi:10.1073/pnas.262443999

Chin, H. G., Pradhan, S., & Este, P. (2005). methyltransferase and p53 modulate expression of p53-repressed promoters.

Chopra, V. S., Hendrix, D. a, Core, L. J., Tsui, C., Lis, J. T., & Levine, M. (2011). The polycomb group mutant esc leads to augmented levels of paused Pol II in the Drosophila embryo. *Molecular cell*, *42*(6), 837–44. doi:10.1016/j.molcel.2011.05.009

Claude Courties, André Vaquer, Marc Troussellier, Jacques Lautier, Marie J. Chrétiennot-Dinet, Jacques Neveux, C. M. & H. C. (1994). smallest eukaryotic orgnanism.pdf.

Cuvelier, M. L., Allen, A. E., Monier, A., McCrow, J. P., Messié, M., Tringe, S. G., Woyke, T., et al. (2010). Targeted metagenomics and ecology of globally important uncultured eukaryotic phytoplankton. *Proceedings of the National Academy of Sciences of the United States of America*, *107*(33), 14679–84. doi:10.1073/pnas.1001665107

De Riso, V., Raniello, R., Maumus, F., Rogato, A., Bowler, C., & Falciatore, A. (2009). Gene silencing in the marine diatom Phaeodactylum tricornutum. *Nucleic acids research*, *37*(14), e96. doi:10.1093/nar/gkp448

Dillon, S. C., Zhang, X., Trievel, R. C., & Cheng, X. (2005). The SET-domain protein superfamily: protein lysine methyltransferases. *Genome biology*, *6*(8), 227. doi:10.1186/gb-2005-6-8-227

Dudley, K. J., Revill, K., Whitby, P., Clayton, R. N., & Farrell, W. E. (2008). Genome-wide analysis in a murine Dnmt1 knockdown model identifies epigenetically silenced genes in primary human pituitary tumors. *Molecular cancer research : MCR* , *6*(10), 1567–74. doi:10.1158/1541-7786.MCR-08-0234

Dunn, D. B., & Smith, J. D. (1958). The occurrence of 6-methylaminopurine in deoxyribonucleic acids. *The Biochemical journal*, *68*(4), 627–36. Retrieved from http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1200409&tool=pmcentrez&rendertype=abstract

Elgin, S. C. R., & Grewal, S. I. S. (2003). Heterochromatin: silence is golden. *Current biology : CB*, *13*(23), R895–8. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/14654010

Endoh, M., Endo, T. a, Endoh, T., Fujimura, Y., Ohara, O., Toyoda, T., Otte, A. P., et al. (2008). Polycomb group proteins Ring1A/B are functionally linked to the core transcriptional regulatory circuitry to maintain ES cell identity. *Development (Cambridge, England)*, *135*(8), 1513–24. doi:10.1242/dev.014340

Exner, V., Aichinger, E., Shu, H., Wildhaber, T., Alfarano, P., Caflisch, A., Gruissem, W., et al. (2009). The chromodomain of LIKE HETEROCHROMATIN PROTEIN 1 is essential for H3K27me3 binding and function during Arabidopsis development. *PloS one*, *4*(4), e5335. doi:10.1371/journal.pone.0005335

Falciatore, a, Casotti, R., Leblanc, C., Abrescia, C., & Bowler, C. (1999). Transformation of Nonselectable Reporter Genes in Marine Diatoms. *Marine biotechnology (New York, N.Y.)*, *1*(3), 239–251. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/10383998

Ficz, G., Branco, M. R., Seisenberger, S., Santos, F., Krueger, F., Hore, T. a, Marques, C. J., et al. (2011). Dynamic regulation of 5-hydroxymethylcytosine in mouse ES cells and during differentiation. *Nature*, *473*(7347), 398–402. doi:10.1038/nature10008

Flanagan, J. F., Mi, L.-Z., Chruszcz, M., Cymborowski, M., Clines, K. L., Kim, Y., Minor, W., et al. (2005). Double chromodomains cooperate to recognize the methylated histone H3 tail. *Nature*, *438*(7071), 1181–5. doi:10.1038/nature04290

Freitag, M., Williams, R. L., Kothe, G. O., & Selker, E. U. (2002). A cytosine methyltransferase homologue is essential for repeat-induced point mutation in Neurospora crassa. *Proceedings of the National Academy of Sciences of the United States of America*, *99*(13), 8802–7. doi:10.1073/pnas.132212899

Fritz, E. L., & Papavasiliou, F. N. (2010). Cytidine deaminases: AIDing DNA demethylation? *Genes & development*, *24*(19), 2107–14. doi:10.1101/gad.1963010

Garcia, B. a, Hake, S. B., Diaz, R. L., Kauer, M., Morris, S. a, Recht, J., Shabanowitz, J., et al. (2007). Organismal differences in post-translational modifications in histones H3 and H4. *The Journal of biological chemistry*, *282*(10), 7641–55. doi:10.1074/jbc.M607900200

Garcia, B. a, Shabanowitz, J., & Hunt, D. F. (2007). Characterization of histones and their post-translational modifications by mass spectrometry. *Current opinion in chemical biology*, *11*(1), 66–73. doi:10.1016/j.cbpa.2006.11.022

Goll, M. G., & Bestor, T. H. (2005). Eukaryotic cytosine methyltransferases. *Annual Review of Biochemistry*, *74*(1), 481–514. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/15952895

Gong, Z., & Zhu, J.-K. (2011). Active DNA demethylation by oxidation and repair. *Cell Research*, *21*(12), 1649–1651. doi:10.1038/cr.2011.140

Hattman, S. (2005). DNA-[adenine] methylation in lower eukaryotes. *Biochemistry Biokhimiia*, *70*(5), 550–8. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/15948708

Hattman, S., Kenny, C., Berger, L., & Pratt, K. (1978). Comparative study of DNA methylation in three unicellular eucaryotes. *Journal of bacteriology*, *135*(3), 1156–7. Retrieved from http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=222496&tool=pmcentrez&rendertype=abstract

Hennig, L., & Derkacheva, M. (2009). Diversity of Polycomb group complexes in plants: same rules, different players? *Trends in genetics : TIG* , *25*(9), 414–23. doi:10.1016/j.tig.2009.07.002

Heo, J. B., & Sung, S. (2011). Vernalization-mediated epigenetic silencing by a long intronic noncoding RNA. *Science (New York, N.Y.)*, *331*(6013), 76–9. doi:10.1126/science.1197349

Hermann, a, Gowher, H., & Jeltsch, a. (2004). Biochemistry and biology of mammalian DNA methyltransferases. *Cellular and molecular life sciences : CMLS* , *61*(19-20), 2571–87. doi:10.1007/s00018-004-4201-1

Holec, S., & Berger, F. (2012). Polycomb group complexes mediate developmental transitions in plants. *Plant physiology*, *158*(1), 35–43. doi:10.1104/pp.111.186445

Hsieh, C. (1999). In Vivo Activity of Murine De Novo In Vivo Activity of Murine De Novo Methyltransferases , Dnmt3a and Dnmt3b, *19*(12).

Ito, S., D'Alessio, A. C., Taranova, O. V., Hong, K., Sowers, L. C., & Zhang, Y. (2010). Role of Tet proteins in 5mC to 5hmC conversion, ES-cell self-renewal and inner cell mass specification. *Nature*, *466*(7310), 1129–33. doi:10.1038/nature09303

Iyer, L. M., Abhiman, S., & Aravind, L. (2011). *Natural history of eukaryotic DNA methylation systems. Progress in molecular biology and translational science* (1st ed., Vol. 101, pp. 25–104). Elsevier Inc. doi:10.1016/B978-0-12-387685-0.00002-0

Jacob, Y., Feng, S., LeBlanc, C. a, Bernatavichute, Y. V., Stroud, H., Cokus, S., Johnson, L. M., et al. (2009). ATXR5 and ATXR6 are H3K27 monomethyltransferases required for chromatin structure and gene silencing. *Nature structural & molecular biology*, *16*(7), 763–8. doi:10.1038/nsmb.1611

Jaenisch, R., & Young, R. (2008). Stem cells, the molecular circuitry of pluripotency and nuclear reprogramming. *Cell*, *132*(4), 567–82. doi:10.1016/j.cell.2008.01.015

Jones, P. a, & Baylin, S. B. (2007). The epigenomics of cancer. *Cell*, *128*(4), 683–92. doi:10.1016/j.cell.2007.01.029

Jurkowski, T. P., Meusburger, M., Phalke, S., Helm, M., Nellen, W., Reuter, G., & Jeltsch, A. (2008). Human DNMT2 methylates tRNA Asp molecules using a DNA methyltransferase-like catalytic mechanism, (1987), 1663–1670. doi:10.1261/rna.970408.transferase

Kangaspeska, S., Stride, B., Métivier, R., Polycarpou-Schwarz, M., Ibberson, D., Carmouche, R. P., Benes, V., et al. (2008). Transient cyclical methylation of promoter DNA. *Nature*, *452*(7183), 112–5. doi:10.1038/nature06640

Karrer, K. M., & VanNuland, T. a. (2002). Methylation of adenine in the nuclear DNA of Tetrahymena is internucleosomal and independent of histone H1. *Nucleic acids research*, *30*(6), 1364–70. Retrieved from http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=101364&tool=pmcentrez&rendertype=abstract

Kerppola, T. K. (2009). Polycomb group complexes--many combinations, many functions. *Trends in cell biology*, *19*(12), 692–704. doi:10.1016/j.tcb.2009.10.001

Klose, R. J., Kallin, E. M., & Zhang, Y. (2006). JmjC-domain-containing proteins and histone demethylation. *Nature reviews. Genetics*, *7*(9), 715–27. doi:10.1038/nrg1945

Kouzminova, E., & Selker, E. U. (2001). dim-2 encodes a DNA methyltransferase responsible for all known cytosine methylation in Neurospora. *The EMBO journal*, *20*(15), 4309–23. doi:10.1093/emboj/20.15.4309

Köhler, C., & Hennig, L. (2010). Regulation of cell identity by plant Polycomb and trithorax group proteins. *Current opinion in genetics & development*, *20*(5), 541–7. doi:10.1016/j.gde.2010.04.015

LeRoy, G., Weston, J. T., Zee, B. M., Young, N. L., Plazas-Mayorca, M. D., & Garcia, B. a. (2009). Heterochromatin protein 1 is extensively decorated with histone code-like post-translational modifications. *Molecular & cellular proteomics MCP* , *8*(11), 2432–42. doi:10.1074/mcp.M900160-MCP200

Lechner, M., Boshoff, C., & Beck, S. (2010). *Cancer epigenome. Advances in genetics* (1st ed., Vol. 70, pp. 247–76). Elsevier Inc. doi:10.1016/B978-0-12-380866-0.60009-5

Li, Y., Bollag, G., Clark, R., Conroy, J. S. L., Fults, D., Ward, K., Friedman, E., et al. (1992). Somatic Mutations in Human Tumors in the Neurofibromatosis 1 Gene, *69*, 275–281.

Lister, R., Pelizzola, M., Dowen, R. H., Hawkins, R. D., Hon, G., Tonti-Filippini, J., Nery, J. R., et al. (2009). Human DNA methylomes at base resolution show widespread epigenomic differences. *Nature*, *462*(7271), 315–22. doi:10.1038/nature08514

Liu, Y., Taverna, S. D., Muratore, T. L., Shabanowitz, J., Hunt, D. F., & Allis, C. D. (2007). RNAi-dependent H3K27 methylation is required for heterochromatin formation and DNA elimination in Tetrahymena. *Genes & Development*, *21*(12), 1530–1545. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/17575054

Lyko, F., Ramsahoye, B. H., Kashevsky, H., Tudor, M., Mastrangelo, M. a, Orr-Weaver, T. L., & Jaenisch, R. (1999). Mammalian (cytosine-5) methyltransferases cause genomic DNA methylation and lethality in Drosophila. *Nature genetics*, *23*(3), 363–6. doi:10.1038/15551

Mabley, J. G., Deb, A., Wallace, R., Elder, R. H., Sciences, B., Building, C., Road, L., et al. (2004). regulating inflammation, *18*, 1–18.

Margueron, Raphael, Li, G., Sarma, K., Blais, A., Zavadil, J., Woodcock, C. L., Dynlacht, B. D., et al. (2008). Ezh1 and Ezh2 maintain repressive chromatin through different mechanisms. *Molecular cell*, *32*(4), 503–18. doi:10.1016/j.molcel.2008.11.004

Margueron, Raphaël, & Reinberg, D. (2011). The Polycomb complex PRC2 and its mark in life. *Nature*, *469*(7330), 343–9. doi:10.1038/nature09784

Maumus, F., Rabinowicz, P., Bowler, C., Rivarola, M., Inserm, U., Nacional, I., Agropecuaria, D. T., et al. (2011). Stemming Epigenetics in Marine Stramenopiles. *Current*.

Morgan, H. D., Dean, W., Coker, H. a, Reik, W., & Petersen-Mahrt, S. K. (2004). Activation-induced cytidine deaminase deaminates 5-methylcytosine in DNA and is expressed in pluripotent tissues: implications for epigenetic reprogramming. *The Journal of biological chemistry*, *279*(50), 52353–60. doi:10.1074/jbc.M407695200

Métivier, R., Gallais, R., Tiffoche, C., Le Péron, C., Jurkowska, R. Z., Carmouche, R. P., Ibberson, D., et al. (2008). Cyclical DNA methylation of a transcriptionally active promoter. *Nature*, *452*(7183), 45–50. doi:10.1038/nature06544

Nowak, A. J., Alfieri, C., Stirnimann, C. U., Rybin, V., Baudin, F., Ly-Hartig, N., Lindner, D., et al. (2011). Chromatin-modifying complex component Nurf55/p55 associates with histones H3 and H4 and polycomb repressive complex 2 subunit Su(z)12 through partially overlapping binding sites. *The Journal of biological chemistry*, *286*(26), 23388–96. doi:10.1074/jbc.M110.207407

Okano, M., Bell, D. W., Haber, D. a, & Li, E. (1999). DNA methyltransferases Dnmt3a and Dnmt3b are essential for de novo methylation and mammalian development. *Cell*, *99*(3), 247–57. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/10555141

Ooi, S. K. T., & Bestor, T. H. (2008). The colorful history of active DNA demethylation. *Cell*, *133*(7), 1145–8. doi:10.1016/j.cell.2008.06.009

Ooi, S. K. T., Qiu, C., Bernstein, E., Li, K., Jia, D., Yang, Z., Erdjument-Bromage, H., et al. (2007). DNMT3L connects unmethylated lysine 4 of histone H3 to de novo methylation of DNA. *Nature*, *448*(7154), 714–7. doi:10.1038/nature05987

Ponger, L., & Li, W.-H. (2005). Evolutionary diversification of DNA methyltransferases in eukaryotic genomes. *Molecular biology and evolution*, *22*(4), 1119–28. doi:10.1093/molbev/msi098

Popp, C., Dean, W., Feng, S., Cokus, S. J., Andrews, S., Pellegrini, M., Jacobsen, S. E., et al. (2010). Genome-wide erasure of DNA methylation in mouse primordial germ cells is affected by AID deficiency. *Nature*, *463*(7284), 1101–5. doi:10.1038/nature08829

Rajasekhar, V. K., & Begemann, M. (2007). Concise review: roles of polycomb group proteins in development and disease: a stem cell perspective. *Stem cells (Dayton, Ohio)*, *25*(10), 2498–510. doi:10.1634/stemcells.2006-0608

Rinn, J. L., Kertesz, M., Wang, J. K., Squazzo, S. L., Xu, X., Brugmann, S. a, Goodnough, L. H., et al. (2007). Functional demarcation of active and silent chromatin domains in human HOX loci by noncoding RNAs. *Cell*, *129*(7), 1311–23. doi:10.1016/j.cell.2007.05.022

Rogers, S. D., Rogers, M. E., Saunders, G., & Holt, G. (1986). Isolation of mutants sensitive to 2-aminopurine and alkylating agents and evidence for the role of DNA methylation in Penicillium chrysogenum. *Current Genetics*, *10*(7), 557–560. Retrieved from http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=2832076

Sandigursky, M., Yacoub, a, Kelley, M. R., Xu, Y., Franklin, W. a, & Deutsch, W. a. (1997). The yeast 8-oxoguanine DNA glycosylase (Ogg1) contains a DNA deoxyribophosphodiesterase (dRpase) activity. *Nucleic acids research*, *25*(22), 4557–61. Retrieved from http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=147074&tool=pmcentrez&rendertype=abstract

Schaefer, M., Pollex, T., Hanna, K., Tuorto, F., Meusburger, M., Helm, M., & Lyko, F. (2010). RNA methylation by Dnmt2 protects transfer RNAs against stress-induced cleavage. *Genes & development*, *24*(15), 1590–5. doi:10.1101/gad.586710

Schuettengruber, B., Martinez, A.-M., Iovino, N., & Cavalli, G. (2011). Trithorax group proteins: switching genes on and keeping them active. *Nature reviews. Molecular cell biology*, *12*(12), 799–814. doi:10.1038/nrm3230

Shaver, S., Casas-Mollano, J. A., Cerny, R. L., & Cerutti, H. (2010). Origin of the polycomb repressive complex 2 and gene silencing by an E(z) homolog in the unicellular alga Chlamydomonas. *Epigenetics : official journal of the DNA Methylation Society*, *5*(4), 301–12. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/20421736

Sheehan, A. M., McGregor, D. K., Patel, A., Shidham, V., Fan, C.-Y., & Chang, C.-C. (2005). Expression of human 8-oxoguanine DNA glycosylase (hOGG1) in follicular lymphoma. *Modern pathology : an official journal of the United States and Canadian Academy of Pathology, Inc*, *18*(11), 1512–8. doi:10.1038/modpathol.3800461

Simon, J. a, & Kingston, R. E. (2009). Mechanisms of polycomb gene silencing: knowns and unknowns. *Nature reviews. Molecular cell biology*, *10*(10), 697–708. doi:10.1038/nrm2763

Suganuma, T., Pattenden, S. G., & Workman, J. L. (2008). Diverse functions of WD40 repeat proteins in histone recognition. *Genes & development*, *22*(10), 1265–8. doi:10.1101/gad.1676208

Vanyushin, B. F. (2005). AND EXPERIMANTAL ARTICLES Adenine Methylation in Eukaryotic DNA. *Text*, *39*(4), 473–481.

Whitcomb, S. J., Basu, A., Allis, C. D., & Bernstein, E. (2007). Polycomb Group proteins: an evolutionary perspective. *Trends in genetics : TIG*, *23*(10), 494–502. doi:10.1016/j.tig.2007.08.006

Williams, K., Christensen, J., Pedersen, M. T., Johansen, J. V., Cloos, P. a C., Rappsilber, J., & Helin, K. (2011). TET1 and hydroxymethylcytosine in transcription and DNA methylation fidelity. *Nature*, *473*(7347), 343–8. doi:10.1038/nature10066

Wu, H., D'Alessio, A. C., Ito, S., Xia, K., Wang, Z., Cui, K., Zhao, K., et al. (2011). Dual functions of Tet1 in transcriptional regulation in mouse embryonic stem cells. *Nature*, *473*(7347), 389–93. doi:10.1038/nature09934

Xhemalce, B., & Kouzarides, T. (2010). A chromodomain switch mediated by histone H3 Lys 4 acetylation regulates heterochromatin assembly. *Genes & development*, *24*(7), 647–52. doi:10.1101/gad.1881710

Xu, C., Bian, C., Yang, W., Galka, M., Ouyang, H., Chen, C., Qiu, W., et al. (2010). Binding of different histone marks differentially regulates the activity and specificity of polycomb repressive complex 2 (PRC2). *Proceedings of the National Academy of Sciences of the United States of America*, *107*(45), 19266–71. doi:10.1073/pnas.1008937107

Youn, C.-K., Song, P. I., Kim, M.-H., Kim, J. S., Hyun, J.-W., Choi, S.-J., Yoon, S. P., et al. (2007). Human 8-oxoguanine DNA glycosylase suppresses the oxidative stress induced apoptosis through a p53-mediated signaling pathway in human fibroblasts. *Molecular cancer research : MCR*, *5*(10), 1083–98. doi:10.1158/1541-7786.MCR-06-0432

Zhang, X., Clarenz, O., Cokus, S., Bernatavichute, Y. V., Pellegrini, M., Goodrich, J., & Jacobsen, S. E. (2007). Whole-genome analysis of histone H3 lysine 27 trimethylation in Arabidopsis. *PLoS biology*, *5*(5), e129. doi:10.1371/journal.pbio.0050129

Zhang, X., Yazaki, J., Sundaresan, A., Cokus, S., Chan, S. W.-L., Chen, H., Henderson, I. R., et al. (2006). Genome-wide high-resolution mapping and functional analysis of DNA methylation in arabidopsis. *Cell*, *126*(6), 1189–201. doi:10.1016/j.cell.2006.08.003

Zhu, J.-K. (2009). Active DNA demethylation mediated by DNA glycosylases. *Annual review of genetics*, *43*, 143–66. doi:10.1146/annurev-genet-102108-134205

Chapter V

Discussion and perspectives

The genome sequences of diatoms have to some extent provided insights that help explain the great success and dominance of these organisms in marine environments by revealing their novel and extraordinary metabolisc capacities Compared to the situation on land, the marine environment is more fluctuating because of the presence of currents and turbulence. Marine organisms have to deal with abiotic factors such as currents, tides, light and UV radiation, temperature, nutrient limitation as well as many other biotic factors. Photosynthetic organisms such as diatomsalso have to cope with changing light environments, such as fluctuating intensities when they drift between different layers of sea water.

Although the melting pot genome of *P. tricornutum* has conferred novel metabolic capabilities. But it is still not clear how these genes with different origins interact together in the genome. Another question is whether the diatom epigenome is more similar to that in plants or animals, or whether it has unique features which do not resemble other well studied models. Investigating diatom epigenetic regulation can enhance our understanding of how diatoms adapt to fluctuating oceanic environment and thrive in the oceans.

In this thesis, I explored the epigenetic world of a phytoplankton organism with evolutionary and ecological importance providing us with an opportunity to understand how an epigenome is organized in a single celled marine alga, what the interactions between the different epigenetic components are, and how this correlates with other well studied models. I pioneered the diatom epigenome research by investigating *P. tricornutum* epigenome specifically by determing genome-wide distribution profile of different histone marks generated by ChIP-seq combined with DNA methylation .In my thesis, Chapter II is about the investigation of DNA methylation in *P. tricornutum*. DNA methylation is the most extensively studied epigenetic component is widely distributed in different branches of life such as mammals and plants, but essentially absent in some model species such as *Saccharomyces cerevisiae*, *Schizosaccharomyces pombe*, *Drosophila melanogaster* and *Caenorhabditis elegans*. Feng et al and Zilberman extended the list of methylomes from model species to various species and investigated the characteristics of DNA methylation distribution in an evolutionary context (Feng & Jacobsen, 2011; Zemach et al., 2010a). However, all the whole genome methylomes that were reported were only confined to Unikont and Archaeplastida. It was therefore worth exploring the methylome of a diatom derived from an important eukaryotic group, the Stramenopiles, and of ecological significance

of diatoms as well as evolutionary. Chapter II reports the genome wide DNA methylation distribution in *P. tricornutum* under normal conditions by utilizing McrBC-ChIP methodology. In generalal *P. tricornutum* was found to have low DNA methylation with relatively extensively methylated TEs and a few methylated genes. Copia type TEs and longer TEs tend be more methylated compared to other groups of TEs which is not surprising because these elements are likely to be young and active TEs and therefore easily mobilized in the genome. Methylation on these TEs can control their mobility and maintain the stability of the genome.

Furthermore, RNA-seq data was generated to evaluate the correlation between gene expression and gene DNA methylation. Even though gene methylation is not significant in *P. tricornutum* (only around 3% of genes) compared to model plant species such as *Arabidopsis thaliana* (about one third of genes are methylated) and mammals (Feng et al., 2010), I found that genes with body methylation tend to be lowly or moderately. Thanks to the availability of EST libraries under 16 conditions, we had a good opportunity to study the correlation between gene methylation and gene expression under different conditions. One group of methylated genes which have heavy and complete methylation attracted our attention: these methylated genes have no expression under normal condition but some of them express under certain specific conditions. In other words, these genes are under tight control and only express when specific environmental cues occur. I speculate that the expression of these genes probably is caused by demethylation or hypomethylation of these genes under specific conditions which has to be confirmed by checking their DNA methylation status. If it is the case, the heavily and completely methylated genes show that they are controlled and regulated by the hypermethylation in response to environmental cues.

It is interesting that *Ectocarpus siliculosus* another Stramenopiles was found not to containing any putative DNA methyltranferases based on its whole genome sequence and nor any detectable DNA methyation when investigated by HPLC (Cock et al., 2010; Jarvis, Dunahay, & Brown, 1992). It will be interesting in the future to explore why these two closely evolutionary related organisms have distinct DNA methylation pattern. The *P. tricornutum* methylome, which is the first methylome from Stramenopiles, will furthermore contribute to our understanding of the evolution of DNA methylation in eukaryotes.

In the future, additional methylomes of different organisms belonging to different lineages can help better understand the evolutionary role of DNA methylation which is considered as an ancient epigenetic mark. Furthermore, genome wide distribution of DNA methylation under different conditions will shed light on whether gene methylation can dynamically regulate gene transcription. Overall, my work established the foundation for further investigation of DNA methylation under different relevant environmental conditions in *P. tricorntutum* but also other diatoms and algae.

Although the genome wide distribution of DNA methylation in *P. tricornutum* has been profiled, the McrBC- ChIP method I used cannot distinguish the different contexts of DNA methylation: CG, CHG and CHH. Bisulfite sequencing, which can generate single base resolution DNA methlyation maps and distinguish CG, CHG and CHH contexts, is now considered as the gold method for investigating DNA methylation. Initially, CG methylation was considered as the only context methylation in mammals, but recently in human stem cells, non CG methylation was also detected (Lister et al., 2009). In plants, all three contexts of DNA methylation have been detected on TEs but only CG methyation is found on genes. CG methylation is the only context of DNA methylation on genes which is conserved among all the organisms where methylomes have been examined. So far only CHG methylation on genes was found in *Chlamydomonas* (Zemach, McDaniel, Silva, & Zilberman, 2010b). For validation purposes, I carried out bisulfite sequencing on 28 loci evenly distributed on the genome. Only CG methylation was found among all the loci I have examined. It therefore seems that CG methylation is the dominant DNA methylation context in *P. tricornutum.* However, among the chosen loci, most of them are genes.Whole genome bisulfite sequencing should therefore be performed to give us a better picture of DNA methylation context in *P. tricornutm*. However, bisulfite sequencing also has its drawbacks: it cannot distinguish 5mc and 5mc-OH and because the unmethylated C is converted to U and finally becomes T, it is a difficult challenge for mapping the regions containing repeated sequences. So bisulfite sequencing with McrBC-ChIP approach would compensate the drawbacks of each other and provide more accurate DNA methylation patterns.

Another major epigenetic component is histone modification. Genome wide studies of epigeneomic landscape have been conducted in yeast (Millar & Grunstein, 2006), *C. elegans* (T. Liu et al., 2011), *Drosophila* (Filion et al., 2010; Kharchenko et al., 2011) *Arabidopsis*

(Roudier et al., 2011a), rice (X. Li et al., 2008), maize (Elling & Deng, 2009), frog (Akkers et al., 2009), and human cells (Ernst et al., 2011; Hawkins et al., 2010; Marks et al., 2012). All these studies demonstrate that there is a relatively low combinatorial complexity of chromatin marks. In *D. melanogaster*, the integrative analysis of distribution of chromatin marks revealed five principal chromatin types (Filion et al., 2010). In *Arabidopsis*, 12 marks along the genomic sequence define four main chromatin states: active genes, repressed genes, silent repeat elements and intergenic regions (Roudier et al., 2011a).

H3K4me2, H3K9me2 and H3K27me3, histone modifications were chosen for genome wide distribution study because H3K4me2 is usually associated with genes while H3K9me2 is associated with TEs and H3K27me3 is associated with repressed gene in plants and metazoans. In *P. tricornutum* we found that these three histone modifications cover almost half the genome which is not the case in plants and metazoans. The high coverage of three histone marks might be due to the fact that the *P. tricornutum* genome is much smaller with shorter gene size and intergenic lengths.

As expected, H3K4me2 marks genes without discrimination for gene expression level while H3K9me2 is mainly associated with TEs in *P. tricornutum*. To our surprise, H3K27me3, which is closely associated with Polycomb protein, has a novel distribution in *P. tricornutum*: it is mainly associated with TEs and not repressed genes as observed in other organisms. H3K27me3 and Polycomb group protein were considered only to exist in multicellular organisms. Our work is the first genome wide investigation of H3K27me3 distribution in a unicellular organism. The results not only demonstrated the existence of H3K27me3 in *P. tricornutum* but also raised the question of its potential functions in unicellular organisms. The majority of H3K27me3-marked regions are TEs, implying that H3K27me3 can probably regulate the activity of TEs and may mark facultative heterochromatin regions. Beside TEs, some genes are marked by H3K27me3. These genes are repressed and differentially expressed under different conditions, a notion which is consistent with previous studies in plants and metazoans. It seems that H3K27me3 in *P. tricornutum* might have a similar function on genes as found in multicellular organisms. H3K27me3 is considered associated with development and cell identity in multicellular organisms. Although in single celled species, cell differentiation does not exist, the transition from one cell type to another such as gametophyte

and sporophyte is very common. It is therefore an interesting question to further investigate the functions of H3K27me3 in unicellular organisms.

Another interesting phenomenon I found is that H3K27me3 and H3K9me2 tend to co-mark heavily methylated genes and TEs. Most of the heavily methylated genes which are marked by H3K9me2 and H3K27me3 are not expressed under except under specific conditions. It seems that these epigenetic marks may cooperatively regulate genes and TEs in *P. tricornutum*. I speculate that in such a small genome, the histone marks probably tend to highly overlap with each other. In such a compact genome, the patterns of combinatorial marks rather than single epigenetic marks may serve like an epigenetic code exerting profound impacts on chromatin states.

P. tricornutum not only has putative histone lysine methyltransferases but other putative enzymes responsible for other histone modifications such as lysine acetylation. It will be of great interest to conduct ChIP-seq studies in order to explore genome wide distribution of different histone modifications. Recent studies show that not only histone modifications but also histone variants can influence transcription (Stroud, Otero, Desvoyes, Ramírez-parra, & Jacobsen, 2012). H3 of *P. tricornutum* has different variants (Pt11841 H3-1a, Pt50872 H3-1b and Pt50695 H3-1c). In the future epigenetic studies in *P. tricornutum*, histone variants should therefore also be considered.

Although, the genome of *P. tricornutum* is small it contains complicated epigenetic regulation systems that are lacking in budding yeast, such as DNA methylation and polycomb group protein regulation systems. In the light of the fact that classical model species such as *S. cerevisiae*, *D. melanogaster* and *C. elegans* have no or very low DNA methylation, the correlation of histone modifications and DNA methylation, cannot be assessed. However, *P. tricornutum*, in spite of being as a unicellular organism displays complex epigenetics features. and Being ploymorphic even more attractive for epigenetic studies. I therefore hope in the future to see its emergence as a model for epigenetic studies in unicellular organisms. Notwithstanding, a crucial step in the near future is to perform mass spectrometry anaylsis of diatom nucleosomes in order to determine the extent of histone modifications in *P. tricornutum*.

The unorthodox distribution of H3K27me3, which is closely associated with Polycomb group proteins, raises the question of their role in unicellular organisms which is discussed in the next chapter. In Chapter IV, I further investigated polycomb group protein components in eukaryotic algae based on comparative genome analyses. The results further confirmed that the PcG system does not only exist in multicellular organisms but also in unicellular organisms because all the unicellular eukaryotic algae I have examined contain putative PRC2 components. I further explored how PcG proteins function in unicellular organisms by knocking down E(z) homolog gene (32817) in *P. tricornutum*. The preliminary results have shown that *P. tricornutum* Ez (32817) knockdown lines have decreased H3K27me3 levels compared to wild type. It is interesting to profile the genome-wide distribution of H3K27me3 in E(z) knockdown line to further investigate the functions of E(z) and H3K27me3 deposition in *P. tricornutum*. Furthermore, profiling other chromatin modifications in E(z) knockdown mutants, such as DNA methylation and H3K4me3 which antagonizes H3K27me3, can help better understand interactions between different chromatin modifications in *P. tricornutum*.

I also investigated the functions of putative DNA methyltransferases in *P. tricornutum* by reverse genetic manipulation methodology. According to preliminary results, two putative methyltransferases (Pt45072 and Pt47357) knockdown lines have shown decreased genome wide DNA methylation and at several specific loci compared to wild type. In the future it will be worth to explore genome wide distribution of DNA methylation in these mutants by bisulfite sequencing to further dissect the functions of these DNA methyltransferases. This will provide answers about which context of DNA methylation they are mainly responsible for and whether they have preferences for genes or TEs.

All together my work as the first epigenome of a Stramenopile has shown novel characteristics revealing some conserved and different mechanisms with respect to other eukaryotic. I hope that my work on *P. tricornutum* epigenome will be the foundation for future diatom epigenetic studies and a great opportunity for comparative epigenomics and the elucidation of relationships between dynamic genome evolution and epigenetic regulation.

Diatoms are also good models for epigenetic variation, genetic variation and their interaction with the environment. It is known that diversity within and between species is primarily driven by genetic variation. Epigenetic variation is due to DNA methylation or histone

modification changes but of contribution of epigenetic variation to the phenotype is still not clear. Three epigenetic alleles have been recently defined: obligate epialleles which display a complete dependency on a genetic variant; the second class called facilitated epialleles is due to the presence of a genetic variant such as a nearby transposon insertion, but their maintenance is not necessarily dependent on this variant; the third class is pure epialleles which are independent of any genetic variation (Schmitz & Ecker, 2012). The "epigenotype" refers to mitotically heritable patterns of DNA methylation at CpG dinucleotides and modifications to chromatin proteins that package DNA (Whitelaw & Whitelaw, 2006). There are still open questions: 1. How is epigenotype influenced by environment and genotype? 2. Between epigenotype and genotype, which has greater impact on the phenotype? 3. How is the epigenotype at any particular locus established and maintained? 4. How do genetic and epigenetic regulations coordinate together in the real world? In this context and knowing that diatoms live in various environments and deal with rapidly fluctuating conditions, reversible and flexible epigenetic regulation is likely to be involved in the process of adapting and coping with the environment. Thus, it will be exciting to explore epigenetic and genetic variation in relation to the environment in diatoms together.

Nowadays, the changing climate greatly influences the global ecology, such as ocean acidification and higher temperatures caused by increased CO_2 release. It will be very exciting to mimic such environmental changes and to profile epigenomic map under such conditions. It will be very interesting to examine the extent of epigenetic and genetic phenomena in diatoms and the crosstalk between these two layers of soft and hard inheritance in natural populations in a real world setting.

Eleven different *P. tricornutum* ecotypes from different locations have been obtained from all over the world with different living conditions. Many previous studies only focused on the genetic diversity of geographically isolated strains of the same diatom species without consideration of epigenetic diversity. It can be envisaged that epigenotype studies on diatom species can bring better understanding of epigenotype, genotype and phenotype in the context of the marine environment. Genome wide distribution analyses of DNA methylation and histone modification combined with SNPs analyses of these strains can enhance our understanding of genetic and epigenetic variation, their crosstalk and relationship to the environment.

It is known that some epigenetic marks can mediate cell differentiation and development in multicellular highly organized organisms (Ahmad et al., 2010; Meissner, 2010). *P. tricornutum* cells, although unicellular, can exist in four different morphotypes, fusiform, triradiate, round and oval. The cells can switch from "fusiform" to "oval" during biofilm formation (Bowler, De Martino, & Falciatore, 2010). It is tempting to propose that epigenetic mechanisms might be involved in morphotype transitions because they occur rapidly. Comparing epigenome profiles of different morphotypes, combined with RNA-seq data may detect expression variation of morphotype-specific genes, pinpoint different epigenetic profiles of these genes, and reveal the networks involved in morphogenesis regulation. It might also provide clues on the degree of conservation of the epigenetic mechanisms involved in cell type variation in unicellular diatoms and multicellular organisms.

Diatoms can form resting spores when the environment becomes inhospitable in order to survive through stressful times. More than 130 diatom species have been found capable of forming resting spores with more heavily silicified frustules compared to normal cells (Xie, 2006). These resting spores are dormant and can be buried in the bottom sediment for a long time until the nutrient conditions are amiable. 100 year old *Skeletonema marinoi* resting spores can still revive (Härnström, Ellegaard, Andersen, & Godhe, 2011). During 100 years, there are no traces at all of genetic impact from the open sea populations on the diatoms in Mariager Fjord where these diatom spores were collected (Härnström et al., 2011). It will be very exciting to compare not only genetic profiles but also epigenetic profiles between resting spores 100 years ago with current populations and to investigate the diatoms' genetic and epigenetic adaptation to climate change. The mechanisms of transition from normal cells to resting spores are still enigmatic. Epigenetic regulation may play a vital role in this transition. *P. tricornutum* may be a good diatom model species to study this because morphotype transitions from fusiform or triradiate to oval may be similar with transitions from normal cells and resting spores in other diatom species.

Although sexual reproduction has not yet been reported in *P. tricornutum*, it is very common among diatom species. However, the frequency of sexual reproduction in diatoms is largely unconstrained, with estimates ranging from once per year to once every 40 years (Significance & Histories, 1997). It has been shown that sexual reproduction occurs when cell size is too small, then the sexual reproduction will bring the diatom cells back to big size

(Armbrust & Galindo, 2001). Sexual reproduction in diatoms may also involve epigenetic regulation as in higher plants and animals. Epigenome studies can be expanded to other diatom species especially those with more ecological relevance. Nowadays, more diatom species are being or will be sequenced. *Pseudo-nitzchia multiseries* (~250Mb), and *Fragilariopsis cylindrus* (~81Mb) are being sequenced and will be completed soon by the Joint Genome Institute. *Amphora sp* CCMP 2378 is also being sequenced (Maumus et al., 2011). More fully sequenced diatom genomes will not only provide an opportunity for diatom genome comparison studies but also will build a basic foundation for diatom epigenome studies.

Pseudo-nitzschia multiseries is a pennate diatom species capable of producing the neurological toxin domoic acid (DA), which can accumulate through food chains and lead to sickness and even mortality in seabirds, sea mammals and human. In 1987, the bloom caused by *P. multiseries* led to the death of three people in Prince Edward in Canada (Mos, 2001). It is still mysterious that DA production capability varies between strains isolated from different locations. Although some studies showed that the interaction between extra bacteria and *P. multiseries* can enhance DA production (Mos, 2001), the mechanism of DA production is still unclear. We can speculate that the epigenetic regulation is implicated in DA production because of the variability in DA production found among strains even though others speculate that this might be due to the presence of bacteria that do the DA production. The coming fully sequenced genome may shed light on these issues.

As my thesis work demonstrates, epigenetic in diatoms a fascinating research topic that is likely to lead to new insights about its role during evolution. I hope my studies will lead the way to further research aimed at exploring epigenetic phenomena in a truly environmnental context.

References:

Ahmad, A., Zhang, Y., & Cao, X.-F. (2010). Decoding the epigenetic language of plant development. *Molecular plant*, *3*(4), 719–28. doi:10.1093/mp/ssq026

Akkers, R. C., van Heeringen, S. J., Jacobi, U. G., Janssen-Megens, E. M., Françoijs, K.-J., Stunnenberg, H. G., & Veenstra, G. J. C. (2009). A hierarchy of H3K4me3 and H3K27me3 acquisition in spatial gene regulation in Xenopus embryos. *Developmental cell*, *17*(3), 425–34. doi:10.1016/j.devcel.2009.08.005

Armbrust, E. V., & Galindo, H. M. (2001). Rapid Evolution of a Sexual Reproduction Gene in Centric Diatoms of the Genus Thalassiosira Rapid Evolution of a Sexual Reproduction Gene in Centric Diatoms of the Genus Thalassiosira, *67*(8). doi:10.1128/AEM.67.8.3501

Bates, S. S. (n.d.). Ecophysiology and metabolism of ASP toxin production. *Journal of wound, ostomy, and continence nursingofficial publication of The Wound, Ostomy and Continence Nurses Society / WOCN*. doi:10.1097/WON.0b013e3181d73aab

Bowler, C., De Martino, A., & Falciatore, A. (2010). Diatom cell division in an environmental context. *Current opinion in plant biology*, *13*(6), 623–30. doi:10.1016/j.pbi.2010.09.014

Cock, J. M., Sterck, L., Rouzé, P., Scornet, D., Allen, A. E., Amoutzias, G., Anthouard, V., et al. (2010). The Ectocarpus genome and the independent evolution of multicellularity in brown algae. *Nature*, *465*(7298), 617–21. doi:10.1038/nature09016

Elling, A. a, & Deng, X. W. (2009). Next-generation sequencing reveals complex relationships between the epigenome and transcriptome in maize. *Plant signaling & behavior*, *4*(8), 760–2. doi:10.1105/tpc.109.065714

Ernst, J., Kheradpour, P., Mikkelsen, T. S., Shoresh, N., Ward, L. D., Epstein, C. B., Zhang, X., et al. (2011). Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature*, *473*(7345), 43–9. doi:10.1038/nature09906

Feng, S., Cokus, S. J., Zhang, X., Chen, P.-Y., Bostick, M., Goll, M. G., Hetzel, J., et al. (2010). Conservation and divergence of methylation patterning in plants and animals.

Proceedings of the National Academy of Sciences of the United States of America, *107*(19), 8689–94. doi:10.1073/pnas.1002720107

Feng, S., & Jacobsen, S. E. (2011). Epigenetic modifications in plants: an evolutionary perspective. *Current opinion in plant biology*, *14*(2), 179–86. doi:10.1016/j.pbi.2010.12.002

Filion, G. J., van Bemmel, J. G., Braunschweig, U., Talhout, W., Kind, J., Ward, L. D., Brugman, W., et al. (2010). Systematic protein location mapping reveals five principal chromatin types in Drosophila cells. *Cell*, *143*(2), 212–24. doi:10.1016/j.cell.2010.09.009

Hawkins, R. D., Hon, G. C., Lee, L. K., Ngo, Q., Lister, R., Pelizzola, M., Edsall, L. E., et al. (2010). Distinct epigenomic landscapes of pluripotent and lineage-committed human cells. *Cell stem cell*, *6*(5), 479–91. doi:10.1016/j.stem.2010.03.018

Härnström, K., Ellegaard, M., Andersen, T. J., & Godhe, A. (2011). Hundred years of genetic structure in a sediment revived diatom population. *Proceedings of the National Academy of Sciences of the United States of America*, *108*(10), 4252–7. doi:10.1073/pnas.1013528108

Jarvis, E. E., Dunahay, T. G., & Brown, L. M. (1992). DNA NUCLEOSIDE COMPOSITION AND METHYLATION IN SEVERAL SPECIES OF MICROALGAE1. *Journal of Phycology*, *28*(3), 356–362.

Kharchenko, P. V., Alekseyenko, A. a, Schwartz, Y. B., Minoda, A., Riddle, N. C., Ernst, J., Sabo, P. J., et al. (2011). Comprehensive analysis of the chromatin landscape in Drosophila melanogaster. *Nature*, *471*(7339), 480–5. doi:10.1038/nature09725

Li, X., Wang, X., He, K., Ma, Y., Su, N., He, H., Stolc, V., et al. (2008). High-resolution mapping of epigenetic modifications of the rice genome uncovers interplay between DNA methylation, histone methylation, and gene expression. *The Plant cell*, *20*(2), 259–76. doi:10.1105/tpc.107.056879

Lister, R., Pelizzola, M., Dowen, R. H., Hawkins, R. D., Hon, G., Tonti-Filippini, J., Nery, J. R., et al. (2009). Human DNA methylomes at base resolution show widespread epigenomic differences. *Nature*, *462*(7271), 315–22. doi:10.1038/nature08514

Liu, T., Rechtsteiner, A., Egelhofer, T. a, Vielle, A., Latorre, I., Cheung, M.-S., Ercan, S., et al. (2011). Broad chromosomal domains of histone modification patterns in C. elegans. *Genome research*, *21*(2), 227–36. doi:10.1101/gr.115519.110

Marks, H., Kalkan, T., Menafra, R., Denissov, S., Jones, K., Hofemeister, H., Nichols, J., et al. (2012). The Transcriptional and Epigenomic Foundations of Ground State Pluripotency. *Cell*, *149*(3), 590–604. doi:10.1016/j.cell.2012.03.026

Maumus, F., Rabinowicz, P., Bowler, C., Rivarola, M., Inserm, U., Nacional, I., Agropecuaria, D. T., et al. (2011). Stemming Epigenetics in Marine Stramenopiles. *Current*.

Meissner, A. (2010). Epigenetic modifications in pluripotent and differentiated cells. *Nature biotechnology*, *28*(10), 1079–1088. doi:10.1038/nbt1684

Millar, C. B., & Grunstein, M. (2006). Genome-wide patterns of histone modifications in yeast. *Nature reviews. Molecular cell biology*, *7*(9), 657–66. doi:10.1038/nrm1986

Mos, L. (2001). Domoic acid: a fascinating marine toxin. *Environmental toxicology and pharmacology*, *9*(3), 79–85. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/11167152

Roudier, F., Ahmed, I., Bérard, C., Sarazin, A., Mary-Huard, T., Cortijo, S., Bouyer, D., et al. (2011). Integrative epigenomic mapping defines four main chromatin states in Arabidopsis. *The EMBO journal*, *30*(10), 1928–38. doi:10.1038/emboj.2011.103

Schmitz, R. J., & Ecker, J. R. (2012). Epigenetic and epigenomic variation in Arabidopsis thaliana. *Trends in plant science*, *17*(3), 149–154. doi:10.1016/j.tplants.2012.01.001

Significance, S., & Histories, D. L. (1997). Ecological, evolutionary, and systematic significance. *North*, *918*, 897–918.

Stroud, H., Otero, S., Desvoyes, B., Ramírez-parra, E., & Jacobsen, S. E. (2012). variants in Arabidopsis thaliana. doi:10.1073/pnas.1203145109/-/DCSupplemental.www.pnas.org/cgi/doi/10.1073/pnas.1203145109

Whitelaw, N. C., & Whitelaw, E. (2006). How lifetimes shape epigenotype within and across generations. *Human molecular genetics*, *15 Spec No*(2), R131–7. doi:10.1093/hmg/ddl200

Xie W, Kang Y, Gao Y (2006). Review on the life history of diatom resting spores. Marine Sciences 30(9), 75-78

Zemach, A., McDaniel, I. E., Silva, P., & Zilberman, D. (2010a). Genome-wide evolutionary analysis of eukaryotic DNA methylation. *Science (New York, N.Y.)*, *328*(5980), 916–9. doi:10.1126/science.1186366

Zemach, A., McDaniel, I. E., Silva, P., & Zilberman, D. (2010b). Genome-wide evolutionary analysis of eukaryotic DNA methylation. *Science (New York, N.Y.)*, *328*(5980), 916–9. doi:10.1126/science.1186366

Abstract:

Diatoms are the most successful, diverse and abundant eukaryotic phytoplankton. They play a vital role in the global ocean primary productivity and are responsible for greater than 40% of carbon sequestration in the ocean sustaining life on our planet. . The whole genome sequence of three diatom species, *Thalassiosira pseudonana*, *Phaeodactylum tricornutum* and *Fragilariopsis cylindrus*) reveals a wealth of information on their genes and how their genome is structured. However, the plasticity of diatoms and the adaptation to different environments implicate not only DNA sequence based regulation but also more reversible and flexible epigenetic changes. DNA methylation and histone modifications are the two main components of the epigenetic code. In my thesis, genome wide distribution of DNA methylation and histone modifications, two major components of epigenetic were investigated in the model pennate diatom *P. tricornutum*.

DNA methylation is the most extensively studied and widely conserved epigenetic mark. Here the first whole genome methylome from a stramenopile, the marine model diatom *P. tricornutum* is reported. In *P. tricornutum*, around 6% of the genome was methylated in a mosaic landscape. Extensive methylation in transposable elements (TEs), especially in recently amplified Copia-like elements was found. Over 320 genes were found methylated occurring in three different genomic contexts: in the proximity of TEs, in clusters of methylated genes, and in single genes. Furthermore, genes extensively and completely methylated correlated strongly with transcriptional silencing and differential expression under specific conditions. Finally, it was found that genes likely acquired by horizontal gene transfer from bacteria were preferentially inserted within TE-rich regions, suggesting a mechanism whereby the expression of foreign genes can be buffered following their insertion in the genome. In general, *P. tricornutum* has low DNA methylation with relatively extensive DNA methylation on TEs and a few methylated genes. This first Stramenopile methylome adds significantly to our understanding of the evolution of DNA methylation in eukaryotes.

As for the histone modifications, genome wide distribution of H3K4me2, H3K9me2 and H3K27me3 were examined in *P. tricornutum*. H3K4me2 is mainly associated with genes while both H3K9me2 and H3K27me3 marks target mainly transposable elements (TEs). The distribution of H3K27me3 is unusual and different from what have been profiled in model species so far. The genes marked by H3K27me3 tend to be lowly and differentially expressed. H3K27me3 and H3K9me2 tend to co-mark not only methylated TEs but also heavily

methylated genes, which appears to be important for maintaining the silencing of differentially expressed genes. The combinatorial analysis of different histone marks and DNA methylation gave us an overview of diatom chromatin landscapes, and will help to define conserved structural and functional features.

Taken together, the work presented here on *P. tricornutum* epigenome significantly contributes to our understanding of the organization of chromatin state of a diatom model species and establishes a foundation for future studies on epigenetic mechanisms underlying diatom adaptation to environmental changes and their ecological success.

Résumé

Les diatomées constituent le groupe d'eucaryotes photosynthétiques le plus diversifié et le plus important. Elles contribuent à environ 40% de la production primaire marine, et produisent presque 1/4 de l'oxygène que nous respirons jouant ainsi un rôle très important dans le maintien de la vie sur notre planète. Le séquençage du génome entier de trois espèces de diatomées, *Thalassiosira pseudonana*, *Phaeodactylum tricornutum* et *Fragilariopsis cylindrus* révèle une mine d'informations sur leurs gènes et comment leur génome est structuré. Toutefois, la plasticité des diatomées et l'adaptation à différents environnements impliquent non seulement la régulation basée sur la séquence d'ADN, mais aussi des changements de nature plus souple et réversibles dits épigénétiques. La méthylation de l'ADN et les modifications des histones sont les deux principales composantes du code épigénétique. Dans ma thèse, la cartographie à l'échelle du génome de la méthylation de l'ADN et des modifications des histones, deux composants majeurs de l'épigénétique ont été étudiés chez la diatomée modèle pennée *P. tricornutum.*

La méthylation de l'ADN est l'une des marques épigénétiques les plus étudiées et est largement conservée. Mes travaux de thèse présentent le premier méthylome d'une diatomée marine *P. tricornutum* qui appartient à la famille des Stramenopiles. *P. tricornutum* présente une méthylation d'environ 6% qui est présente en mosaïque sur l'ensemble du génome. Une méthylation importante a été retrouvé chez les éléments transposables, en particulier les éléments amplifiés récemment de type Copia. L'analyse met en évidence plus de 320 gènes méthylés dans trois contextes génomiques différents : à proximité des éléments transposables, en grappes de gènes méthylés, et dans des gènes uniques. En outre, les gènes largement et complètement méthylés ont été trouvé fortement corrélés avec le silencing transcriptionnel et l'expression différentielle dans des conditions spécifiques. Enfin, il a été constaté que les gènes susceptibles d'avoir été acquis par transfert horizontal de gènes bactériens étaient préférentiellement insérés dans des régions riches en éléments transposables, ce qui suggère un mécanisme par lequel l'expression de gènes étrangers peut être tamponnée à la suite de leur insertion dans le génome. En résumé, *P. tricornutum* a une faible méthylation de l'ADN et une methylation relativement importante des éléments transposables et seulement quelques gènes méthylés. Ce premier méthylome d'une diatomée Stramenopile ajoute de manière significative à notre compréhension de l'évolution de la méthylation de l'ADN chez les eucaryotes. En ce qui concerne les modifications des histones, la distribution des marques

H3K4me2, H3K9me2 et H3K27me3 a été examinée chez *P. tricornutum*. H3K4me2 est principalement associée à des gènes alors que les deux marques H3K9me2 et H3K27me3 ciblent principalement des éléments transposables. La répartition de H3K27me3 est inhabituelle et différente de ce qui a été observé chez les espèces modèles étudiées à ce jour. Les gènes marqués par H3K27me3 ont tendance à être faiblement exprimés et de façon différentielle. H3K27me3 et H3K9me2 ont tendance à co-marquer non seulement les éléments transposables méthylés, mais aussi des gènes fortement méthylés, ce qui semble être important pour le maintien du silencing des gènes différentiellement exprimés. L'analyse combinatoire de différentes marques d'histones et la méthylation de l'ADN nous a donné un aperçu du paysage de la chromatine chez les diatomées, et aidera à définir les caractéristiques structurales et fonctionnelles conservées. Considéré dans son ensemble, le travail présenté ici sur l'épigénome de *P. tricornutum* contribue de manière significative à notre compréhension de l'organisation de la chromatine chez une espèce modèle des diatomées et établit une fondation pour les études futures sur les mécanismes épigénétiques sous-jacents à l'adaptation des diatomées aux changements environnementaux et leur succès écologique.