Evolutionary analysis of DNA methyltransferases in microeukaryotes: Insights from the model diatom Phaeodactylum tricornutum

Antoine Hoguin, Ouardia Ait, Chris Bowler, Auguste Genovesio, Fabio Rocha Jimenez Vieira, Leila Tirichine

Evolutionary analysis of DNA methyltransferases in microeukaryotes: Insights from the model diatom *Phaeodactylum tricornutum*

Antoine Hoguin[1], Ouardia Ait Mohamed[1†], Chris Bowler[1], Auguste Genovesio[1], Fabio Rocha Jimenez Vieira[1¥*] and Leila Tirichine[2*]

[1]Institut de biologie de l'Ecole normale supérieure (IBENS), Ecole normale supérieure, CNRS, INSERM, PSL Université Paris 75005 Paris, France

[2]Université de Nantes, CNRS, UFIP, UMR 6286, F-44000 Nantes, France

[†]Current affiliation: Immunity and Cancer Department, Institut Curie, PSL Research University, INSERM U932, 75005 Paris, France

[¥]Current affiliation: Medical Affairs Department, AFM-Téléthon, Evry, France

* Authors for correspondence: tirichine-l@univ-nantes.fr, fabiorjvieira@gmail.com

Abstract

Cytosine DNA methylation is an important epigenetic mark in eukaryotes that is involved in the transcriptional control of mainly transposable elements in mammals, plants, and fungi. Eukaryotes encode a diverse set of DNA methyltransferases that were iteratively acquired and lost during evolution. The Stramenopiles-Alveolate-Rhizaria (SAR) lineages are a major group of ecologically important marine microeukaryotes that include the main phytoplankton classes such as diatoms and dinoflagellates. However, little is known about the diversity of DNA methyltransferases and their role in the deposition and maintenance of DNA methylation in microalgae. We performed a phylogenetic analysis of DNA methyltransferase families found in marine microeukaryotes and show that they encode divergent DNMT3, DNMT4, DNMT5 and DNMT6 enzymes family revisiting previously established phylogenies. Furthermore, we reveal a novel group of DNMTs with three classes of enzymes within the DNMT5 family. Using a CRISPR/Cas9 strategy we demonstrate that the loss of the DNMT5 gene correlates with a global depletion of DNA methylation and overexpression of transposable elements in the model diatom *Phaeodactylum tricornutum*. The study provides a pioneering view of the structure and function of a DNMT family in the SAR supergroup.

Background

In eukaryotes the methylation of the fifth carbon of cytosines (5mC) is a recognized epigenetic mark involved in chromosome X inactivation, genomic imprinting, stability of repeat rich centromeric and telomeric regions as well as transposable elements (TEs) repression[1–4]. 5mC is deposited by DNA methyltransferases (DNMTs) capable of *de novo* methylation and is propagated through subsequent cell division by maintenance enzymes. Eukaryotes have acquired a diverse set of DNA methyltransferases (DNMTs) by horizontal gene transfer of bacterial DNA Cytosine methyltransferase (DCM) enzymes involved in the Restriction Methylation system[5]. All DNMTs are composed of a catalytic protein domain composed of ten conserved motifs (annotated I to X) that provide binding affinity to DNA substrate and the methyl donor cofactor S-Adenosyl methionine (SAM) to process the transfer of a methyl group to unmethylated cytosines [6,7]. DNMTs have further diversified over evolutionary time scales in eukaryote lineages and acquired chromatin associated recognition and binding domains giving rise to a wide diversity of DNA methylation patterning [8,9]. The loss and gain of DNMTs have been associated with profound divergence in cell biology and in the control of gene expression. To date, six main eukaryotic DNMT families have been described and named DNMT1, DNMT2, DNMT3, DNMT4, DNMT5 and DNMT6[10,11]. In metazoans, the concert activity of the DNMT3 family and DNMT1 enzymes allow the deposition and the maintenance of DNA methylation pattern during the successive developmental waves of DNA demethylation and remethylation[12]. In fungi, the DNA methylation machinery consists in a maintenance activity by DNMT1/DIM2, as in *Neurospora crassa* [13], or by the activity of ATPase-DNMT5 enzymes as reported in *Cryptococcus neoformans* [11,14]. Fungal DNMT4 relatives are involved in the DNA methylation related process known as Repeat-Induced Point Mutation (RIP) and Methylation Induced Premeiotically (MIP) that leads to transposable elements extinction and or stage specific repression as observed in *Aspergillus* and *Neurospora* species [15–18]. Successive gains and loss of DNMT1 and DNMT3 occurred during insect evolution leading to secondary loss of global 5mC methylation as in diptera lineages[19]. In plants, the acquisition of new DNMT1 proteins named Chromomethylases (CMTs) and the divergence of the DNMT3 family led to the spreading of the asymmetrical non-CG patterns of DNA methylation that is extensively found in angiosperms [20–22]. In *Micromonas pusilla*, the pelagophyte *Aureococcus annophagefferens* and the haptophyte *Emiliania huxleyi*, the presence of

a DNMT5 enzyme correlates with a heavy histone linker DNA methylation landscape[11]. Importantly, 5mC have been now reported in more diverse eukaryotes of the *SAR* lineages as in dinoflagellates [23], diatoms [24] and kelps [25]. However, due to the severe underrepresentation of marine unicellular eukaryotes in modern sequencing databases, our understanding of the DNA methylation machinery in microalgae remains scarce.

Diatoms are a dominant, abundant, and highly diverse group of unicellular brown micro-algae (from 2 to 200 µm) of the stramenopile lineage. It is estimated that diatoms are responsible for nearly 20% of earth primary production [26,27]. They are known to dominate marine polar areas and are major contributors of phytoplankton oceanic blooms. To date, 5mC has been reported in four diatoms, namely the centrics *Thalassiosira pseudonana* [11] and *Cyclotella cryptica* [28], as well as in *Fragilariopsis cylindrus* [11] and *Phaeodactylum tricornutum* [11,24], a polar and temperate pennate diatom respectively. Diatoms methylation pattern strongly contrasts with the patterns observed in animals but also dinoflagellates and plants [29]. Firstly in *P. tricornutum*, *T. pseudonana* and *F. cylindrus*, total levels of DNA methylation range from 8% to as low as 1% of Cytosines in the CG context [11] over repeats and transposable elements usually (but not exclusively) concentrated in telomeric regions [11,24]. Non-CG methylation is also detected but is scarce. Diatom genomes are thus mainly composed of isolated highly CG methylated TE islands in an otherwise unmethylated genome and to that regards are remarkably similar to fungi methylation profiles. In all diatoms studied so far, methylated TEs often have low expression [11,24,28]. This is very consistent with the repressive role of DNA methylation in other eukaryotes and further traces back 5mC mediated control of TE expression to the last eukaryotic common ancestor. Nonetheless, direct evidence of the repressive role of 5mC at transposable elements in diatoms is lacking.

The recent advance in high throughput RNA sequencing technologies applied to microeukaryotes led to the development of the Microbial Eukaryote Transcriptome Sequencing Project (MMETSP) consortium. The MMETSP concatenates more than 650 transcriptomes from diverse microeukaryote lineages as diatoms and dinoflagellates making it the biggest sequence database for marine microeukaryotes. Here using the newly defined enhanced Domain Architecture Framework (eDAF) methodology[30], we first explored the structural and phylogenetic diversity of DNMTs

sequences in marine microeukaryotes from the publicly available MMETSP sequencing databases. Using an integrative approach with available genomes and phylogenetic studies, we provide a DNMT phylogeny focused on the structural and domain diversity found in microeukaryote enzymes and discuss their evolutionary origins. We found that marine microeukaryotes encode a wide set of DNMTs that lack chromatin domains such as in DNMT6 and in divergent DNMT3 enzymes. We report an unprecedented diversity in the DNMT5 family and define the DNMT5a, b and c enzymes based on structure and phylogenetic assessment. We found that this diversity remarkably contrasts with the apparent lack of DNMT1 in most of the MMETSP and microeukaryotes databases. Secondly, using a CRIPSR/Cas9 knock out approach, we present the direct experimental characterization of the newly defined DNMT5a family of enzymes in the model diatom *P. tricornutum* demonstrating, for the first time in any SAR, the role of the DNMT5 family in the maintenance of DNA methylation and the role of DNA methylation in the repression of transposable elements in an early diverging eukaryote lineage.

Results

1. Diversity of DNA methyltransferases in microeukaryotes

In order to capture the diversity of 5-cytosines DNA methyltransferases encoded in microalga, we applied a relaxed HMMER search (e-value=1 as the cut-off threshold) for the PFAM DNMT (PF00145) domain on the whole transcriptomes of the MMETSP database. We complemented MMETSP with 7 diatoms transcriptomes and genomes from the top 20 most abundant diatoms found in TARA Oceans database (courtesy of the GENOSCOPE sequencing project). Next, we eliminated the expected high number of false positives using eDAF[30]. We also performed reciprocal BLAST best hit analysis on a phylogenetically optimized genomic and transcriptomic database as previously described[31]. We kept enzymes that showed conserved DNMT domains and depicted their domain structures by eDAF curation[30]. We build a representative phylogeny of DNA methyltransferases based on the alignment of DNMT motifs of conserved DNMTs (Figure 1a, supplementary Figure 1, supplementary table S1). We also compiled the observed structure of microalgal enzymes based on the analysis of DAMA[32]/CLADE[33] guided eDAF annotation of representative enzymes used for phylogeny (Figure 1b, Supplementary table S2). In this study we focused on the DNMT1, DNMT3, DNMT4, DNMT5 and DNMT6 gene families that are known or putative DNA modifying enzymes. Within lineages, a species level summary of presence absence of each gene family can be found in supplementary table S3**.**

Our results reaffirm and strengthen previously established phylogenies[10,11] and we report the six mains eukaryotic DNMT families accordingly (Figure 1a, supplementary Figure 1). We found however that the DNMT5 families branches into three highly supported subfamilies of enzymes that we hereafter named DNMT5a, b and c (Figure 1a, supplementary Figure 1). To our knowledge these gene families have never been described. The DNMT domain of DNMT5 families diverged at motifs I and IV (Figure 3). The DNMT5a and b families share an SNF2 related C-terminal domain composed of two helicases complemented or not by a RING finger domain (Figure 1b, supplementary table S2) while the DNMT5c family is composed of enzymes sharing a long (~1kb) N-terminal amino acid sequence with no annotated functional domains (Figure 1b, supplementary table S2). We found that diatoms DNMT5b enzymes contain an N-terminal laminin B receptor domain as in *T. pseudonana* (Figure 1b,

supplementary table S2). Other DNMT5b contain N-terminal CpG methyl binding domains, HAND and TUDOR domains (supplementary table S2). In addition, the DNMT domain of DNMT5b enzymes is longer than other DNMT5s between motifs VI and VII. These differences in structures probably highlight profound functional diversification between the DNMT5 subfamilies relative to ATPase activity and substrate specificity.

We found 76 species with at least one DNMT5 orthologue. This includes green algal DNMT5s as in *Tetraselmis astigmatica* that are similar to the published *Micromonas pusilla* enzyme (Figure 1b, supplementary table S2). Green algal DNMT5s are found within the DNMT5a subfamily (supplementary figure S1). We also found a DNMT5a transcript in the marine photosynthetic excavate euglenozoa *Eutreptiella gymnastica* closely related to green algal DNMT5as (Figure 2, supplementary Figure S1, supplementary table S3). The DNMT5 family is also widespread in diatom species in which DNMT5a enzymes are mainly found in pennate genomes while DNMT5b is found in both pennates and centrics. Interestingly, our data do not strongly support the presence of the DNMT5b family outside diatoms (Figure 2). *Emiliania huxleyi* has two annotated DNMT5s, one DNMT5a and one DNMT5b enzyme, but we could not find DNMT5 enzymes in other haptophytes of the MMETSP database suggesting horizontal acquisition of this enzyme family. DNMT5 transcripts identified in dinoflagellates transcriptomes fall into the DNMT5c family, and reciprocally, no DNMT5c enzymes are found in other lineages (Figure 2, supplementary table S3). Overall, our data support the absence of SNF2 containing DNMT5s in dinoflagellates with a few exceptions of partial DNMT5a/b enzymes (supplementary figure 1, supplementary table S2). Within DNMT5a, we identify at least 3 monophyletic group of enzymes: diatoms DNMT5as, Pelagophyte/Dictyochophyte DNMT5as and green algal DNMT5as. However, fungal enzymes are not monophyletic, and we failed to recover monophyletic relationship between ochrophyte stramenopiles suggesting a complex phylogenetic history of DNMT5 subfamilies. In addition, DNMT5 is absent from non-photosynthetic stramenopiles such as labyrinthumycetes and raphidophytes (Figure 2).

In our phylogeny the DNMT4 and DNMT1 clade form a moderately supported gene family as previously described[11,34] (Figure 1a, supplementary Figure 1). DNMT1s often associate DNMT catalytic domain with chromatin binding domains as bromo-

adjacent homology (BAH) domains, Plant Homeo Domains (PHDs), chromodomains and domains required for the interaction with accessory proteins. We isolated DNMT1/MET1 transcripts similar to the plant MET1 enzyme in 7 green algae species of the MMETSP database as in *Chlamydomonas* species (Figure 1b, Figure 2, supplementary Figure 1, supplementary table S3), reaffirming that the DNMT1 family was acquired early during plant evolution. CMT enzymes are divergent DNMT1 related enzymes that likely appeared in the very first plant lineages. Its conservation in green algae is more contrasted as many partial CMT enzymes as well as misannotated DNMT1 transcripts were found in early reports [35]. Accordingly, we could not find *bona fide* CMT transcripts in green algae species of the MMETSP data base and instead found two partial CMT transcripts in *Polytomella parva* and *Chlamydomonas chlamydogama* that lack BAH, PHD and chromodomains (supplementary Figure 1, supplementary table S2). Interestingly, we found a DNMT1 related enzyme in three haptophytes species (*Gephyrocapsa oceanica*, *Isochrysis.sp-CCMP1324* and *Coccolithus pelagicus*) of the MMETSP database that cluster with annotated CMTs found in the coccolithophore *Emiliania huxleyi* (Figure 1a, supplementary Figure 1). Haptophytes enzymes distantly relate to the conserved green algae CMTs (hCMT2) enzymes (Figure 1a, supplementary Figure S1). We found that the enzymes of *Gephyrocapsa oceanica* (CAMPEP_0188208858), *Isochrysis-CCMP1324* (CAMPEP_0188844028) and *Emiliania huxleyi* (jgi_458278) have DNMT1-like structures with a Replication Foci Domain (RFD) followed by a BAH (in *Emiliana huxleyi* only) and a conserved DNMT domain (Figure 1b, supplementary table S2). We also found two putative DNMT1-like enzymes in the transcriptomic database of *Raphidophyceae* brown microalgae *Heterosigma akashiwo* and *Chatonella subsala*. They are composed of a conserved DNMT domain and a plant homeodomain (PHD) (Figure 1b, supplementary figure 1, Figure 2, supplementary table S2). Raphidophycea enzymes poorly define a DNMT family by themselves and distantly relate to the DNMT1 enzymes (Figure 1a, Supplementary Figure 1). We could not find compelling evidence for *bona fide* DNMT1s in any other microalgae. Our data support the hypothesis that the DNMT1 maintenance machinery is absent from the main microeukaryote lineages that therefore rely on DNMT1-independent mechanisms for the maintenance of their DNA methylation pattern as previously suggested[11].

DNMT4 enzymes have been identified in fungi and are involved in the MIP or RIP processes. Closely related to DIM2 enzymes in fungi[34]; two DNMT4 enzymes were also found in the pennate diatom *F. cylindrus* and the centric diatom *T. pseudonana*[10]. Accordingly, we identified a monophyletic diatom enzyme family that relate to the fungal DNMT4 gene family (Figure 1a, Supplementary Figure 1). Diatoms and RID/DMTA enzymes as well as DNMT1 proteins form a moderately supported monophyletic clade. A total of 31 diatoms, pennate and centric species expresses or encodes at least one DNMT4 related transcript (supplementary table S4). This indicates that the DNMT4 family was acquired early during diatoms evolution. Most diatoms DNMT4 enzymes are composed of a single DNMT domain as in *T. pseudonana*, which contrasts with fungal enzymes (Figure 1b – supplementary table S2). However, 9 diatom's DNMT4 proteins possess an additional N-ter chromodomain as observed in *Thalassiosira miniscula* (Fig. 1b, supplementary table S2 and S4). We found the microalgal DNMT4 family is restricted to diatoms. The DNMT4 family is thus only observed in two very distantly related microeukaryote lineages questioning the evolutionary origin of this gene family in eukaryotes.

Our data indicate that the DNMT3 family is not particularly abundant in microalgae (Figure 2, supplementary table S3). DNMT3 is absent in most stramenopiles except for diatoms; in which both genomic and transcriptomic data strongly support its presence (supplementary table S4). DNMT3 seems absent in haptophytes (Figure 2, supplementary table S3). Only one transcript from the cryptomonad *Goniomonas pacifica* could be annotated as DNMT3. In addition, we could not identify DNMT3 enzymes in any green algae of the MMETSP database, but it is present in red algae as found in the genomes of *Cyanidioschyzon merolae* and *Galdieria sulphuraria* (Figure 2, supplementary table S3). We also report few additional DNMT3 transcripts in dinoflagellates as previously described[23] (Figure 2, supplementary table S3). Upon alignment, dinoflagellates DNMT3 enzymes (including former annotated enzymes[23]) and *Goniomonas pacifica* DNMT3s are closely related to red algal genes but diverged compared to other DNMT3s while diatoms form their own DNMT3 family (Supplementary figure 1). This suggest that the DNMT3 family was iteratively lost and acquired several times during microalgae evolution. As observed in *P. tricornutum*, DNMT3 enzymes found in microalgae all lack chromatin associated

domains (Figure 1b, supplementary table S2). This contrasts with mammalian DNMT3s[36] that interact with histone post-translational modifications.

We found that the DNMT6 enzymes are among the most widespread DNMTs in microeukaryotes. We found a DNMT6 transcript in the MMETSP transcriptomes of 3 *Tetraselmis* green algae and 7 dinoflagellates (Figure 2 – supplementary table S3). In addition, DNMT6 is extensively found in stramenopiles as in *Dictyochophyceae*, *Crysophyceae* and *Pelagophyceae* (Figure 2, supplementary table S3). In diatoms, DNMT6 is very abundant (Supplementary table S4). DNMT6 is also present in the non-photosynthetic labyrinthulomycetes *Aplanochytrium stocchinoi* and probably in *Aplanochytrium keurgelense* (Figure 2, supplementary table S3). In addition, our data also strongly support the presence of DNMT6 orthologue in the major *Chromalveolata* lineage of *Rhizaria* (Figure 2, supplementary table S3) as suggested in previous reports [23]. DNMT6 enzymes are for most part homogeneous and do not contain chromatin associated signatures, as in *P. tricornutum* DNMT6 and DNMT3 enzymes (Figure 1b, supplementary table S2). Monophyletic relationships within the DNMT6 family and between micro-eukaryotes could not be resolved (supplementary figure 1).

2. Functional study of DNMT5 in the model diatom *Pheaodactylum tricornutum*

To our knowledge no DNMTs have been functionally characterized in SARs. The pennate *P. tricornutum* is the model species that we and others used to decipher the epigenomic landscape in diatoms, shedding light into the conservation and divergence of the methylation patterns in early diverging eukaryotes [24,37]. *P. tricornutum* encodes a DNMT3, DNMT6 and DNMT5a orthologues in single copy but lacks a DNMT4 and DNMT5b orthologues found in other diatoms (supplementary table S4). We asked whether any of those DNMTs had DNA methylation function(s) *in vivo*. Using a CRISPR/Cas9 mediated knock-out approach we screened *P. tricornutum* genome for DNMTs loss of function mutants (see material and methods). In this work, we report two independent mutants with homozygous out of frame deletions generating premature STOP codons in the coding sequence of DNMT5a (Figure 4). DNMT5:KOs will be hereafter referred to as the ‘M23’ and ‘M25’ lines respectively. No DNMT3 or DNMT6 mutations could be generated using our CRISPR/Cas9 strategy.

Using sets of primer pairs targeting the DNMT domain as well as the DEADX helicase-SNF2 like domain of DNMT5 transcripts, we monitored by quantitative RT-

PCR a 4 to 5-fold loss in mRNAs levels in both M23 and M25 lines (Supplementary Figure 2). 5mC dot blot screening revealed that all DNMT5:KOs had a 4-3 fold loss of DNA methylation compared to the Pt18.6 reference ('wild-type') and Cas9:mock controls (Supplemental Figure 2) consistent with the putative role of DNMT5 in maintaining DNA methylation patterns in diatoms.

To get a quantitative single base resolution of DNA methylation loss in DNMT5:KOs, we performed whole genome bisulfite sequencing on the M23, M25 and the reference Pt18.6 line. To assess highly methylated regions, we first filtered cytosines by coverage and methylation levels considering a 5X coverage threshold and 60% methylation levels in all lines. In the reference Pt18.6, we identified 10349 methylated CGs (0.3% of all analyzed CGs), 129 methylated CHHs (3/10000) and 12 methylated CHGs (<5/10^6). About 76% of methylated cytosines are found within TEs confirming that repeated sequences are the main targets of CG methylation in *P. tricornutum* genome (Figure 5a). In comparison only 9% of methylated cytosines overlap with genes (Figure 5a). As a result, 876/3406 (25%) of transposable elements and 170/9416 (1,8%) of genes are marked by at least one methylated CG in Pt18.6. Among methylated TEs, 711 (81%) were also found methylated in the reference strain in earlier reports[24] and reciprocally (supplementary Figure 3). Gene methylation was less consistent between already published reports[24] as 230 (63%) of known methylated genes are not found methylated in the present study (supplementary Figure 3), which might be due to differences in culture conditions and/or coverage depths and methods used to profile DNA methylation. In DNMT5:KOs, a maximum of 2 cytosines in any context are found methylated. Methylated cytosines identified in the reference strain show background levels of DNA methylation in DNMT5:KOs (Figure 5 a). The knockout of DNMT5 thus generated a global loss of all cytosine methylation in *P. tricornutum*.

We asked whether the extensive loss of DNA methylation in DNMT5:KOs could define differentially methylated regions (DMRs) and computed DMRs between DNMT5:KOs and WT lines using the bins built-in DMRcaller[38] tools (material and method). DMRs corresponding to hypo methylated regions in DNMT5:KOs cover ~0.8% of *P. tricornutum* genome. In addition, 96% of DMRs found in M23 and M25 are shared while about 4% of DMRs are found in one mutant only, showing the consistent loss of DNA methylation in independent cell lines upon knockout of DNMT5 (supplementary Figure S3). We found that ~83% of DMRs overlap with transposable

elements and their regulatory regions (~5%) (Figure 5 b). Reciprocally, 584 (~15%) annotated transposable elements are found within DMRs (+/-500bp) in DNMT5:KOs lines (Figure 5 c, Supplementary table 7). Among them 544 showed at least one highly methylated CG (Figure 5c). We found that 70% of all methylated cytosines in the reference strain concentrate in TEs that overlap with DMRs and 90% of all methylation found at TEs are found within DMR-TEs (supplementary figure S3). Outside DMRs, methylated TEs show less than 2 methylated cytosines while within DMRs TEs show extensive methylation (supplementary Figure 3). Our DMR analysis thus identified a significant loss of DNA methylation at transposons that also concentrate most of densely methylated cytosines in *P. tricornutum*. There is a low overlap of DMRs with genes or their regulatory regions and only ~0.5% of all protein coding genes (30% of genes with at least one highly methylated CGs) are found within DMRs (Figure 4c). Nonetheless, genes within DMRs concentrate more than 80% of methylation found in genes in the reference strain. As for TEs, genes identified in hypoDMRs are also among the most methylated genes in the reference strain. We then asked whether DMRs associate with known regions marked by histone posttranslational modifications. Interestingly, between 80% and 90% of all DMRs overlap with known regions marked by H3K27me3, H3K9me3 or H3K9me2 defined in the reference Pt18.6 line[37] (Figure 5d). We found that 70% of DMRs are found within regions co-marked by all three repressive histone marks in association or not with active marks. Our DMR analysis thus retrieved the denser and highly heterochromatic regions of *P. tricornutum* genome that are reported targets of CG methylation in the reference strain. Our data are consistent with a global loss of DNA methylation in DNMT5:KOs at TE-rich DNA methylated-H3K27me3, H3K9me2 and H3K9me3 rich regions of *P. tricornutum* genome. Of note, one hypermethylated region was found in the mutant M23 in the promoter region of the gene Phatr3_J46344. Consistent with the low level of CHH and CHG methylation in the reference strain we found no CHG and CHH hypo nor hyper DMRs.

The control of transposable elements by the DNA methyltransferases family is a key unifying feature within eukaryotes[2]. We hence monitored the transcriptional effect of the loss of DNMT5 in M23/M25 by whole RNA high throughput sequencing (Material and Methods). To identify the most significant changes in mRNA levels, we focused our analysis on genes showing a significant 2 folds induction or reduction of

expression in mutants compared to the reference line (LFC > |1| and an FDR< 0.01). We also analyzed expression of genes overlapping with TE annotations (TE-genes) as previously described[39]. In both DNMT5:KOs, we observed a global overexpression of genes and TE-genes and comparatively fewer downregulation (Figure 6a). TE-genes show higher level of overexpression than genes that show milder inductions suggesting that primary targets of DNA methylation in the reference strains are also the first to respond to the loss of DNA methylation in DNMT5:KOs. Gene overexpression is inconsistent between lines (Figure 6b). At the transcriptional level, M23 shows up to 1133 up regulated genes and M25 shows 393 upregulated genes. Only 219 genes have consistent upregulation in both mutants. Transposable element showed more overlap. We found that 339 TE-genes with up-regulation in at least one mutant are consistently up-regulated in both (Figure 6c). We performed GOterm enrichment using TopGO tools[40] and found that overexpressed genes in both mutants are enriched for GOs associated with protein folding responses as well as nucleotide phosphate metabolism and nucleotide binding activity. This is typified by the overexpression of chaperone DnaJ domain containing proteins and Hsp90 like proteins (Supplementary tables S10 and S11). The recent advances in high through output RNA sequencing technologies led to the recent identification of gene co-regulatory networks in *P. tricornutum*[41]. Genes with overexpression in DNMT5:KOs are enriched within the “salmon” and “steelblue” coregulatory gene modules. The ‘salmon’ module contains genes associated with chloroplast biology as well as amino acid related enzymes which might highlight metabolic defects. The ‘steelBlue’ regulatory module associates with protein translation as well as nucleotide metabolism and is enriched in genes involved in purine and pyrimidine cellular pathways[41]. Accordingly, some upregulated proteins are involved in ribosome biosynthesis and RNA processing. The chaperone like enzymes found upregulated in DNMT5:KOs are part of both the ‘salmon’ and ‘steelblue’ pathway (supplementary table S12). Only 36 genes showed consistent down-regulation in both lines in a DNA methylation independent manner (Supplementary figure 4). Among those genes we found back DNMT5a, that shows a 3-fold loss in both mutants confirming qpcr analysis and suggesting that premature stop codons triggered mRNA degradation (Supplementary table S13). We could not find compelling GO enrichment nor coregulation network enrichment for downregulated genes in DNMT5:KOs. A handful of TE genes show loss of expression

in DNMT5:KOs that is not associated with DNA methylation loss (Supplementary figure 4).

We asked if we could directly or indirectly link upregulation of genes and TEs to DNA hypomethylation in DNMT5:KOs. Interestingly, we found that gene upregulation is not linked to gene loss of DNA methylation since 31 (60%) of genes in shared hypo DMR do not show changes in mRNA levels in any mutants (Supplementary figure S4). However, 66% (226) of TE-genes with upregulation in both DNMT5:KOs also overlap with hypoDMRs (Supplementary figure S4). Reciprocally, half of TE-genes overlapping with *bona fide* TEs found in hypoDMRs are upregulated in at least one DNMT5:KOs.. Consistent with the association of TEs and DMRs, we found that up regulated TE-genes fall into regions that typify the repressive epigenetic landscape of *P. tricornutum* that also overlap with DMRs identified in M23 (figure 7a, supplementary table S9). Notably, 80% of TE-genes marked by H3K27me3 and in hypoDMRs are overexpressed in both DNMT5:KOs (Figure 7a and b – category F). Up-regulation is also associated with hypoDMR regions co-marked by H3K27me, H3K9me3, H3K9me2, H3K4me3 and to a lesser extent H3K19_14Ac (Figure 7a and b - category B, D and E). Interestingly, a subset of TEs in regions that were not highly methylated in the reference strain but marked by H3K27me3, H3K9me2 and H3K9me3 also show upregulation (Figure 7a, b – category C). TEs that overlap with hypoDMRs and heterochromatin regions of the genome are among the most highly expressed TEs in our datasets (Figure 7b – category B, D, E and F). Between mutants, M23 shows the highest number of TEs with significant changes in expression per epigenetic categories. However, levels of inductions are not statistically different except for the C and E categories (Figure 7b). In conclusion, our data show that loss of DNA methylation primarily associates with reactivation of a subset of highly repressed TEs, but not genes, found in heterochromatic regions of *P. tricornutum* genome.

Discussion

This work extends earlier studies in the field and further demonstrates that DNA methyltransferases are much diversified in microalgae. Studies on the evolutionary history of DNMTs have established that the DNA methylation machinery diverged between eukaryotes along with their respective epigenetic landscape. Using the MMETSP we found that secondary endosymbionts rather encode a combination of

DNMT5, DNMT3 and DNMT6 enzymes. We nonetheless identified CMTs with DNMT1-like features in haptophytes as well as new DNMT1-like enzyme in *Raphidophyceae.* This suggests that CMTs and DNMT1 type of enzymes might have originated prior to the *chlorophyta-embryophyta* divergence and in the common heterotrophic ancestor but were extensively lost in a lineage specific manner. Alternatively, horizontal gene transfer of DNMT1 type of enzymes between microalgae, such as during endosymbiotic events, could have sustained continuous flow of this gene family in secondary and tertiary endosymbiotic lineages such as haptophytes and stramenopiles.

In our phylogeny, the RID/DMTA and the diatoms DNMT4 enzymes are clearly related as shown by Huff and Zilberman[11] and Punger and Li[10]. Within fungi, RID proteins were found closely related to DIM2 enzymes[34] for which we have not found any related orthologue in diatoms. We make the hypothesis that the DNMT4 family is closely related to DNMT1 enzymes and arose by convergence in the diatom and fungi lineages. The function of DNMT4 in diatoms is unknown. Whether RID or MIP process occur in any DNMT4 containing diatom also remain an open question. The presence of chromodomains that are known to bind histone post translational modifications as in CMT enzymes[1,42] nonetheless suggests that diatoms DNMT4 might be functional as either *de novo* or maintenance enzymes.

DNMT6 enzymes are very well detected in diatoms and dinoflagellates. Our study independently reports the same DNMT6 enzymes found in *Bigelowella natans* and *Aplanochytrium stochhinoi* by earlier works although not specified by the authors[23]. As it is reported in trypanosomes[10], we suggest that DNMT6 likely emerged prior to the *Chromalveolata* radiation. It is the only known DNMT enzyme in Trypanosomes that does not show extensive DNA methylation pattern in any context[43]. DNMT6 is not active in *Leishmania* species[43]. Its presence in several lineages, hence, does not predict DNA methylation *per se* and must be further investigated.

We showed that the DNMT domain of the different DNMT5s are conserved within each other's and diverged compared to other DNMTs supporting a common evolutionary origin for all DNMT5 enzymes. It was proposed that DNMT5s are ancestral to eukaryotes. However, we found that DNMT5s are more diverse than previously thought as showed by the structural divergence between DNMT5s

subfamilies apart of their DNMT domain. It was established that DNMT5s rely on the ATPase activity of their SNF2 like domains to methylate hemi-methylated cytosines[14]. In that regards DNMT5c that are also dinoflagellate specific, lacks ATPases. The DNMT5 family also likely diverged in diatom species as DNMT5b enzymes specially in centric diatoms. DNMT5b might be multifunctional related DNMT5 enzymes as it is suggested by the presence of N-ter HAND domains, which are found in chromatin remodelers[44], as well as TUDOR domains that are found in histone modifying enzymes and readers[45] and small RNA interacting proteins[4647]. In the genomes of *Fragilariopsis cylindrus* species as well as in *Synedra* both DNMT5a and b are found. This supports a scenario in which both DNMT5a and b were acquired early in diatoms and might be paralogues. In other diatoms, DNMT5a is however phylogenetically restricted. It is almost exclusively found in annotated genomes of pennate diatoms. *Corethron pennatum* is the only centric diatom with a DNMT5a transcript. Besides transcriptomic evidence, the absence of DNMT5a in centric diatoms is further supported by genomic data in 5 centric species in which at least another DNMT could be detected, especially DNMT5b (table S4). The diatoms DNMT5a enzymes might originate from a pennate specific horizontal gene transfer with another stramenopiles encoding DNMT5a and or within diatoms themselves.

The independent accession of DNMT5s in ochrophyte algae is supported by our phylogenies in which diatoms and pelagophytes/dictyochophytes enzymes are not monophyletic. DNMT5 enzymes are thus related gene families that show lineage specific divergence. The origin of the DNMT5c is also intriguing since it is both highly divergent compared to other DNMT5s and dinoflagellate specific. In DNMT5 enzymes, we noticed that the DNMT motifs I; involved in the binding of S-Adenosyl-methionine; and IV; the catalytic site; are highly divergent compared to other DNMTs suggesting that DNMT5 enzymes might require accessory proteins aside of their ATPase activity. In addition, the lack of chromatin associated domains in DNMT3 proteins, but also DNMT6 and DNMT4 proteins suggest that the link, if any, between DNA methylation and histone modifications is more indirect than observed in plants and mammals (more discussed below) and might also require the activity of accessory proteins such as UHRF1 type[48] or DNMT3-like[49] enzymes that we have not investigated in this study.

Several species presented in this study are not known to have DNA methylation. Crosschecking the diversity in DNA methyltransferases with *bona fide in vivo* functions

should retain much attention. In this context, we investigated the role of DNMT5a which is conserved in the pennate diatom *P. tricornutum*. As discussed above, the DNMT5a enzyme of *P. tricornutum* is not shared with many other diatoms but is an orthologue of the DNMT5a protein from *Cryptococcus neoformans* which is involved in the maintenance of DNA methylation patterns[11,14]. In laboratory conditions, *P. tricornutum* propagates via mitotic division. Our study demonstrates that the 'somatic' (i.e. clonal) loss of DNMT5 was sufficient alone to generate a global loss of CG methylation upon repeated cellular divisions. For this reason, we state here that we monitored a loss due to defective maintenance activity. TEs are major targets of DNA methylation in numerous eukaryote lineages including dinoflagellates and diatoms anchoring these features to the origin of eukaryotes[23,24]. Our DMR analysis *de novo* identified regions that show extensive methylation in the reference strain i.e TEs that were also reported in earlier studies[24] (supplementary figure S3).

We further demonstrate that only 15% of TEs concentrate more than 70% of DNA methylation in *P. tricornutum*. We found that most TEs associate with active epigenetic marks alone. It strongly suggests that although TE methylation is indeed observed, TE regulation in diatoms is for most part not directly dependent of DNA methylation in general. Genes showed low overlap with both DMRs and previous reports (supplementary figure 3) strongly suggesting that gene methylation is more labile even within independent lines of *P. tricornutum*. In addition, the loss of DNA methylation did not trigger extensive changes in gene expression but a reactivation of transposable elements. This is a strong support for the hypothesis that 5mC was recruited to keep TEs directly or indirectly at bay in the very first eukaryotic radiations. Nonetheless, our data also show that not all TEs are responsive to the loss of DNA methylation independently of their methylation levels. In *P.tricornutum,* chromatin immunoprecipitation studies followed by high throughput sequencing demonstrated that a subset of TEs are co-marked by DNA methylation as well as repressive histone modifications such as H3K27me3 and H3K9me2/3 suggesting redundancy or dependency within the epigenetic code of diatoms[37]. Interestingly, we found that upregulation is also associated with TEs in highly heterochromatic regions including transposable elements co-marked by DNA methylation and H3K27me3. This combination of repressive marks at TEs is remarkable. DNA methylation is known to interact with histone modifications. The interdependency of DNA methylation and

H3K9 methylation is well documented in plants[50], fungi[51] and mammals[52],[53]. In *Arabidopsis thaliana*, upon loss of CG methylation in *met-1* (DNMT1) mutants, H3K27me3 is relocated at specific hypomethylated TEs[54]. Similar observations were made in *ddm1* mutants defective for heterochromatic cytosine methylation[55]. In mammalian systems, compensatory relocation mechanisms of H3K27me3, but also H3K9me3, occur upon extensive DNA hypomethylation to maintain repression in a loci specific but also TE family specific manner[56]. We make the hypothesis that, aside direct repressive roles, the loss of DNA methylation also triggered genome wide changes within the epigenetic landscape of *P. tricornutum* in turn controlling the expression of a subclass of TEs. It was found that a subset of transposons are responsive to nitrogen depletion in conjunction with hypomethylation in *P. tricornutum*[57]. It is thus possible that TE expression also relies on condition specific transcription factor binding that are not fully recapitulated in our growth conditions.

Previous studies hinted for a potential RNA dependent DNA methylation system,[58] however, the mechanism of *de novo* methylation, if any, in diatoms is elusive. We did not found evidences for loss or gain of expression of other DNMTs nor putative DICER and AGO proteins in DNMT5:KOs. In addition, given that no regions retained DNA methylation upon loss of DNMT5:KOs alone we did not put in evidence direct compensatory *de novo* DNA methylation mechanisms in *P. tricornutum* DNMT5:KOs. It is possible that *de novo* methylation activity is performed by DNMT3 and or DNMT6 and or also directly or indirectly require DNMT5 activity. We could not generate DNMT3 or DNMT6 KOs. Complementation studies using ectopic expression of DNMT3, DNMT6 and or heterologous DNMT4 enzymes in *P. tricornutum* DNMT5:KOs loss of methylation background are conceivable experiments that could shed light into the mechanisms behind the setting or resetting of DNA methylation in diatoms.

DNMT5 mutant lines are viable in conditions used in routine for *P. tricornutum* culture which indicate that in optimal conditions loss of DNA methylation is not associated with drastic biological effects. Metabolic changes in DNMT5:KOs should be further determined in future experiments under stress conditions. In mammals, the release of TE repression by loss of DNMT3 enzymes triggered strong meiotic defects due to genomic instability. DNA methylation mediated TE control is also linked to the sexual cycle of cells as in MIP processes. Contrary to *P. tricornutum*, diatom species amenable for direct genetic approaches are also known to undergo sexual

reproduction. It is however unknown if DNA methylation plays a critical role in protecting the genome of diatom cells during meiosis. We would also like to highlight that *P. tricornutum* encode a unique set of DNMTs even within diatoms. In addition, closely related diatom species often encode a different combination of enzymes as we show for DNMT5b and DNMT5a clades between centrics and pennates. Within pennates, DNMT4 and DNMT3 were likely lost multiple times. We make the hypothesis that diatoms show extensive diversity in their DNA methylation machinery.

In conclusion, we provide the first report of the function of DNMT5 in Stramenopiles and update our knowledge about its diversity in eukaryotes with the identification of three new subclasses within this family of enzymes. As for the conservation of DNMT5 in other ochrophyte and green algae, we firmly hypothesize that its function is also conserved. DNMT5 is likely a major functional DNMT family that has expanded in SAR lineages as did DNMT1 in metazoans and land plants.

Material and Methods

Phylogenetic analysis of DNMTs in microeukaryotes

HMMER and RBH analysis

We performed an extensive scan of the MMETSP database by HMMER-search using the model PF00145 to fetch any DNMT-like, including partial transcripts, sequence within micro eukaryotes. We ran HMMER in a no-stringent fashion aiming to do not miss positives DNMT sequences. We used eDAF approach to filter the expected high number of false positives. It is worth noting that we initially use HMMER for screening instead of the built-in module of eDAF due to the time complexity of the latter for extensive searches (tens to hundreds of times slower than HMMER). Reciprocal BLAST best hit analysis was performed as previously described[31]. The DNMT3 (Phatr3_J47136), DNMT4 (*Thaps3_11011*), DNMT5 (*Phatr3_EG02369*) and DNMT6 (Phatr3_J47357) orthologues found in *P. tricornutum* or *T. pseudonana* (for DNMT4) were blasted on a phylogenetically optimized database that include MMETSP transcriptomes. Upon reciprocal BLAST, putative DNMT sequence hits giving back the corresponding enzyme (DNMT3, DNMT4, DNMT5 or DNMT6) at the threshold of e-

value of 1 × 10^-5 in the corresponding diatom were retained. Candidate enzymes were then analyzed using eDAF.

eDAF-guided domain architecture analysis

enhanced Domain Architecture Framework (eDAF) is a four module computational tool for gene prediction, gene ontology and functional domain predictions[30]. As previously described for Polycomb and Trithorax enzymes [30], candidate DNMTs identified by RBH and HMMER-search were analyzed using the DAMA-CLADE guided built-in functional domain architecture. The domain architecture of representative enzymes used in this study can be found in Supplementary table S2.

Phylogenetic tree analysis

The DNMT domain of candidate enzymes were aligned using ClustalΩ[59]. The alignment was manually curated and trimmed using trimAL (removing >25% gap column) to align corresponding DNMT motifs in all gene families. A convergent phylogenetic tree was then generated using the online CIPRES Science gateway program[60] using MrBAYES built-in algorithm. Default parameters were used with the following specifications for calculation of the posterior probability of partition: sumt.burninfraction =0.5, sump.burningfraction = 0.5.

Cell cultures

Axenic *Phaeodactylum tricornutum* CCMP2561 clone Pt18.6 cultures were obtained from the culture collection of the Provasoli-Guillard National Center for Culture of Marine Phytoplankton (Bigelow Laboratory for Ocean Sciences, USA.). Cultures were grown in autoclaved and filtered (0.22 µM) Enhanced Sea Artificial Water (ESAW - https://biocyclopedia.com/index/algae/algal_culturing/esaw_medium_composition.php) medium supplemented with f/2 nutrients and vitamins without silica under constant shaking (100rpm). Cultures were maintained in flasks at exponential state in a controlled growth chamber at 19°C under cool white fluorescent lights at 100 µE m−2 s−1 with a 12h photoperiod. For RNA-sequencing and bisulfite experiments, WT and DNMT5 mutant cultures were seeded in duplicate at 10.000 cells/ml and grown side by side in 250ml flasks until early-exponential at 1.000.000 cells/ml. Culture growth was followed using a hematocytometer (Fisher Scientific, Pittsburgh, PA, USA). Pellets were collected by centrifugation (4000rpm) washed twice with marine PBS

(http://cshprotocols.cshlp.org/content/2006/1/pdb.rec8303) and flash frozen in liquid nitrogen. Cell pellets were kept at -80°C until use. For bisulfite sequencing, technical duplicates were finally pooled to get sufficient materials.

CRISPR/Cas9 mediated gene extinction

The CRSIPR/Cas9 knock outs were performed as previously described[61]. Our strategy consisted in the generation of short deletions and insertions to disrupt the open reading frame of putative DNMTs of *P. tricornutum*. We introduced by biolistic the guide RNAs independently of the Cas9 and ShBle plasmids, conferring resistance to Phleomycin, into the reference strain Pt18.6 (referred hereafter as 'reference line' or 'wild-type'-WT). Briefly, specific target guide RNAs were designed in the first exon of Phatr3_EG02369 (DNMT5), Phatr3_J47357 (DNMT6) and Phatr3_J36137 (DNMT3) using the PHYTO/CRISPR-EX[62] software and cloned into the pU6::AOX-sgRNA plasmid by PCR amplification. For PCR amplification, plasmid sequences were added in 3' of the guide RNA sequence (minus –NGG) which are used in a PCR reaction with the template pU6::AOX-sgRNA. Forward primer – sgRNA seq + GTTTTAGAGCTAGAAATAGC. Reverse primer - sequence to add in 3' reverse sgRNA seq + CGACTTTGAAGGTGTTTTTTG. This will amplify a new pU6::AOX-(your_sgRNA). The PCR product is digested by the enzyme DPN1 (NEB) in order to remove the template plasmid and cloned in TOPO10 *E. coli*. The sgRNA plasmid, the pDEST-hCas9-HA and the ShBLE Phleomycin resistance gene cloned into the plasmid pPHAT-eGFP were co-transformed by biolistic in the Pt18.6 'Wild Type' strain as described in[61]. We also generated a line that was transformed with pPHAT-eGFP and pDEST-hCas9-HA but no guide RNAs. This line is referred as the Cas9:Mock line.

RNA and DNA extraction

Total RNA extraction was performed by classical TRIZOL/Chloroform isolations and precipitation by isopropanol. Frozen cell pellets were extracted at a time in a series of 3 technical extraction/duplicates and pooled. RNA was DNAse treated using DNAse I (ThermoFisher) as per manufacturer's instructions. DNA extraction was performed using the Invitrogen™ Easy-DNA™ gDNA Purification Kit following 'Protocol #3' instructions provided by the manufacturer. Extracted nucleic acids were measured using QUBIT fluorometer and NANODROP spectrometer. RNA and gDNA Integrity were controlled by electrophoresis on 1% agarose gels.

RT-QPCR and analysis

qPCR primers were designed using the online PRIMER3 program v0.4.0 defining 110-150 amplicon size and annealing temperature between 58°C and 62°C. Primer specificity was checked by BLAST on *P. tricornutum* genome at ENSEMBL. For cDNA synthesis, 1µg of total RNA was reverse transcribed using the SuperScript™ III First-Strand (Invitrogen) protocol. For quantitative reverse transcription polymerase chain reaction (RT-qPCR) analysis, cDNA was amplified using SYBR Premix ExTaq (Takara, Madison, WI, USA) according to manufacturer's instructions. CT values for genes of interest were generated on a Roche lightcycler® 480 qpcr system. CT values were normalized on housekeeping genes using the deltaCT method. Normalized CT values for amplifications using multiple couple of primers targeting several cDNA regions of the genes of interest were then averaged and used as RNA levels proxies.

Dot blot

gDNA samples were boiled at 95°C for 10 min for denaturation. Samples were immediately placed on ice for 5 min, and 250-500 ng were loaded on regular nitrocellulose membranes. DNA was then autocrosslinked in a UVC 500 crosslinker – 2 times at 1200uj (*100). The membranes were blocked for 1 hr in 5% PBST-BSA. Membranes were probed for 1 hr at room temperature or overnight at 4°C with 1:1000 dilution of 5mC antibody (OptimAbtm Anti-5-Methylcytosine – BY-MECY 100). 5mC signals were revealed using 1:5000 dilution of HRP conjugated antirabbit IgG secondary antibody for 1 hr at room temperature followed by chemo luminescence. Loading was measured using methylene blue staining.

RNA sequencing and Bisulfite sequencing

RNA sequencing was performed by the FASTERIS Company (https://www.fasteris.com/dna/).Total RNAs were polyA purified and libraries were prepared for illumina NextSeq sequencing technologies. Two technical replicates per biological samples were performed. A Pt18.6 line was sequenced in the same run as a control. Bisulfite libraries and treatments were performed by the FASTERIS Company and DNA was sequenced on an Illumina NextSeq instrument. 150bp Pair-end reads were generated with 30X coverage. A new 5mC map was also generated in the reference Pt18.6 line as a control.

Bisulfite sequencing analysis

Bisulfite analysis was performed using Bismark-bowtie 2 (babraham bionfinformatics). We used the default Bowtie2 implements of bismark with the specifications that only uniquely mapped reads should be aligned. We also clearly specified that no discordant pairs of the pair-end reads should be aligned.

R version and tools

Analysis and graphs were made using R version 4.0.3. DMR calling analysis was performed using the DMRcaller package v1.22.0. UpSet plots were computed using UpSetR v1.4.0. Genomic overlaps were computed using genomation v1.22.0

DMR calling

Differentially methylated regions were called using the DMRcaller Rpackage[38]. Given the low level of correlation of DNA methylation observed in *P. tricornutum*[11,24] and low sequencing coverage in Pt18.6 (supplementary Figure S3), only cytosines with >=5X coverage in all three lines were kept for further analysis (supplementary Figure S3) and the bins strategy was favored over other built-in DMRcaller tools. DMRs were defined as 100bp regions with at least an average 20% loss/gain of DNA methylation in either one of the DNMT5:KOs compared to the reference strain. The 'Score test' method was used to calculate statistical significance and threshold was set at p.value <0.01. In addition, to separate isolated differentially methylated cytosines from regions with significant loss of DNA methylations, each hypoDMRs must contain at least 4 methylated cytosines in the reference strain.

Overlap with histone post translational modifications and genomic annotations.

Percentage overlaps between DMRs as well as the overlap of gene and TEs coordinates with histone posttranslational modifications and DMRs were calculated using the genomation R package computation[63] and the 'annotateWithFeature' and 'getMembers' functions. DMRs were extended 500bp in 5' and 3' to account for promoter and regulatory region marking. For RNAseq analysis, we analyzed the expression of TEgenes previously defined[39]. To define TE genes in DMRs we crosscheck overlapping TEgenes annotations with *bona fide* TEs in DMRs using 'annotatewithFeature' function

RNAseq analysis

150bp pair-end sequenced reads were aligned on the reference genome of *P. tricornutum* (Phatr3) using STAR (v2.5.3a). Expression levels were quantified using HTseq v 0.7.2. Differentially expressed genes were analyzed using DESeq2 v1.19.37. FDR values are corrected P. values using the Benjamin and Hochberg method.

Legend

Figure 1

A. Convergent phylogenetic tree DNMT domains found in eukaryotes enriched with micro-eukaryote sequences of the MMETSP and reference genome databases. Numbers represent MrBAYES branching probability values for n=600000 generations. Grey branches represent bacterial and viral DCM enzymes. We indicate the main lineages found within each gene families using their corresponding colors next to the tree. **B**. Schematic representation of the DAMA/CLADE structure of representative DNMT enzymes. DNMT: DNA methyltransferase; RING: Ring zinc finger domain; DX : Dead box helicase; Hter : Cterminus-Helicase; LBR: Laminin B receptor; RFD: Replication Foci Domain; BAH: Bromo-Adjacent Homology; Chromo : Chromodomain; PHD: plant HomeoDomain

Figure 2

Lineage summary of DNMT families found in microeukaryotes. Plein crosses reports the presence of a given gene family within lineages. Dashed lines and crosses indicate the uncertainty in the eukaryotic phylogeny as well as low support presence of a given DNMT family within lineages. Fungi that share DNMT families with other eukaryotes presented in this study are shown for comparison purposes. SAR: Stramenopile Alveolate Rhizaria lineage. Ochrophyte are secondary endosymbiont, photosynthetic lineages of stramenopiles.

Figure 3

Alignment of the DNMT domain of representative DNMT5 enzymes. We labelled DNMT motifs using roman numbers. Motifs in brackets are divergent compared to other DNMTs. We propose an annotation for the motif I : TxCSGTD(A/S)P and IV : TSC; that are highly divergent compared to other DNMT motifs I (DXFXGXG) and IV (PCQ); based on their conservation in other DNMT5s and their position relatively to the other conserved DNMT motifs. Other motifs are well conserved and amino acid

with DNA binding and S-Adenosyl-Methionine binding activity are annotated accordingly.

Figure 4

Homozygous mutations generated by CRISPR/Cas9 in M23 and M25 lines at two independent target sequences. In M25, the mutation consists in 16 base pair out of frame deletion around CRISPR/Cas9 cutting sites that generates a loss of amino acids 28 to 34 leading to a premature STOP codon at amino acid 280. M23 has a 11 base pair out of frame deletion that generates a loss of amino acids 58 to 60/61 followed by a premature STOP codon at amino acid position 179-180 from ATG

Figure 5

A. Distribution of DNA methylation levels as a fraction of methylated reads per cytosines in the reference strain and in DNMT5:KOs. Levels are shown for n=10349 methylated cytosines found in the reference strain compared to their levels in DNMT5:KOs. The Pie chart show the distribution of CG methylation on genes, transposable elements (TEs) and their regulatory regions ‘reg.Genes’ and ‘reg.TEs’ defined as the 500bp upstream or downstream sequence of annotated coding sequences. Levels for non-CG methylation are also showed. **B.** Overlap between hypoDMRs and genomic features in DNMT5:KOs M23 and M25. Legend is as for **A.** **C.** Intersection of genes and TEs found in hypoDMRs and methylated genes and TEs in the reference strain identified in **A**. **D.** Association between CG hypomethylated regions in DNMT5:KO M23 and regions targeted by histone post-translational modifications representative of the epigenetic landscape of *P. tricornutum*. The number of overlapping hypoDMRs is showed for each histone modification and each combination of histone modifications.

Figure 6

A. Volcano plot of gene and TE genes overexpression in DNMT5:Kos M25 and M23 show upregulation of transposons. **B.** Overlap between gene and **C.** TEs overexpression between DNMT5:KOs.

Figure 7

A. Association between annotated transposable elements and regions targeted by histone post-translational modifications in the reference strain and regions with hypo methylation in DNMT5:KO M23. For each category, the total number of TEs is indicated. Proportion of TEs with consistent overexpression in both DNMT5:KOs is showed in pink. Red circles indicate regions overlapping with hypoDMRs in M23. We labelled combination of epigenetic landscape for which significant changes in expression are pictured in the following panel **B.** For epigenetic categories defined in **A.** We show distribution of expression of TE-genes as a log2 of fold change in RNA levels in DNMT5:KOs compared to the WT strain. Only fold change levels of TE-genes with a false discovery rate (FDR) <0.01 are used. We compared levels of expression between the different epigenetic categories within each mutant. Distributions sharing the same letters are statistically different based on the Mann-Withney U-test at p.value <0.01. Within categories and between mutants M23 showed lower induction level in the 'C' (H3K27me3 + H3K9me2 +H3K9me3) and 'E' (hypoDMRs+H3K27me3) categories.

Legend of supplemental information

Figure S1

Cladogram of the phylogenies of DNMTs (Figure 1a). D2 : DNMT2 family; D3 : DNMT3 family; D4 : DNMT4 family; D5(abc): DNMT5 subfamilies; D6: DNMT6 family; D1: DNMT1 family; CMT: chromomethylase family. Sequences are colored by lineage assignment. The phylogenetic tree depicted in figure 2 is also showed.

Figure S2

A. Quantitative PCR analysis of DNMT5 mRNA levels in the mutants compared to the reference Pt18.6 line (WT). Average fold loss is calculated by the ratio of CTs, normalized on the RPS and TUB genes (see material and methods), between mutants and WT. Normalized ratios were then averaged on n=2 biological replicates (*2 technical replicates per biological replicates) per lines for 6 primers targeting all the DNMT5 transcripts. Error bars represent the standard deviation between biological replicates. DNMT5:KO M26 is an independent DNMT5:KO line showing a deletion at the same position of DNMT5:KO M23 and is not further described in this manuscript **B**. as for A but compared to Cas9:Mock control line. **C.** Dot blot analysis of DNMT5 mutants compared to the Pt18.6 reference line (WT) and the Cas9:Mock control. 7C4

and 7C6 are DNMT5:KOs lines not further investigated in this study. We could not detect DNA methylation at comparable levels to the reference strain in any DNMT5:KO lines. **D.** As for C with serial dilutions of DNMT5:KOs M23 genomic DNA. Background level of DNA methylation are observed. Loading control is obtained by methylene blue staining.

Figure S3

A. And **B.** Intersection between genes and TEs found methylated in this study compared to previously identified methylated genes and TEs (Veluchamy et al. 2014, Nat Comm). **C**. Correlation of methylation levels for CG methylation in the reference Pt18.6 and DNMT5:KOs. DNA methylation levels show low correlation after a distance of 100bp in the reference strain suggesting a sparse methylation pattern as previously observed (see Huff and Zilberman 2014. Cell). No correlation is found in DNMT5:KOs. **D**. Coverage of cytosines after bisulfite treatment and illumina sequencing in Pt18.6 and M23/M25 show stronger cytosine sequencing coverage for mutants. Number of covered cytosines quickly drop off in the reference strain above 5X and this threshold was chosen for subsequent analysis. **E.** %overlap between DMRs identified in M23 and M25.

Figure S4

A. Intersection of genes found in hypoDMRs in both DNMT5:KOs and genes with >2 fold gain of mRNA levels in the respective mutants. **B.** As in A. for transposable elements overlapping with TE genes annotations **D. and E.** As for A. and B. but for loss of expression genes and TEs respectively.

Supplementary table S1

List of putative DNMTs fetched by HMMER in the MMETSP database ("HMMER hits") and reciprocal best hits search in the phylogenetically optimized database described in [31]. When already reported in other studies we added references. We also added enzymes found in reference database and literature in the tree for comparison purposes. References – MMETSP[64] Dorrell et al.[31]; De Mendoza et al. 2018[23]; GENOSCOPE (courtesy of GENOSCOPE sequencing project); Bewick et al. 2016[35]

Supplementary table S2

Representative microeukaryote enzymes used for phylogenetic analysis (Figure 1a) and their structure as given by eDAF pipeline.

Supplementary table S3

Species level summary of DNMTs found in microeukaryotes. Presence “1” or absence “x” of at least one DNMT transcript or gene per species.

Supplementary table S4

Summary of number of DNMTs per diatom species. Numbers indicate the number of putative paralogues for each gene family in each species. The common ancestor of diatoms likely possessed a DNMT3, DNMT6, DNMT5b and DNMT2 enzyme while DNMT5a is restricted to pennate diatoms. Iterative lineage specific loss of DNMTs is observed as between *Fistulifera solaris*, *Fragilariopsis cylindrus* and *P. tricornutum*. *P. tricornutum* (in red) is the reference species for epigenomics in diatoms and lack DNMT5b and DNMT4 orthologues. c- chromodomain containing-DNMT4 ; g – enzyme found in annotated genomes.

Supplementary table S5 and S6

Summary of DMR analysis in DNMT5:KOs M23 and M25 respectively. For each mutants, DMR coordinates, number of methylated reads/unmethylated reads and pooled %level of DNA methylation are showed. loss = hypoDMRs, gain = hyperDMRs

Supplementary table S7

Methylation of genes and transposable elements in the reference strain and overlap with hypoDMRs in M23/M25. For each gene and transposable element, the number of methylated cytosines and associated average DNA methylation level, methylated and unmethylated read counts are showed.

Supplementary table S8

Summary of RNAseq analysis.

Supplementary table S9

TE genes association with the epigenetic landscape of *P.tricornutum* in respect to upregulation in D5:Kos

Supplementary table 10

TopGO results of molecular function gene ontology enrichment of genes upregulated in both DNMT5:KOs. For each GO, the number of significant genes and the associated Fischer p.value is indicated. Only GOs with Fisher exact p.value <0.05 are shown.

Supplementary table S11

TopGO results of biological process gene ontology enrichment of genes upregulated in both DNMT5:KOs. For each GO, the number of significant genes and the associated Fischer p.value is indicated. Only GOs with p.value <0.05 are shown.

Supplementary table S12

Summary and list of upregulated genes in DNMT5:KOs.

Supplementary table S13

Summary and list of downregulated genes in DNMT5:KOs.

Acknowledgements

We thank Catherine Cantrel form IBENS for media preparation. LT acknowledges funds from the CNRS, the region of Pays de la Loire (ConnecTalent EPIALG project) and Epicycle ANR project (ANR-19-CE20- 0028-02). CB acknowledges funding from the ERC Advanced Award Diatomite.

Author Contributions

A.H., F.R.J.V and L.T. conceived and designed the study. LT supervised and coordinated the study. A.H. performed the experiments. O.A.M. performed the bioinformatic analysis of RNAseq, gene ontology and bisulfite data. A.H. analysed the bisulfite sequencing data and performed the DMR analysis. F.R.J.V and A.H analysed HMMER, DAMA/CLADE and eDAF data. All authors analysed and interpreted the data. A.H. and L.T. wrote the manuscript with input from all authors.

References

1. Kato, M., Miura, A., Bender, J., Jacobsen, S. E. & Kakutani, T. Role of CG and non-CG methylation in immobilization of transposons in Arabidopsis. *Curr. Biol.* (2003) doi:10.1016/S0960-9822(03)00106-4.

2. Zilberman, D. The evolving functions of DNA methylation. *Current Opinion in Plant Biology* (2008) doi:10.1016/j.pbi.2008.07.004.

3. Barlow, D. P. & Bartolomei, M. S. Genomic imprinting in mammals. *Cold Spring Harb. Perspect. Biol.* (2014) doi:10.1101/cshperspect.a018382.

4. Galupa, R. & Heard, E. X-Chromosome Inactivation: A Crossroads Between Chromosome Architecture and Gene Regulation. *Annu. Rev. Genet.* (2018) doi:10.1146/annurev-genet-120116-024611.

5. Bestor, T. H. DNA methylation: evolution of a bacterial immune function into a regulator of gene expression and genome structure in higher eukaryotes. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences* (1990) doi:10.1098/rstb.1990.0002.

6. Kumar, S. *et al.* The DNA (cytosine-5) methyltransferases. *Nucleic Acids Research* (1994) doi:10.1093/nar/22.1.1.

7. Cheng, X., Kumar, S., Klimasauskas, S. & Roberts, R. J. Crystal structure of the HhaI DNA methyltransferase. in *Cold Spring Harbor Symposia on Quantitative Biology* (1993). doi:10.1101/SQB.1993.058.01.039.

8. Zemach, A. & Zilberman, D. Evolution of eukaryotic DNA methylation and the pursuit of safer sex. *Current Biology* (2010) doi:10.1016/j.cub.2010.07.007.

9. Zemach, A., McDaniel, I. E., Silva, P. & Zilberman, D. Genome-wide evolutionary analysis of eukaryotic DNA methylation. *Science (80-.).* (2010) doi:10.1126/science.1186366.

10. Ponger, L. & Li, W. H. Evolutionary diversification of DNA methyltransferases in eukaryotic genomes. *Mol. Biol. Evol.* (2005) doi:10.1093/molbev/msi098.

11. Huff, J. T. & Zilberman, D. Dnmt1-independent CG methylation contributes to nucleosome positioning in diverse eukaryotes. *Cell* (2014) doi:10.1016/j.cell.2014.01.029.

12. Greenberg, M. V. C. & Bourc'his, D. The diverse roles of DNA methylation in mammalian development and disease. *Nature Reviews Molecular Cell Biology* (2019) doi:10.1038/s41580-019-0159-6.

13. Kouzminova, E. & Selker, E. U. Dim-2 encodes a DNA methyltransferase responsible for all known cytosine methylation in Neurospora. *EMBO J.* (2001) doi:10.1093/emboj/20.15.4309.

14. Dumesic, P. A., Stoddard, C. I., Catania, S., Narlikar, G. J. & Madhani, H. D. ATP Hydrolysis by the SNF2 Domain of Dnmt5 Is Coupled to Both Specific Recognition and Modification of Hemimethylated DNA. *Mol. Cell* (2020) doi:10.1016/j.molcel.2020.04.029.

15. Gladyshev, E. Repeat-Induced Point Mutation and Other Genome Defense Mechanisms in Fungi. *Microbiol. Spectr.* (2017) doi:10.1128/microbiolspec.funk-0042-2017.

16. Galagan, J. E. & Selker, E. U. RIP: The evolutionary cost of genome defense. *Trends in Genetics* (2004) doi:10.1016/j.tig.2004.07.007.

17. Yang, K. *et al.* The DmtA methyltransferase contributes to Aspergillus flavus conidiation, sclerotial production, aflatoxin biosynthesis and virulence. *Sci. Rep.* (2016) doi:10.1038/srep23259.

18. Malagnac, F. *et al.* A gene essential for de novo methylation and development in ascobolus reveals a novel type of eukaryotic DNA methyltransferase structure. *Cell* (1997) doi:10.1016/S0092-8674(00)80410-9.

19. Bewick, A. J., Vogel, K. J., Moore, A. J. & Schmitz, R. J. Evolution of DNA methylation across insects. *Mol. Biol. Evol.* (2017) doi:10.1093/molbev/msw264.

20. Feng, S. *et al.* Conservation and divergence of methylation patterning in plants and animals. *Proc. Natl. Acad. Sci. U. S. A.* (2010) doi:10.1073/pnas.1002720107.

21. Vanyushin, B. F. & Ashapkin, V. V. *DNA methylation in plants*. *DNA Methylation in Plants* (2011). doi:10.1146/annurev.arplant.49.1.223.

22. Zhong, X. *et al.* Molecular mechanism of action of plant DRM de novo DNA methyltransferases. *Cell* (2014) doi:10.1016/j.cell.2014.03.056.

23. De Mendoza, A. *et al.* Recurrent acquisition of cytosine methyltransferases into eukaryotic retrotransposons. *Nat. Commun.* (2018) doi:10.1038/s41467-018-03724-9.

24. Veluchamy, A. *et al.* Insights into the role of DNA methylation in diatoms by genome-wide profiling in Phaeodactylum tricornutum. *Nat. Commun.* (2013) doi:10.1038/ncomms3091.

25. Fan, X. *et al.* Single-base methylome profiling of the giant kelp Saccharina japonica reveals significant differences in DNA methylation to microalgae and plants. *New Phytol.* (2020) doi:10.1111/nph.16125.

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

27. Malviya, S. *et al.* Insights into global diatom distribution and diversity in the world's ocean. *Proc. Natl. Acad. Sci. U. S. A.* (2016) doi:10.1073/pnas.1509523113.

28. Traller, J. C. *et al.* Genome and methylome of the oleaginous diatom Cyclotella cryptica reveal genetic flexibility toward a high lipid phenotype. *Biotechnol. Biofuels* (2016) doi:10.1186/s13068-016-0670-3.

29. Lister, R. *et al.* Human DNA methylomes at base resolution show widespread epigenomic differences. *Nature* (2009) doi:10.1038/nature08514.

30. Zhao, X. *et al.* Probing the Diversity of Polycomb and Trithorax Proteins in Cultured and Environmentally Sampled Microalgae. *Front. Mar. Sci.* (2020) doi:10.3389/fmars.2020.00189.

31. Dorrell, R. G. *et al.* Chimeric origins of ochrophytes and haptophytes revealed through an ancient plastid proteome. *Elife* (2017) doi:10.7554/eLife.23717.

32. Bernardes, J. S., Vieira, F. R. J., Zaverucha, G. & Carbone, A. A multi-objective optimization approach accurately resolves protein domain architectures. *Bioinformatics* (2016) doi:10.1093/bioinformatics/btv582.

33. Bernardes, J., Zaverucha, G., Vaquero, C. & Carbone, A. Improvement in Protein Domain Identification Is Reached by Breaking Consensus, with the Agreement of Many Profiles and Domain Co-occurrence. *PLoS Comput. Biol.* (2016) doi:10.1371/journal.pcbi.1005038.

34. Bewick, A. J. *et al.* Diversity of cytosine methylation across the fungal tree of life. *Nat. Ecol. Evol.* (2019) doi:10.1038/s41559-019-0810-9.

35. Bewick, A. J. *et al.* The evolution of CHROMOMETHYLASES and gene body DNA methylation in plants. *Genome Biol.* **18**, (2017).

36. Jurkowska, R. Z., Jurkowski, T. P. & Jeltsch, A. Structure and Function of Mammalian DNA Methyltransferases. *ChemBioChem* (2011) doi:10.1002/cbic.201000195.

37. Veluchamy, A. *et al.* An integrative analysis of post-translational histone modifications in the marine diatom Phaeodactylum tricornutum. *Genome Biol.* (2015) doi:10.1186/s13059-015-0671-8.

38. Catoni, M., Tsang, J. M. F., Greco, A. P. & Zabet, N. R. DMRcaller: A versatile R/Bioconductor package for detection and visualization of differentially methylated regions in CpG and non-CpG contexts. *Nucleic Acids Res.* (2018)

doi:10.1093/nar/gky602.

39. Rastogi, A. *et al.* Integrative analysis of large scale transcriptome data draws a comprehensive landscape of Phaeodactylum tricornutum genome and evolutionary origin of diatoms. *Sci. Rep.* (2018) doi:10.1038/s41598-018-23106-x.

40. Alexa, A. & Rahnenführer, J. Gene set enrichment analysis with topGO. *Bioconductor Improv.* (2007).

41. Ait-Mohamed, O. *et al.* PhaeoNet: A Holistic RNAseq-Based Portrait of Transcriptional Coordination in the Model Diatom Phaeodactylum tricornutum. *Front. Plant Sci.* (2020) doi:10.3389/fpls.2020.590949.

42. Stroud, H. *et al.* Non-CG methylation patterns shape the epigenetic landscape in Arabidopsis. *Nat. Struct. Mol. Biol.* (2014) doi:10.1038/nsmb.2735.

43. Cuypers, B. *et al.* C-5 DNA methyltransferase 6 does not generate detectable DNA methylation in Leishmania. *bioRxiv* (2019) doi:10.1101/747063.

44. Grüne, T. *et al.* Crystal structure and functional analysis of a nucleosome recognition module of the remodeling factor ISWI. *Mol. Cell* (2003) doi:10.1016/S1097-2765(03)00273-9.

45. Lu, R. & Wang, G. G. Tudor: A versatile family of histone methylation 'readers'. *Trends in Biochemical Sciences* (2013) doi:10.1016/j.tibs.2013.08.002.

46. Pek, J. W., Anand, A. & Kai, T. Tudor domain proteins in development. *Dev.* (2012) doi:10.1242/dev.073304.

47. Tóth, K. F., Pezic, D., Stuwe, E. & Webster, A. The pirna pathway guards the germline genome against transposable elements. in *Advances in Experimental Medicine and Biology* (2016). doi:10.1007/978-94-017-7417-8_4.

48. Bostick, M. *et al.* UHRF1 plays a role in maintaining DNA methylation in mammalian cells. *Science (80-.).* (2007) doi:10.1126/science.1147939.

49. Ooi, S. K. T. *et al.* DNMT3L connects unmethylated lysine 4 of histone H3 to de novo methylation of DNA. *Nature* (2007) doi:10.1038/nature05987.

50. Jackson, J. P., Lindroth, A. M., Cao, X. & Jacobsen, S. E. Control of CpNpG

DNA methylation by the KRYPTONITE histone H3 methyltransferase. *Nature* (2002) doi:10.1038/nature731.

51. Honda, S. & Selker, E. U. Direct Interaction between DNA Methyltransferase DIM-2 and HP1 Is Required for DNA Methylation in Neurospora crassa. *Mol. Cell. Biol.* (2008) doi:10.1128/mcb.00823-08.

52. Estève, P. O. *et al.* Direct interaction between DNMT1 and G9a coordinates DNA and histone methylation during replication. *Genes Dev.* (2006) doi:10.1101/gad.1463706.

53. Ikegami, K. *et al.* Genome-wide and locus-specific DNA hypomethylation in G9a deficient mouse embryonic stem cells. *Genes to Cells* (2007) doi:10.1111/j.1365-2443.2006.01029.x.

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

55. Rougée, M. *et al.* Polycomb mutant partially suppresses DNA hypomethylation-associated phenotypes in Arabidopsis. *Life Sci. Alliance* (2021) doi:10.26508/LSA.202000848.

56. Walter, M., Teissandier, A., Pérez-Palacios, R. & Bourc'His, D. An epigenetic switch ensures transposon repression upon dynamic loss of DNA methylation in embryonic stem cells. *Elife* (2016) doi:10.7554/eLife.11418.001.

57. Maumus, F. *et al.* Potential impact of stress activated retrotransposons on genome evolution in a marine diatom. *BMC Genomics* (2009) doi:10.1186/1471-2164-10-624.

58. Rogato, A. *et al.* The diversity of small non-coding RNAs in the diatom Phaeodactylum tricornutum. *BMC Genomics* (2014) doi:10.1186/1471-2164-15-698.

59. Madeira, F. *et al.* The EMBL-EBI search and sequence analysis tools APIs in 2019. *Nucleic Acids Res.* (2019) doi:10.1093/nar/gkz268.

60. Miller, M. A., Pfeiffer, W. & Schwartz, T. Creating the CIPRES Science Gateway for inference of large phylogenetic trees. in *2010 Gateway Computing*

Environments Workshop, GCE 2010 (2010). doi:10.1109/GCE.2010.5676129.

61. Nymark, M., Sharma, A. K., Sparstad, T., Bones, A. M. & Winge, P. A CRISPR/Cas9 system adapted for gene editing in marine algae. *Sci. Rep.* (2016) doi:10.1038/srep24951.

62. Rastogi, A., Murik, O., Bowler, C. & Tirichine, L. PhytoCRISP-Ex: A web-based and stand-alone application to find specific target sequences for CRISPR/CAS editing. *BMC Bioinformatics* (2016) doi:10.1186/s12859-016-1143-1.

63. Akalin, A., Franke, V., Vlahoviček, K., Mason, C. E. & Schübeler, D. Genomation: A toolkit to summarize, annotate and visualize genomic intervals. *Bioinformatics* (2015) doi:10.1093/bioinformatics/btu775.

64. Keeling, P. J. *et al.* The Marine Microbial Eukaryote Transcriptome Sequencing Project (MMETSP): Illuminating the Functional Diversity of Eukaryotic Life in the Oceans through Transcriptome Sequencing. *PLoS Biol.* (2014) doi:10.1371/journal.pbio.1001889.