Genetics

Volume 2, Issue 3, 173-191.
DOI: 10.3934/genet.2015.3.173
Received date 19 May 2015,
Accepted date 13 July 2015,
Published date 20 July 2015

http://www.aimspress.com/

Review

Probing the evolutionary history of epigenetic mechanisms: what can we learn from marine diatoms

Achal Rastogi[1], Xin Lin[1,2], Bérangère Lombard[3], Damarys Loew[3] and Leïla Tirichine[1,*]

[1] Ecology and Evolutionary Biology Section, Institut de Biologie de l'École Normale Supérieure (IBENS), CNRS UMR8197 INSERM U1024, 46 rue d'Ulm 75005 Paris, France
[2] State Key Laboratory of Marine Environmental Science, Xiamen University, Xiamen 361005, China
[3] Institut Curie, PSL Research University, Centre de Recherche, Laboratoire de Spectrométrie de Masse Protéomique, 26 rue d'Ulm 75248 Cedex 05 Paris, France

* **Correspondence:** Email: tirichin@biologie.ens.fr; Tel: 33 1 44 32 35 34.

Abstract: Recent progress made on epigenetic studies revealed the conservation of epigenetic features in deep diverse branching species including Stramenopiles, plants and animals. This suggests their fundamental role in shaping species genomes across different evolutionary time scales. Diatoms are a highly successful and diverse group of phytoplankton with a fossil record of about 190 million years ago. They are distantly related from other super-groups of Eukaryotes and have retained some of the epigenetic features found in mammals and plants suggesting their ancient origin. *Phaeodactylum tricornutum* and *Thalassiosira pseudonana*, pennate and centric diatoms, respectively, emerged as model species to address questions on the evolution of epigenetic phenomena such as what has been lost, retained or has evolved in contemporary species. In the present work, we will discuss how the study of non-model or emerging model organisms, such as diatoms, helps understand the evolutionary history of epigenetic mechanisms with a particular focus on DNA methylation and histone modifications.

Keywords: diatoms; *Phaeodactylum tricornutum*; *Thalassiosira pseudonana*; epigenetics; DNA methylation; histone modifications; non-coding RNA; comparative epigenetics; evolution

1. Introduction

Research in the field of epigenetics has taken off in the last decade as evidenced by the growing

number of published literature and scientific meetings. This is obviously due to numerous findings of its critical role in diseases such as cancer, development and responses to environmental cues in a wide range of species. Epigenetics means in addition to or above genetics implying changes in gene expression without altering the DNA sequence. These changes are inherited from cell to cell and trans-generationally from parent to offspring. Such changes involve chemical modifications of the DNA such as methylation, histone post-translational modifications leading to chromatin modifications, remodeling and attachment to the nuclear matrix, packaging of DNA around nucleosomes and RNA mediated gene silencing. Epigenetic mediated modifications are usually influenced by environmental cues, including diet, physical stresses such as temperature, or chemicals such as toxins and can also be stochastic due to random effects. A striking example is seen in Agouti mice exposed to bisphenol A, a ubiquitous chemical found in our environment. These are genetically identical twins but have a different size and fur color. In slim healthy brown mice, Agouti gene is prevented from transcription by DNA methylation while in yellow obese mice which are prone to diabetes and cancer, the same gene is not methylated resulting in its expression [1,2]. This is a fine example of the trans-generational inheritance of an epigenetic state where the Agouti locus escaped the usual resetting of epigenetic states during reproduction.

In the fruit fly *Drosophila melanogaster*, temperature treatment changes the eye color from white to red, and the treated individual flies pass on the change to their offspring over several generations without further requirement of temperature treatment [3]. The DNA sequence of the gene responsible for eye color remained the same for white eyed parents and red eyed offspring and the change was attributed to a specific histone modification [3]. Consistent with the work described above, a more recent study in *Drosophila* showed that the fission yeast homolog of activation transcription factor 2 (ATF2) that usually contributes to heterochromatin formation becomes phosphorylated leading to its release from heterochromatin upon heat shock or osmotic stress [4]. This new heterochromatin state that does not involve any DNA sequence change is transmitted over multiple generations [4].

In an ecological context, variation of DNA methylation was observed in a wild population of *Viola cazorlensis* which is a perennial plant [5]. Using a modeling approach on data collected over many years, the authors have 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 [5].

The above-discussed examples show a remarkable conservation of the function of epigenetic mechanisms in regulating gene expression among mammals, plants and invertebrates. This conservation goes beyond these species including early diverging single celled organisms such as microalgae. In this work, we will discuss how the study of non-model or emerging model organisms such as diatoms helps understand the evolutionary history of epigenetic mechanisms with a particular focus on DNA methylation and histone modifications.

2. Diatoms, what are they?

Diatoms are photosynthetic eukaryotic algae with cell sizes that usually range between 10 and 200 μm. They are found in all aquatic habitats including fresh and marine waters. These single celled species belong to Stramenopiles, which are part of the supergroup, Chromalveolates, containing also

the Alveolata, the Haptophyta and Cryptophyceae (Figure 1, [6,7]). Diatoms are one of the most diverse and widespread phytoplankton with more than 100,000 extant species which are divided into two orders: centric that are round with radial symmetry and pennate that are elongate with bilateral symmetry (Figure 2). Fossil evidence suggests that diatoms originated during or before the early Jurassic period (~ 210–144 Mya). They are hypothesized to be derived from successive endosymbiosis where a heterotrophic eukaryotic host engulfed cells, phylogenetically close to red and green alga [8], combining therefore features from both green and red algae predecessors [9]. The diversity of diatoms increased further via the horizontal transfer of bacterial genes [10]. Diatoms and bacteria have indeed co-occurred in common habitats throughout the oceans for more than 200 million years, fostering interactions between these two diverse groups over evolutionary time scales [11]. Diatoms are at the base of the food web contributing to one fifth of the planet's oxygen and representing 40% of primary marine productivity [12]. They therefore play a critical role sustaining life not only in the oceans but also on Earth as a whole through their role in the global carbon cycle. Diatoms are also important for human society, providing food through the aquatic food chain and high value compounds for cosmetic, pharmaceutical and industrial applications.

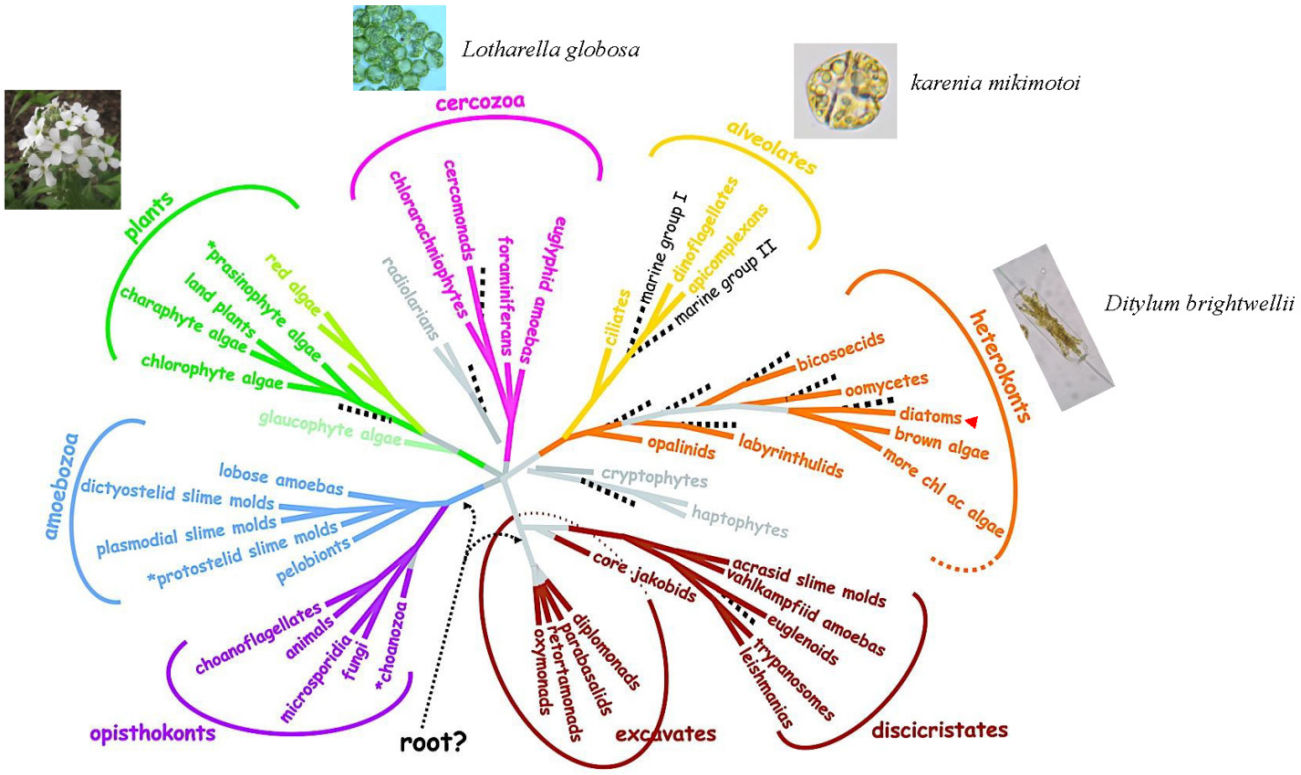

Figure 1. Eukaryote phylogenetic tree. The tree is derived from different molecular phylogenetic and ultrastructural studies (adapted from [13]). Images courtesy of NCMA, the Culture Collection of Marine Phytoplankton at Bigelow Laboratory for Ocean Sciences, and for dinoflagellates (image courtesy of Richard Dorrell). Red arrow head points to diatoms.

Several diatom genome sequences are now available including the two centrics, *Thalassiosira*

pseudonana (32 Mbp), (http://genome.jgi-psf.org/Thaps3/Thaps3.home.html) [14] and *Thalassiosira oceanica* (81.6 Mbp, [15]) *Phaeodactylum tricornutum* (27 Mbp) [10], (http://genome.jgi-psf.org/Phatr2/Phatr2.home.html) (Figure 2), the polar, cold-loving species *Fragilariopsis cylindrus* (80 Mbp; http://genome.jgi-psf.org/Fracy1/Fracy1.home.html), the toxigenic coastal species *Pseudo-nitzschia multiseries* (300 Mb; by the Joint Genome Institute) and the high lipid content diatom *Fistulifera* sp. strain JPCC DA058 [16]. The ecological success of diatoms suggests that they have developed sophisticated ways to cope with changing environments. Complete sequencing of *P. tricornutum* genome [17] showed that it has an unusual genetic composition, which arose through successive endosymbioses and horizontal gene transfers from bacteria. These events have provided diatoms with several unusual metabolic pathways, such as the urea cycle which was previously considered to exist only in animals [14,18]. The ability of diatoms to survive in rapidly changing environments with all the fluctuating conditions (UV radiations, temperature, salinity, toxins, nutrients, grazing pressure etc.) is also attributable to another layer of regulation known as epigenetics [19]. It was previously shown using McrBC, an enzyme sensitive to methylated DNA, that there is an induction of LTR-retrotransposon called Blackbeard (*Bkb*) with a decrease in cytosine methylation under nitrate limitation suggesting that nitrate depletion induces demethylation and upregulation of *Bkb* [20]. Although de novo insertion of *Bkb* was not shown in this study, its distribution with two other retrotransposons was analyzed in thirteen different accessions of *P. tricornutum.* The work showed clear differences in the distribution of the three retrotransposons among the tested accessions demonstrating their transposition in natural environments [20]. Besides this experimental clue of the occurrence of DNA methylation, our recent work [21-24] revealed an amazing conservation of the epigenetic machinery in this model diatom. *P. tricornutum* possesses histone modifying enzymes, small RNA [25,26] as well as DNA methylation which is absent in the multicellular brown algae *Ectocarpus siliculosus* which belongs to Stramenopiles [27].

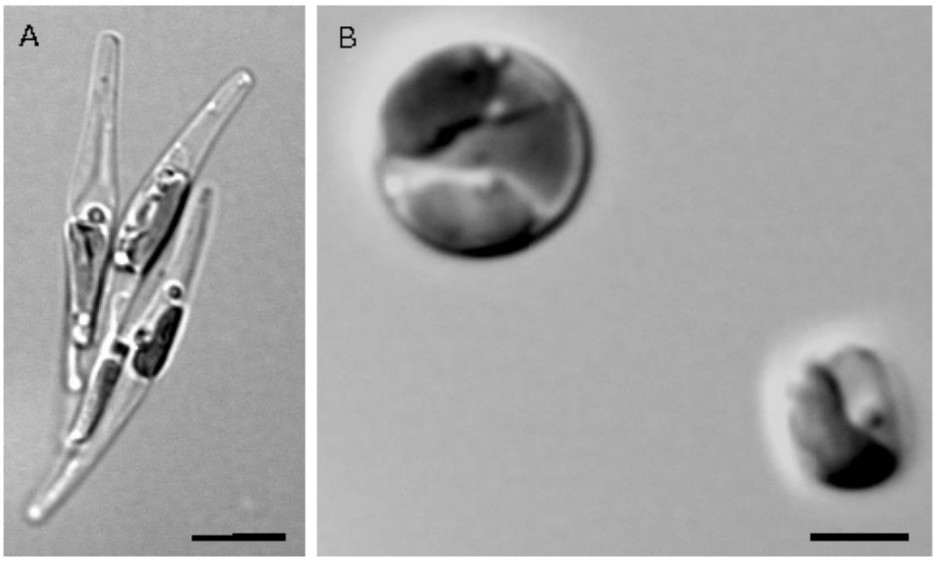

Figure 2. Light microscopy micrographs of representative model diatoms. A. The pennate diatom *Phaeodactylum tricornutum*. Scale bar 5 μm. B. The centric diatom *Thalassiosira pseudonana*. Scale bar 2 μm.

3. DNA methylation

Cytosine 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 five (5meC) common to all three super kingdoms. Cytosine methylation 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 transposons and regulation of gene expression [28,29]. DNA methylation is widespread among protists, plants, fungi and animals [30,31]. 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 algae *Ectocarpus siliculosus* [27,32].

With the advent of sequencing technologies and their increasing quality in terms of resolution and depth, our view and understanding of DNA methylation in the main supergroups of eukaryotes, plants and animals starts to emerge. The recently published methylome of *P. tricornutum* [23], which is phylogenetically distant from classic model organisms in the animal and green plant groups as well as diverse protists [31,33], drew a better picture and brought more insights into the evolutionary history of DNA methylation. With 27 Mb genome size, *P. tricornutum* shows a low level of DNA methylation compared to other eukaryotes such as human, *Arabidopsis* and the sea squirt *Ciona intestinalis* [31,33,34] (Figure 3). This is not correlated to the size of the genome as evidenced by the higher methylation occurrence of *Ostreoccocus* [33] that have much smaller genome and the low methylation in honey bee [31] whose genome is nearly ten times bigger than *P. tricornutum*. Although few species are compared in Figure 3, increase in cytosine DNA methylation seems to correlate with the average content of transposable elements, which presumably are kept silenced, and the complexity of the genome. Comparative epigenomics or methylomics provide some insights into the genes that might have impacted species evolutionary fate. A striking example are the differentially methylated genic regions (DMRs) found in human and its closely related primates such as chimpanzees, gorillas and orangutans which encode neurological functions suggesting species divergence correlated with developmental specialization [35,36]. In line with these observations, comparative epigenetic analysis of the two diatoms, the pennate *P. tricornutum* and the centric *T. pseudonana* [33], revealed no major differences in the fraction of the genome that is methylated or the context (Figure 4). However, out of 6199 shared genes, 408 are methylated only in *P. tricornutum* versus 461 only in *T. pseudonana*. DMRs between the two species are subsequently reflected in different GO categories enrichment [33] (Figure S1). Investigating further these genes might shed light on the history of their evolutionary divergence.

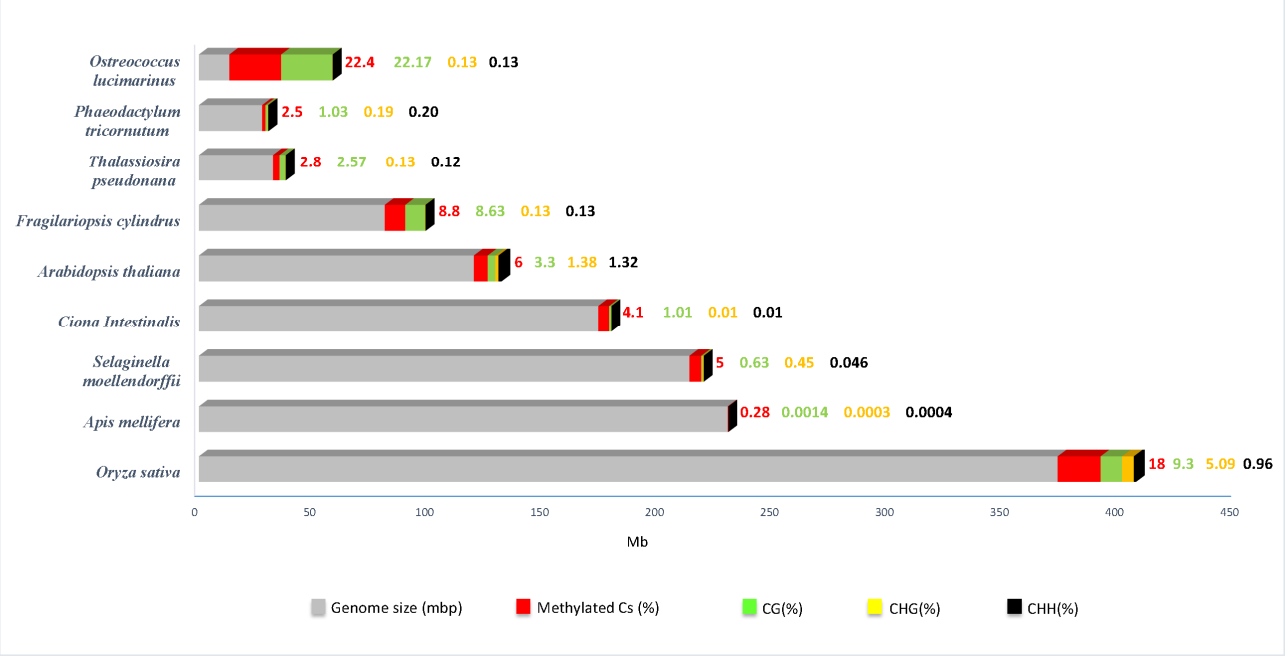

Figure 3. DNA methylation in diverse Eukaryotes. Graphical representation of genome-wide percentages of cytosine DNA methylation as well as in different contexts [C (red), CG (green), CHG (orange) and CHH (black)]. Species names are represented on the Y-axis. All the stated elements are represented as stacks over gray bar indicating the size of each genome measured as mega base pairs (Mbp). For comparison, the human genome methylation data is given: genome size (3381,94), methylated Cs (75%). Data was taken from [33,37,38], http://genome.jgi-psf.org/, http://phytozome.jgi.doe.gov/pz/portal.html.

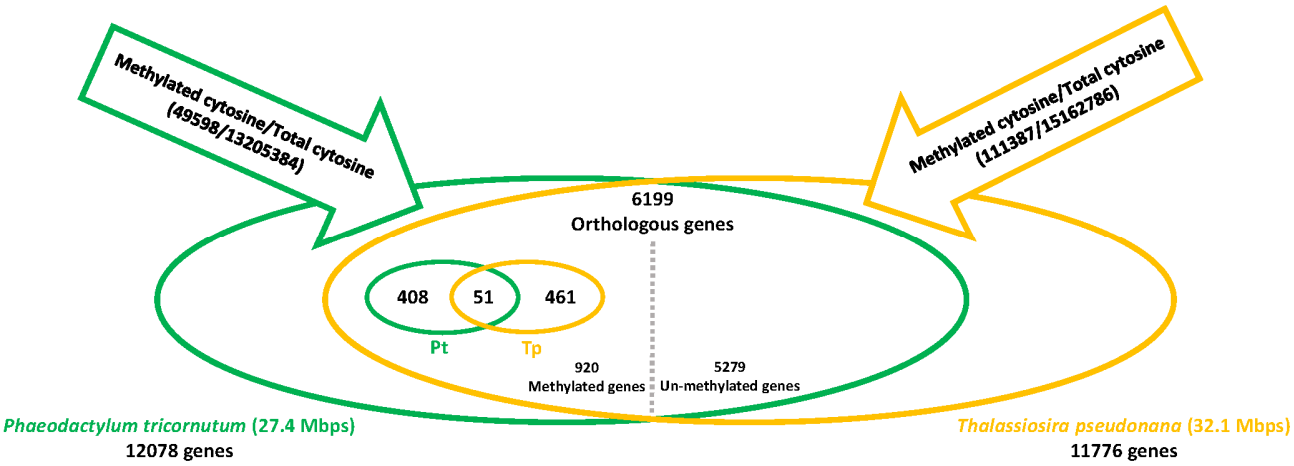

Figure 4. The orthologous gene body cytosine methylation analysis. The genes that are differentially methylated between *Phaeodactylum tricornutum* (Pt) and *Thalassiosira pseudonana* (Tp) are represented. Qualitative analysis of gene body cytosine methylation

on the orthologous genes between Pt and Tp genome. Using reciprocal best-hit BLAST approach, orthologous genes between Pt and Tp genomes are found. Out of 6199 orthologues, 459 genes are methylated in Pt whereas 512 genes are found methylated in Tp genome. The Venn comparison of these genes shows the conservation of gene body cytosine methylation over 51 genes while 408 and 461 genes are specifically methylated in Pt and Tp genomes, respectively. SRA accessions: Tp = GSM1134628; Pt = GSM1134626.

DNA methylation can occur in different contexts including CG, CHG and CHH where H can be any nucleotide except G. In *P. tricornutum,* DNA methylation was found in all contexts suggesting that CHG and CHH is not a plant innovation but existed already in a common ancestor and was lost from certain lineages. Indeed, Eukaryotes have evolved and/or retained different DNA methyltransferase complements responsible for the different context of methylation. Metazoans commonly encode DNMT1 and DNMT3 proteins, while higher plants additionally have plant-specific chromomethylase (CMT). On the other hand, fungi have DNMT1, Dim-2, DNMT4, and DNMT5 [39,40]. Previous phylogenetic analysis suggests that *P. tricornutum* genome encodes a peculiar set of DNMTs as compared to other eukaryotes [41]. DNMT1 appears to be absent in *P. tricornutum* as well as putative proteins coding for plant specific DNA methyltransferase CMT3 and DRM, which are responsible for non CG methylation. *P. tricornutum* encodes DNMT2 (Pt16674), which is an RNA methyltransferase that shows strong sequence similarities with DNA cytosine C5 methyltransferases. In addition to DNMT3 (Pt 46156), diatom genomes also encode DNMT5 (Pt45072) and DNMT6 (Pt36049) proteins as well as a bacterial-like DNMT (Pt47357) [41]. 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 *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 [42]. Conservation of this gene in diatoms over this length of time suggests that it is functional. 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 [39,40]. 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 the absence of DNA methyltransferases in the stramenopile *E. siliculosus* [27]. *P. tricornutum* encodes three putative DNA demethylases (Pt46865, Pt48620, Pt12645) with ENDO domain similar to the *Arabidopsis* DNA demethylases ROS1 domain suggesting similar mechanisms for DNA demethylation.

Dnmt5 was reported in a wide range of Eukaryotic single celled species that lack Dnmt1 but nevertheless retain CG methylation which was shown to be catalyzed by Dnmt5 [33]. In this work, the authors used *Cryptococcus neoformans* that has Dnmt5 as a unique DNA methylatransferase and showed that CG methylation is entirely lost when DNMT5 is deleted [33]. However, the authors did not exclude that another unknown methyltransferase catalyzes CG methylation and uses Dnmt5 as a required accessory or regulatory protein [33]. As mentioned above, typical Dnmt1 does not exist in *P. tricornutum* but our in-silico analysis revealed the presence of a gene which 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. In higher Eukaryotes, Dnmt1 is the enzyme that catalyzes CG methylation and the activity of its catalytic domain is regulated by the N terminal region of the protein. Indeed an isolated Dnmt1 catalytic domain was proven to be inactive [43,44]. Interestingly, both BAH and cysteine rich domains are found within the N terminal region of Dnmt1 in higher eukaryotes. A tempting hypothesis would be that *P. tricornutum* Dnmt1-like is the accessory protein that might interact with Dnmt5 to catalyze CG methylation. It is tempting to think that these two domains that are as independent proteins in *P. tricornutum* fused through evolutionary time in a single polypeptide protein in higher Eukaryotes and gave rise to the eukaryotic Dnmt1. We are currently using a reverse genetic approach to determine the function of Dnmts and the putative accessory protein in *P. tricornutum.* The work will help to better understand their role in processes such as maintenance and *de novo* DNA methylation as well as context specificities which will ultimately shed light on the function of DNMTs in an evolutionary context.

P. tricornutum methylome discussed in various studies [23,30,31] confirms the conservation of gene body methylation as an ancient feature and its methylation preference for exons over introns in all Eukaryotic genomes where it has been examined including *Arabidopsis*, *Ciona intestinalis*, honey-bee and human. Several hypotheses were made to explain this specific pattern and interestingly, in-silico analysis of *P. tricornutum* genome revealed few evidences that support them. *P. tricornutum* encodes ROS1 related glycolysases that were thought present only in *Arabidopsis* where they were shown to specifically remove DNA methylation from gene ends [45]. A more universal factor that might explain gene body methylation pattern is the histone mark H3K4me that antagonizes DNA methylation and is distributed around the transcription start site in the genomes where it has been examined. In *P. tricornutum*, H3K4me2 does not localize with DNA methylation and maps around the translation start site [24], which is in line with its potential contribution to DNA methylation pattern at gene bodies.

A conserved function for gene-body methylation at the whole-genome level has not yet been established. When examined, sets of body-methylated genes were found to be expressed constitutively at moderate levels such as in angiosperms and most invertebrates [34,46-48]. Nevertheless, in the silkworm, gene-body methylation correlates positively with gene expression levels [49]. In human, gene body methylation was shown to be involved in X chromosome activation [50] while it was recently reported that methylation of the first exon of autosomal genes correlates with transcriptional silencing [51]. It was also proposed that gene body methylation in human regulates the activity of intragenic alternative promoters [52]. In this line, a recent study [53] has established that body-methylated genes in *A. thaliana* are functionally more important, as measured by phenotypic effects of insertional mutants, than unmethylated genes. Using a probabilistic approach, the authors have reanalyzed single-base resolution bisulfite sequence data from *A. thaliana*. They demonstrated that body methylated genes are likely involved in either suppressing expression from cryptic promoters within coding regions and/or in enhancing accurate splicing of primary transcripts [53]. Interestingly, these functions were already proposed by previous studies [54-56], and the recent comparative study of honey-bee methylome has also established a link between gene-body methylation and splicing [57]. In our study, we found that gene-body methylation in *P. tricornutum* correlates positively with gene length and exon number. It is thus tempting to infer that intragenic methylation in *P. tricornutum* may play a role in avoiding aberrant transcription and/or mis-splicing. Furthermore, functional annotation of body-methylated genes reveals the presence of important

functional classes such as (1) transferases and catalytic enzymes that play important role in cell wall assembly and its rearrangement which is crucial for cell integrity, (2) hydrolase activity which is important in stress responses, and (3) transporter activity necessary for metabolites shuttling such as silicic acid. Considering previous studies and in light of our recent work in *P. tricornutum*, gene body methylation does not suppress expression but rather correlates with low to moderate transcriptional activity. This might have the putative function of preventing aberrant transcription from intragenic promoters and appears to be a common and ancestral eukaryotic feature as reported previously [31,54].

4. Histones and their modifications

Eukaryotic chromosomes are packaged in the nucleus by wrapping the DNA around an octamer of four core histone proteins H2A, H2B, H3 and H4 forming the basic unit of chromatin, the nucleosome. Further compaction is achieved by the interaction of the nucleosome to the linker histone H1. This phenomenon seems to be conserved among all Eukaryotes and even archaea, where the nucleosomes are formed of only a tetramer of two H3 and H4 histones found in the cell, as archaea do not have a nucleus. Furthermore, nucleosome occupancy was found similar in two species of Archaea with depletion over transcriptional start sites as well as a conservation of nucleosome positioning code [58,59]. This demonstration of similarities between Eukaryotes and Archaea chromatin, suggests that histones and chromatin architecture evolved before the divergence of Archaea and Eukarya. This also suggests that the initial function of nucleosomes and chromatin formation might have been for the regulation of gene expression rather than the packaging of DNA, which is an Eukaryotic invention [58].

Histones are subject to a variety of post-translational modifications (PTMs) that have an important role in several processes such as transcription, replication and DNA repair. Histone PTMs in particular at the N terminus include acetylation, methylation, phosphorylation and ubiquitination, which were extensively studied in diverse species, along with modifications like sumoylation, glycosylation, biotinylation, carbonylation, and ADP ribosylation for which little is known [60]. Histone PTMs function either by altering the accessibility of genes to the transcriptional machinery, or by binding to effector proteins via specialized chromatin domains that deposit or erase these histone modifications. PTMs function in a combinatorial pattern known as the histone code, which confers active or repressive chromatin states to specific chromosomal regions of the genome [60,61].

P. tricornutum possesses 14 histone genes encoding 9 histone proteins. They are dispersed throughout five chromosomes with most in clusters of two to six genes as seen for most Eukaryotes. *P. tricornutum* histones belong to the five known classes, histone H1, H3, H4, H2A and H2B. These histones are conserved among diatoms and eukaryotic species. With the exception of histones H4 and H2B, *P. tricornutum* encodes variants for each histone H1, H3 and H2A. Sequence alignment of histone H3 shows the presence of canonical and replacement histones similar to human, H3.2 and H3.3. Additionally, *P. tricornutum* expresses a centromere specific variant commonly called CenH3 that varies considerably from the rest of H3 histones especially in the N terminal tail. CenH3 is essential for recruitment of kinetochores components ensuring correct segregation of chromosomes during mitosis and meiosis [62].

H2A histone members constitute the most diverse group of histones with the greatest number of variants. *P. tricornutum* is no exception as it encodes two copies of the canonical H2A but also both

H2AZ (Pt28445) and H2AX variants while this latter is missing from *C. elegans* and protozoan parasites such as *Plasmodium* and *Trypanosomes*. The presence of the conserved motif SQE/D in the C terminal of *P. tricornutum* H2AX suggests a putative role of this histone in the maintenance of genome integrity via its contribution in the repair of double stranded DNA breaks. *P. tricornutum* encodes two histone H1 variants, which share nearly 50% identity. Interestingly, one of them (Pt44318), is expressed only in stress conditions such as high light which suggests its putative role in DNA repair as found previously in yeast and vertebrates [63,64]. The diversity of histone variants in *P. tricornutum* is interesting and suggests an adaptive evolution to the life history of diatoms via their chromatin interface to acquire new abilities to cope with the changing environment.

P. tricornutum and *T. pseudonana* genome sequencing revealed a long list of histone modifying and demodifying enzymes that are summarized in Table 1. This shows the great conservation of the writers and erasers of histone modification marks in diatoms and their ancient origin. Furthermore, Mass spectrometry analysis (MS) of PTMs in *P. tricornutum* showed similarities to that of plants and mammals including acetylation and/or methylation of several lysines on the N terminal tail of histones H2A, H2B, H3 and H4 and mono, di and tri-methylation of lysines 4, 9, 27 and 36 of histone H3 suggesting the early divergence of these PTMs and their important role in transcriptional regulation of many biological processes (Table 2). Interestingly, *P. tricornutum* combines histone PTMs found in both mammals and plants such as acetylation and mono-di methylation of lysine 79 of histone H3 found only in human and yeast [65] but not in *Arabidopsis* [66] underlying *P. tricornutum* genome diversity and the divergence of histone modifications among species throughout evolution. Another interesting example is the acetylation of lysine 20 of histone H4 which is shared with *Arabidopsis* but different from human where the residue is only methylated [66]. H4K20me which is known to be a repressive mark was detected neither by mass spectrometry nor by western blot using an antibody that recognizes this modification in *Arabidopsis* (data not shown). Furthermore, mono and dimethylation of lysine 79 of histone H4 are modifications that *P. tricornutum* shares only with *Toxoplasma gondii* which is an obligate intracellular parasitic protozoan belonging to Alveolates, a superphylum closely related to Stramenopiles [24]. A non-exhaustive mass spectrometry analysis of histones from an early diverging diatom *Thalassiosira pseudonana* shows the presence of similar histone PTMs (Figure 5), which points to the important role that histone PTMs might have had in shaping diatom genomes and ultimately in the diversification of eukaryotes.

Table 1. **Histone modifications enzymes in two diatom species**. Proteins encoding putative enzymes responsible for histone modification which are identified in *P. tricornutum* and *T. pseudonana.* New gene models are given for *P. tricornutum* (http://protists.ensembl.org/Phaeodactylum_tricornutum/Location/Genome).

Histone Modifiers	**Residues Modified**	**Homologs in *P. tricornutum* (Phatr2)**	**Homologs in *P. tricornutum* (Phatr3)**	**Homologs in *T. pseudonana***
Lysine Acetyltransferases (KATs)				
HAT1 (KAT1)	H4 (K5, K12)	54343	Phatr3_J54343	1397, 22580
GCN5 (KAT2)	H3 (K9, K14, K18, K23, K36)	46915	Phatr3_J2957	15161
Nejire (KAT3);	H3 (K14, K18,	45703, 45764,	Phatr3_J45703, Phatr3_J45764,	24331, 269496,

CBP/p300 (KAT3A/B)	K56) H4 (K5, K8); H2A (K5) H2B (K12, K15)	54505	Phatr3_J54505	263785
MYST1 (KAT8)	H4 (K16)	24733, 24393	Phatr3_J51406, Phatr3_J3062	37928, 36275
ELP3 (KAT9)	H3	50848	Phatr3_J50848	9040
Unknown				
RPD3 (Class I HDACS)	H2, H3, H4	51026, 49800	Phatr3_J51026, Phatr3_J49800	41025, 32098, 261393
HDA1 (Class II HDACS)	H2, H3, H4	45906, 50482, 35869	Phatr3_J45906, Phatr3_J50482, Phatr3_J35869	268655, 269060, 3235, 15819
NAD+ dependent (Class III HDACS)	H4 (K16)	52135, 45850, 24866, 45909, 52718, 21543, 39523	Phatr3_J52135, Phatr3_J45850, Phatr3_J8827, Phatr3_J12305, Phatr3_J16589, Phatr3_J21543, Phatr3_J39523	269475, 264809, 16405, 35693, 264494, 16384, 35956
Lysine Methyltransferases				
MLL	H3 (K4)	40183, 54436, 42693, 47328, 49473, 49476, 44935	Phatr3_EG00277, Phatr3_EG02316, Phatr3_J6915, Phatr3_J47328, Phatr3_EG00277, Phatr3_15913, Phatr3_J44935	35182, 35531, 22757
ASH1/WHSC1	H3 (K4)	43275	Phatr3_6093	264323
SETD1	H3 (K36), H4 (K20)	not found	not found	not found
SETD2	H3 (K36)	50375	Phatr3_EG02211	35510
SETDB1	H3 (K9)	not found	not found	not found
SETMAR	H3 (K4, K36)	not found	not found	not found
SMYD	H3 (K4)	bd1647, 43708	Phatr3_J1647, Phatr3_J43708	23831, 24988
TRX-related		not found	not found	not found
E(Z)	H3 (K9, K27)	32817	Phatr3_J6698	268872
EHMT2	H3 (K9, K27)	not found	not found	not found
SET+JmjC	Unknown	bd1647	Phatr3_J1647	not found
Lysine Demethylases (KDM)				
LSD1 (KDM1)	H3 (K4, K9)	51708, 44106, 48603	Phatr3_J51708, Phatr3_J44106, Phatr3_J48603	not found
FBXL (KDM2)	H3 (K36)	42595	Phatr3_J42595	not found
JMJD2 (KDM4)/JARID	H3 (K9, K36)	48747	Phatr3_J48747	2137
JMJ-MBT	Unknown	48109	Phatr3_J48109	22122
JMJ-CHROMO	Unknown	40322	Phatr3_J40322	1863

Table 2. Diversity of histone PTMs in *P. tricornutum*. Examples of PTMs of histones present in *P. tricornutum* but absent or not detected (ND) in representative of two major lineages, animals and plants. Data taken from [24,66,67].

Histone PTM	*P. tricornutum*	*H. sapiens*	*A. thaliana*
H4K31	present	ND	ND
H4K59Ac	present	ND	ND
H4K59me	present	ND	ND
H4K79me	present	ND	ND
H4K79me2	present	ND	ND
H4K20Ac	present	ND	present
H4K20me	present	present	ND
H3K79me	present	present	ND
H3K79me2	present	present	ND
H2BK107Ac	present	ND	ND

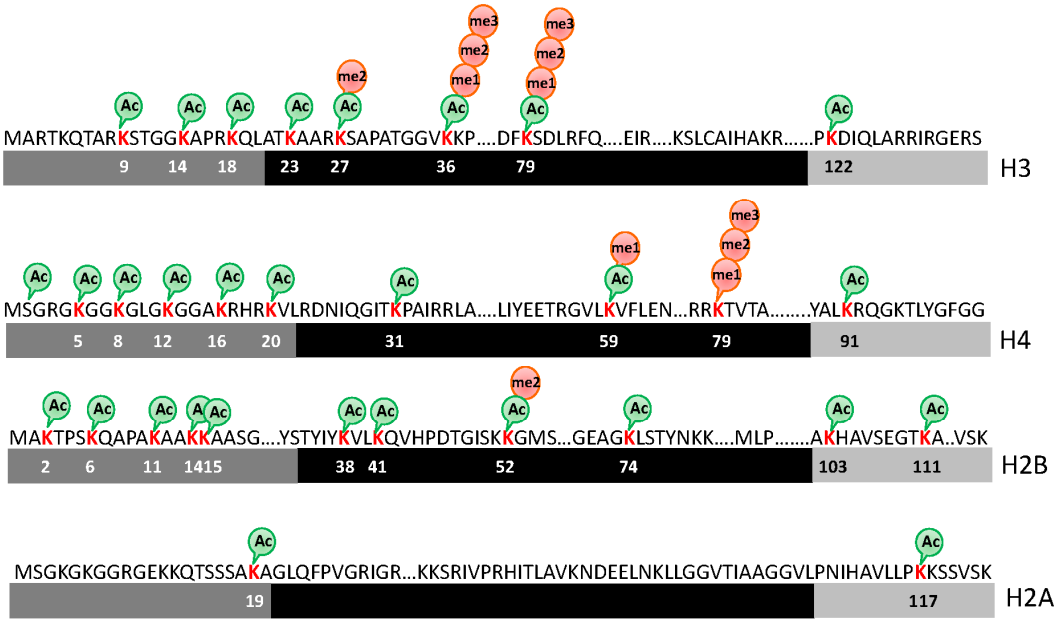

Figure 5. Histone PTMs in *T. pseudonana.* Diagram showing sites of PTMs of core and variant histones identified in *Thalassiosira pseudonana* by mass spectrometry. Amino acid residue number is indicated below the peptide sequence. Dark gray, black and light gray boxes indicate N-terminal, globular core and C-terminal domains, respectively. Acetylation and methylation are indicated in green and red respectively.

5. Non-coding RNA

Non-coding RNA is found in all kingdoms of life with fractions varying from 8% for bacteria to more than 98% for human genome (Figure 6). This non-coding fraction comprises functional non-coding RNAs such as transfer, ribosomal and regulatory RNAs as well as DNA that remains

untranscribed or gives rise to RNA molecules of unknown function. Genome size correlates positively with the amount of non-coding DNA and evolutionary age of the species suggesting that the smaller and early diverging the species are, the less non-coding fraction of their genome they have (Figure 6). This also suggests that non-coding RNAs arose with the complexity of species and the plethora of subsequent novel functions. Although initially argued to be spurious transcriptional noise or accumulated evolutionary debris arising from the early assembly of genes and/or the insertion of mobile genetic elements, we have now evidence suggesting that the previously named "junk DNA" may play a major biological role in cellular development, physiology and pathologies [68]. It is also argued that not all of it will be functional as the transcription machinery is not perfect and will generate non-coding RNA with no fitness advantage and simply tolerating them would be more feasible than evolving and maintaining more rigorous control mechanisms that could prevent their production [69]. Non-coding RNAs that appear to have an epigenetic function including heterochromatin formation, DNA methylation, histone modifications and transcriptional silencing can be divided into two main categories based on their length: short non-coding RNAs (< 30 nts) and long non-coding RNAs (> 200 nts). Short interfering RNAs (siRNA) of 21 nucleotides are produced by long double stranded RNA through a cleavage by the endonuclease Dicer and are bound by an Argonaute protein. They recognize and silence their target mRNAs by perfect sequence complementarity which is in contrast to micro RNAs (miRNAs, 20 to 23 nts) which silence their target sequences by incomplete homology and act primarily at the translational level. Long non-coding RNAs (lncRNAs) have been reported in several eukaryotic genomes including mouse [70], human [71], *Arabidopsis* [72] and Zebrafish [73].

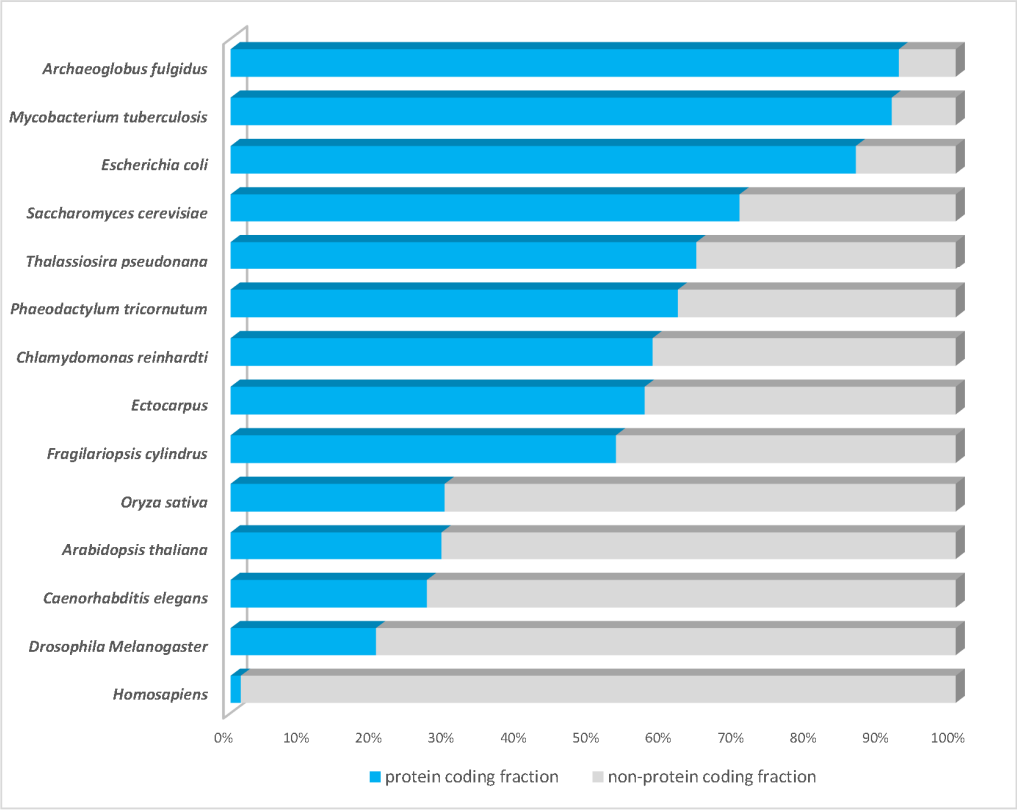

Figure 6. The percentage of coding fraction of several Eukaryotic and bacterial genomes (Adapted from [68]).

Non-coding RNAs are highly diverse and new classes are constantly being discovered. For an exhaustive list of known non-coding RNAs, refer to [74]. Non-coding RNA are known to occur in a wide range of species including human, insects, fish, plants, yeast, protists, even bacteria and archaea, underlying a conserved phenomenon. In *Chlamydomoans reinhardtii*, two studies reported the existence of miRNA that are reminiscent of the miRNAs of multicellular organisms as well as the phased transacting siRNAs (tasiRNAs) of plants. *Chlamydomonas* miRNA do not seem to have sequence homology to any known miRNAs in animals or plants, suggesting that miRNA genes may have evolved independently in the lineages leading to animals, plants and green algae [75,76]. The discovery of small RNA in diatoms and cocolithophores further confirmed the early divergence of such molecules [25,77,78].

6. Conclusions and future perspectives

Although epigenetics is recognized for its fundamental role in diseases such as cancer, there is still a long way to go before we appreciate its importance in shaping species genomes through evolutionary time scales. Epigenetics allows individuals and populations to cope with biotic and abiotic stresses and respond to environmental cues through its dynamic regulation of genes but also provides progenies with a better fitness when the parents experience a particular stress affecting therefore their evolutionary potential. This is exemplified by DNA methylation that acts as an inducer of mutations in DNA sequences via the deamination process impacting therefore genome nucleotide sequences. These mutations in chromosomal DNA might have an effect on the fitness and evolution of individuals and populations. Using model or non-model single celled eukaryotes such as diatoms which constitute an early diverging branch in the evolutionary tree will provide a solid complement to multicellular organisms to enhance our understanding of the impact and true contribution of epigenetics to biological processes and ultimately to their evolutionary history. It is becoming clear now that it is important to include epigenetics and its impact on the evolutionary biology of species in our way of thinking and designing of experiments in biology.

Acknowledgments

AR is a PhD student funded by the MEMO LIFE International PhD program.

Conflict of interest

All authors declare no conflicts of interest in this paper.

Supplementary

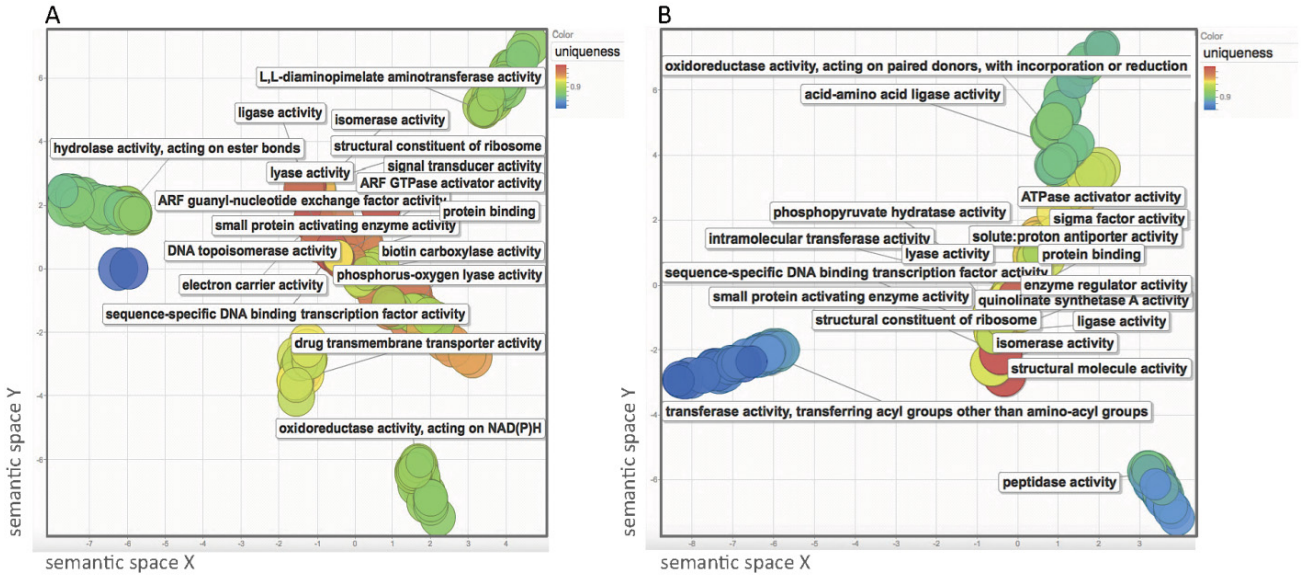

Figure S1. Gene Ontology (GO) enrichment analysis based on semantic clustering of molecular function (MF) associated to *P. tricornutum-T. pseudonana* orthologous genes which are (A) methylated only in *P. tricornutum* and (B) methylated in *T. pseudonana.* X and the Y axis represent the pairwise semantic similarity scores. Color in the sphere represents the uniqueness of each term when compared semantically to the whole list of molecular functions. More unique terms tends to be less dispensable. The graph was generated using Revigo [79].

References

1. Dolinoy DC (2008) The agouti mouse model: an epigenetic biosensor for nutritional and environmental alterations on the fetal epigenome. *Nutr Rev* 66 Suppl 1: S7-11.
2. Dolinoy DC (2007) Epigenetic gene regulation: early environmental exposures. *Pharmacogenomics* 8: 5-10.
3. Tariq M, Nussbaumer U, Chen Y, et al. (2009) Trithorax requires Hsp90 for maintenance of active chromatin at sites of gene expression. *Proc Natl Acad Sci U S A* 106: 1157-1162.
4. Seong KH, Li D, Shimizu H, et al. (2011) Inheritance of stress-induced, ATF-2-dependent epigenetic change. *Cell* 145: 1049-1061.
5. Herrera CM, Bazaga P (2011) Untangling individual variation in natural populations: ecological, genetic and epigenetic correlates of long-term inequality in herbivory. *Mol Ecol* 20: 1675-1688.
6. Dorrell RG, Smith AG (2011) Do red and green make brown?: perspectives on plastid acquisitions within chromalveolates. *Eukaryot Cell* 10: 856-868.
7. Walker G, Dorrell RG, Schlacht A, et al. (2011) Eukaryotic systematics: a user's guide for cell biologists and parasitologists. *Parasitology* 138: 1638-1663.
8. Archibald JM (2009) The puzzle of plastid evolution. *Curr Biol* 19: R81-88.

9. Moustafa A, Beszteri B, Maier UG, et al. (2009) Genomic footprints of a cryptic plastid endosymbiosis in diatoms. *Science* 324: 1724-1726.
10. Bowler C, Allen AE, Badger JH, et al. (2008) The Phaeodactylum genome reveals the evolutionary history of diatom genomes. *Nature* 456: 239-244.
11. Amin SA, Parker MS, Armbrust EV (2012) Interactions between diatoms and bacteria. *Microbiol Mol Biol Rev* 76: 667-684.
12. Falkowski PG, Barber RT, Smetacek VV (1998) Biogeochemical Controls and Feedbacks on Ocean Primary Production. *Science* 281: 200-207.
13. Baldauf SL (2008) An overview of the phylogeny and diversity of eukaryotes. *J Syst Evol* 46: 263-273.
14. Armbrust EV, Berges JA, Bowler C, et al. (2004) The genome of the diatom Thalassiosira pseudonana: ecology, evolution, and metabolism. *Science* 306: 79-86.
15. Lommer M, Specht M, Roy AS, et al. (2012) Genome and low-iron response of an oceanic diatom adapted to chronic iron limitation. *Genome Biol* 13: R66.
16. Tanaka T, Maeda Y, Veluchamy A, et al. (2015) Oil Accumulation by the Oleaginous Diatom Fistulifera solaris as Revealed by the Genome and Transcriptome. *Plant Cell* 27: 162-176.
17. Bowler C, De Martino A, Falciatore A (2010) Diatom cell division in an environmental context. *Curr Opin Plant Biol* 13: 623-630.
18. Allen AE, Dupont CL, Obornik M, et al. (2011) Evolution and metabolic significance of the urea cycle in photosynthetic diatoms. *Nature* 473: 203-207.
19. Tirichine L, Bowler C (2011) Decoding algal genomes: tracing back the history of photosynthetic life on Earth. *Plant J* 66: 45-57.
20. Maumus F, Allen AE, Mhiri C, et al. (2009) Potential impact of stress activated retrotransposons on genome evolution in a marine diatom. *BMC Genomics* 10: 624.
21. Lin X, Tirichine L, Bowler C (2012) Protocol: Chromatin immunoprecipitation (ChIP) methodology to investigate histone modifications in two model diatom species. *Plant Methods* 8: 48.
22. Tirichine L, Lin X, Thomas Y, et al. (2014) Histone extraction protocol from the two model diatoms Phaeodactylum tricornutum and Thalassiosira pseudonana. *Mar Genomics* 13: 21-25.
23. Veluchamy A, Lin X, Maumus F, et al. (2013) Insights into the role of DNA methylation in diatoms by genome-wide profiling in Phaeodactylum tricornutum. *Nat Commun* 4: 2091.
24. Veluchamy A, Rastogi A, Lin X, et al. (2015) An integrative analysis of post-translational histone modifications in the marine diatom Phaeodactylum tricornutum. *Genome Biol* 16: 102.
25. Rogato A, Richard H, Sarazin A, et al. (2014) The diversity of small non-coding RNAs in the diatom Phaeodactylum tricornutum. *BMC Genomics* 15: 698.
26. 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.
27. Cock JM, Sterck L, Rouze P, et al. (2010) The Ectocarpus genome and the independent evolution of multicellularity in brown algae. *Nature* 465: 617-621.
28. Suzuki MM, Bird A (2008) DNA methylation landscapes: provocative insights from epigenomics. *Nat Rev Genet* 9: 465-476.
29. Law JA, Jacobsen SE (2010) Establishing, maintaining and modifying DNA methylation patterns in plants and animals. *Nat Rev Genet* 11: 204-220.

30. Feng S, Cokus SJ, Zhang X, et al. (2010) Conservation and divergence of methylation patterning in plants and animals. *Proc Natl Acad Sci U S A* 107: 8689-8694.
31. Zemach A, McDaniel IE, Silva P, et al. (2010) Genome-wide evolutionary analysis of eukaryotic DNA methylation. *Science* 328: 916-919.
32. Goll MG, Bestor TH (2005) Eukaryotic cytosine methyltransferases. *Annu Rev Biochem* 74: 481-514.
33. Huff JT, Zilberman D (2014) Dnmt1-independent CG methylation contributes to nucleosome positioning in diverse eukaryotes. *Cell* 156: 1286-1297.
34. Zhang X, Yazaki J, Sundaresan A, et al. (2006) Genome-wide high-resolution mapping and functional analysis of DNA methylation in arabidopsis. *Cell* 126: 1189-1201.
35. Molaro A, Hodges E, Fang F, et al. (2011) Sperm methylation profiles reveal features of epigenetic inheritance and evolution in primates. *Cell* 146: 1029-1041.
36. Zeng J, Konopka G, Hunt BG, et al. (2012) Divergent whole-genome methylation maps of human and chimpanzee brains reveal epigenetic basis of human regulatory evolution. *Am J Hum Genet* 91: 455-465.
37. Honeybee Genome Sequencing C (2006) Insights into social insects from the genome of the honeybee Apis mellifera. *Nature* 443: 931-949.
38. Satou Y, Mineta K, Ogasawara M, et al. (2008) Improved genome assembly and evidence-based global gene model set for the chordate Ciona intestinalis: new insight into intron and operon populations. *Genome Biol* 9: R152.
39. Goll MG, Kirpekar F, Maggert KA, et al. (2006) Methylation of tRNAAsp by the DNA methyltransferase homolog Dnmt2. *Science* 311: 395-398.
40. Ponger L, Li WH (2005) Evolutionary diversification of DNA methyltransferases in eukaryotic genomes. *Mol Biol Evol* 22: 1119-1128.
41. Maumus F, Rabinowicz P, Bowler C, et al. (2011) Stemming Epigenetics in Marine Stramenopiles. *Current Genomics* 12: 357-370.
42. Bowler C, Vardi A, Allen AE (2010) Oceanographic and biogeochemical insights from diatom genomes. *Ann Rev Mar Sci* 2: 333-365.
43. Zimmermann C, Guhl E, Graessmann A (1997) Mouse DNA methyltransferase (MTase) deletion mutants that retain the catalytic domain display neither de novo nor maintenance methylation activity in vivo. *Biol Chem* 378: 393-405.
44. Fatemi M, Hermann A, Pradhan S, et al. (2001) The activity of the murine DNA methyltransferase Dnmt1 is controlled by interaction of the catalytic domain with the N-terminal part of the enzyme leading to an allosteric activation of the enzyme after binding to methylated DNA. *J Mol Biol* 309: 1189-1199.
45. Penterman J, Zilberman D, Huh JH, et al. (2007) DNA demethylation in the Arabidopsis genome. *Proc Natl Acad Sci U S A* 104: 6752-6757.
46. Cokus SJ, Feng S, Zhang X, et al. (2008) Shotgun bisulphite sequencing of the Arabidopsis genome reveals DNA methylation patterning. *Nature* 452: 215-219.
47. Hunt BG, Brisson JA, Yi SV, et al. Functional conservation of DNA methylation in the pea aphid and the honeybee. *Genome Biol Evol* 2: 719-728.
48. Foret S, Kucharski R, Pittelkow Y, et al. (2009) Epigenetic regulation of the honey bee transcriptome: unravelling the nature of methylated genes. *BMC Genomics* 10: 472.

49. Xiang H, Zhu J, Chen Q, et al. (2010) Single base-resolution methylome of the silkworm reveals a sparse epigenomic map. *Nat Biotechnol* 28: 516-520.
50. Hellman A, Chess A (2007) Gene body-specific methylation on the active X chromosome. *Science* 315: 1141-1143.
51. Brenet F, Moh M, Funk P, et al. DNA methylation of the first exon is tightly linked to transcriptional silencing. *PLoS One* 6: e14524.
52. Maunakea AK, Nagarajan RP, Bilenky M, et al. (2010) Conserved role of intragenic DNA methylation in regulating alternative promoters. *Nature* 466: 253-257.
53. Takuno S, Gaut BS (2012) Body-Methylated Genes in Arabidopsis thaliana Are Functionally Important and Evolve Slowly. *Mol Biol Evol* 29: 219-227.
54. Zilberman D, Gehring M, Tran RK, et al. (2007) Genome-wide analysis of Arabidopsis thaliana DNA methylation uncovers an interdependence between methylation and transcription. *Nat Genet* 39: 61-69.
55. Lorincz MC, Dickerson DR, Schmitt M, et al. (2004) Intragenic DNA methylation alters chromatin structure and elongation efficiency in mammalian cells. *Nat Struct Mol Biol* 11: 1068-1075.
56. Luco RF, Pan Q, Tominaga K, et al. (2010) Regulation of alternative splicing by histone modifications. *Science* 327: 996-1000.
57. Lyko F, Foret S, Kucharski R, et al. (2010) The honey bee epigenomes: differential methylation of brain DNA in queens and workers. *PLoS Biol* 8: e1000506.
58. Ammar R, Torti D, Tsui K, et al. (2012) Chromatin is an ancient innovation conserved between Archaea and Eukarya. *Elife* 1: e00078.
59. Nalabothula N, Xi L, Bhattacharyya S, et al. (2013) Archaeal nucleosome positioning in vivo and in vitro is directed by primary sequence motifs. *BMC Genomics* 14: 391.
60. Mersfelder EL, Parthun MR (2006) The tale beyond the tail: histone core domain modifications and the regulation of chromatin structure. *Nucleic Acids Res* 34: 2653-2662.
61. Cosgrove MS, Boeke JD, Wolberger C (2004) Regulated nucleosome mobility and the histone code. *Nat Struct Mol Biol* 11: 1037-1043.
62. Lermontova I, Schubert V, Fuchs J, et al. (2006) Loading of Arabidopsis centromeric histone CENH3 occurs mainly during G2 and requires the presence of the histone fold domain. *Plant Cell* 18: 2443-2451.
63. Hashimoto H, Sonoda E, Takami Y, et al. (2007) Histone H1 variant, H1R is involved in DNA damage response. *DNA Repair (Amst)* 6: 1584-1595.
64. Maheswari U, Jabbari K, Petit JL, et al. (2010) Digital expression profiling of novel diatom transcripts provides insight into their biological functions. *Genome Biol* 11: R85.
65. Bheda P, Swatkoski S, Fiedler KL, et al. (2012) Biotinylation of lysine method identifies acetylated histone H3 lysine 79 in Saccharomyces cerevisiae as a substrate for Sir2. *Proc Natl Acad Sci U S A* 109: E916-925.
66. Zhang K, Sridhar VV, Zhu J, et al. (2007) Distinctive core histone post-translational modification patterns in Arabidopsis thaliana. *PLoS One* 2: e1210.
67. Tan M, Luo H, Lee S, et al. (2011) Identification of 67 histone marks and histone lysine crotonylation as a new type of histone modification. *Cell* 146: 1016-1028.
68. Sana J, Faltejskova P, Svoboda M, et al. (2012) Novel classes of non-coding RNAs and cancer. *J Transl Med* 10: 103.

69. Ulitsky I, Bartel DP (2013) lincRNAs: genomics, evolution, and mechanisms. *Cell* 154: 26-46.
70. Okazaki Y, Furuno M, Kasukawa T, et al. (2002) Analysis of the mouse transcriptome based on functional annotation of 60,770 full-length cDNAs. *Nature* 420: 563-573.
71. Cabili MN, Trapnell C, Goff L, et al. (2011) Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. *Genes Dev* 25: 1915-1927.
72. Liu J, Jung C, Xu J, et al. (2012) Genome-wide analysis uncovers regulation of long intergenic noncoding RNAs in Arabidopsis. *Plant Cell* 24: 4333-4345.
73. Pauli A, Valen E, Lin MF, et al. (2012) Systematic identification of long noncoding RNAs expressed during zebrafish embryogenesis. *Genome Res* 22: 577-591.
74. Cech TR, Steitz JA (2014) The noncoding RNA revolution-trashing old rules to forge new ones. *Cell* 157: 77-94.
75. Molnar A, Schwach F, Studholme DJ, et al. (2007) miRNAs control gene expression in the single-cell alga Chlamydomonas reinhardtii. *Nature* 447: 1126-1129.
76. Zhao T, Li G, Mi S, et al. (2007) A complex system of small RNAs in the unicellular green alga Chlamydomonas reinhardtii. *Genes Dev* 21: 1190-1203.
77. Lopez-Gomollon S, Beckers M, Rathjen T, et al. (2014) Global discovery and characterization of small non-coding RNAs in marine microalgae. *BMC Genomics* 15: 697.
78. Norden-Krichmar TM, Allen AE, Gaasterland T, et al. (2011) Characterization of the small RNA transcriptome of the diatom, Thalassiosira pseudonana. *PLoS One* 6: e22870.
79. Supek F, Bosnjak M, Skunca N, et al. (2011) REVIGO summarizes and visualizes long lists of gene ontology terms. *PLoS One* 6: e21800.

AIMS Press