



HAL
open science

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

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

► **To cite this version:**

Antoine Hoguein, Ouardia Ait, Chris Bowler, Auguste Genovesio, Fabio Rocha Jimenez Vieira, et al..
Evolutionary analysis of DNA methyltransferases in microeukaryotes: Insights from the model diatom
Phaeodactylum tricornutum. BioRxiv, In press. hal-03388012

HAL Id: hal-03388012

<https://hal.science/hal-03388012>

Submitted on 20 Oct 2021

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

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

3

4 Antoine Hoguein¹, Ouardia Ait Mohamed^{1†}, Chris Bowler¹, Auguste Genovesio¹, Fabio
5 Rocha Jimenez Vieira^{1‡*} and Leila Tirichine^{2*}

6

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

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

10

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

13 ‡Current affiliation: Medical Affairs Department, AFM-Téléthon, Evry, France

14

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

16

17

18

19

20

21

22

23

24

25

26 **Abstract**

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

44

45

46 **Background**

47 In eukaryotes the methylation of the fifth carbon of cytosines (5mC) is a
48 recognized epigenetic mark involved in chromosome X inactivation, genomic
49 imprinting, stability of repeat rich centromeric and telomeric regions as well as
50 transposable elements (TEs) repression¹⁻⁴. 5mC is deposited by DNA
51 methyltransferases (DNMTs) capable of *de novo* methylation and is propagated
52 through subsequent cell division by maintenance enzymes. Eukaryotes have acquired
53 a diverse set of DNA methyltransferases (DNMTs) by horizontal gene transfer of
54 bacterial DNA Cytosine methyltransferase (DCM) enzymes involved in the Restriction
55 Methylation system⁵. All DNMTs are composed of a catalytic protein domain composed
56 of ten conserved motifs (annotated I to X) that provide binding affinity to DNA substrate
57 and the methyl donor cofactor S-Adenosyl methionine (SAM) to process the transfer
58 of a methyl group to unmethylated cytosines^{6,7}. DNMTs have further diversified over
59 evolutionary time scales in eukaryote lineages and acquired chromatin associated
60 recognition and binding domains giving rise to a wide diversity of DNA methylation
61 patterning^{8,9}. The loss and gain of DNMTs have been associated with profound
62 divergence in cell biology and in the control of gene expression. To date, six main
63 eukaryotic DNMT families have been described and named DNMT1, DNMT2, DNMT3,
64 DNMT4, DNMT5 and DNMT6^{10,11}. In metazoans, the concert activity of the DNMT3
65 family and DNMT1 enzymes allow the deposition and the maintenance of DNA
66 methylation pattern during the successive developmental waves of DNA demethylation
67 and remethylation¹². In fungi, the DNA methylation machinery consists in a
68 maintenance activity by DNMT1/DIM2, as in *Neurospora crassa*¹³, or by the activity of
69 ATPase-DNMT5 enzymes as reported in *Cryptococcus neoformans*^{11,14}. Fungal
70 DNMT4 relatives are involved in the DNA methylation related process known as
71 Repeat-Induced Point Mutation (RIP) and Methylation Induced Premeiotically (MIP)
72 that leads to transposable elements extinction and or stage specific repression as
73 observed in *Aspergillus* and *Neurospora* species¹⁵⁻¹⁸. Successive gains and loss of
74 DNMT1 and DNMT3 occurred during insect evolution leading to secondary loss of
75 global 5mC methylation as in diptera lineages¹⁹. In plants, the acquisition of new
76 DNMT1 proteins named Chromomethylases (CMTs) and the divergence of the DNMT3
77 family led to the spreading of the asymmetrical non-CG patterns of DNA methylation
78 that is extensively found in angiosperms²⁰⁻²². In *Micromonas pusilla*, the pelagophyte
79 *Aureococcus anophagefferens* and the haptophyte *Emiliania huxleyi*, the presence of

80 a DNMT5 enzyme correlates with a heavy histone linker DNA methylation landscape¹¹.
81 Importantly, 5mC have been now reported in more diverse eukaryotes of the SAR
82 lineages as in dinoflagellates²³, diatoms²⁴ and kelps²⁵. However, due to the severe
83 underrepresentation of marine unicellular eukaryotes in modern sequencing
84 databases, our understanding of the DNA methylation machinery in microalgae
85 remains scarce.

86 Diatoms are a dominant, abundant, and highly diverse group of unicellular
87 brown micro-algae (from 2 to 200 µm) of the stramenopile lineage. It is estimated that
88 diatoms are responsible for nearly 20% of earth primary production^{26,27}. They are
89 known to dominate marine polar areas and are major contributors of phytoplankton
90 oceanic blooms. To date, 5mC has been reported in four diatoms, namely the centrics
91 *Thalassiosira pseudonana*¹¹ and *Cyclotella cryptica*²⁸, as well as in *Fragilariopsis*
92 *cylindrus*¹¹ and *Phaeodactylum tricornutum*^{11,24}, a polar and temperate pennate
93 diatom respectively. Diatoms methylation pattern strongly contrasts with the patterns
94 observed in animals but also dinoflagellates and plants²⁹. Firstly in *P. tricornutum*, *T.*
95 *pseudonana* and *F. cylindrus*, total levels of DNA methylation range from 8% to as low
96 as 1% of Cytosines in the CG context¹¹ over repeats and transposable elements
97 usually (but not exclusively) concentrated in telomeric regions^{11,24}. Non-CG
98 methylation is also detected but is scarce. Diatom genomes are thus mainly composed
99 of isolated highly CG methylated TE islands in an otherwise unmethylated genome and
100 to that regards are remarkably similar to fungi methylation profiles. In all diatoms
101 studied so far, methylated TEs often have low expression^{11,24,28}. This is very consistent
102 with the repressive role of DNA methylation in other eukaryotes and further traces back
103 5mC mediated control of TE expression to the last eukaryotic common ancestor.
104 Nonetheless, direct evidence of the repressive role of 5mC at transposable elements
105 in diatoms is lacking.

106 The recent advance in high throughput RNA sequencing technologies applied
107 to microeukaryotes led to the development of the Microbial Eukaryote Transcriptome
108 Sequencing Project (MMETSP) consortium. The MMETSP concatenates more than
109 650 transcriptomes from diverse microeukaryote lineages as diatoms and
110 dinoflagellates making it the biggest sequence database for marine microeukaryotes.
111 Here using the newly defined enhanced Domain Architecture Framework (eDAF)
112 methodology³⁰, we first explored the structural and phylogenetic diversity of DNMTs

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

128

129 Results

130 1. Diversity of DNA methyltransferases in microeukaryotes

131 In order to capture the diversity of 5-cytosines DNA methyltransferases encoded
132 in microalga, we applied a relaxed HMMER search (e-value=1 as the cut-off threshold)
133 for the PFAM DNMT (PF00145) domain on the whole transcriptomes of the MMETSP
134 database. We complemented MMETSP with 7 diatoms transcriptomes and genomes
135 from the top 20 most abundant diatoms found in TARA Oceans database (courtesy of
136 the GENOSCOPE sequencing project). Next, we eliminated the expected high number
137 of false positives using eDAF³⁰. We also performed reciprocal BLAST best hit analysis
138 on a phylogenetically optimized genomic and transcriptomic database as previously
139 described³¹. We kept enzymes that showed conserved DNMT domains and depicted
140 their domain structures by eDAF curation³⁰. We build a representative phylogeny of
141 DNA methyltransferases based on the alignment of DNMT motifs of conserved DNMTs
142 (Figure 1a, supplementary Figure 1, supplementary table S1). We also compiled the
143 observed structure of microalgal enzymes based on the analysis of DAMA³²/CLADE³³
144 guided eDAF annotation of representative enzymes used for phylogeny (Figure 1b,
145 Supplementary table S2). In this study we focused on the DNMT1, DNMT3, DNMT4,
146 DNMT5 and DNMT6 gene families that are known or putative DNA modifying enzymes.
147 Within lineages, a species level summary of presence absence of each gene family
148 can be found in supplementary table S3.

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

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

167 We found 76 species with at least one DNMT5 orthologue. This includes green
168 algal DNMT5s as in *Tetraselmis astigmatica* that are similar to the published
169 *Micromonas pusilla* enzyme (Figure 1b, supplementary table S2). Green algal
170 DNMT5s are found within the DNMT5a subfamily (supplementary figure S1). We also
171 found a DNMT5a transcript in the marine photosynthetic excavate euglenozoa
172 *Eutreptiella gymnastica* closely related to green algal DNMT5as (Figure 2,
173 supplementary Figure S1, supplementary table S3). The DNMT5 family is also
174 widespread in diatom species in which DNMT5a enzymes are mainly found in pennate
175 genomes while DNMT5b is found in both pennates and centrics. Interestingly, our data
176 do not strongly support the presence of the DNMT5b family outside diatoms (Figure
177 2). *Emiliania huxleyi* has two annotated DNMT5s, one DNMT5a and one DNMT5b
178 enzyme, but we could not find DNMT5 enzymes in other haptophytes of the MMETSP
179 database suggesting horizontal acquisition of this enzyme family. DNMT5 transcripts
180 identified in dinoflagellates transcriptomes fall into the DNMT5c family, and
181 reciprocally, no DNMT5c enzymes are found in other lineages (Figure 2,
182 supplementary table S3). Overall, our data support the absence of SNF2 containing
183 DNMT5s in dinoflagellates with a few exceptions of partial DNMT5a/b enzymes
184 (supplementary figure 1, supplementary table S2). Within DNMT5a, we identify at least
185 3 monophyletic group of enzymes: diatoms DNMT5as, Pelagophyte/Dicthyochophyte
186 DNMT5as and green algal DNMT5as. However, fungal enzymes are not monophyletic,
187 and we failed to recover monophyletic relationship between ochrophyte stramenopiles
188 suggesting a complex phylogenetic history of DNMT5 subfamilies. In addition, DNMT5
189 is absent from non-photosynthetic stramenopiles such as labyrinthumycetes and
190 raphidophytes (Figure 2).

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

194 adjacent homology (BAH) domains, Plant Homeo Domains (PHDs), chromodomains
195 and domains required for the interaction with accessory proteins. We isolated
196 DNMT1/MET1 transcripts similar to the plant MET1 enzyme in 7 green algae species
197 of the MMETSP database as in *Chlamydomonas* species (Figure 1b, Figure 2,
198 supplementary Figure 1, supplementary table S3), reaffirming that the DNMT1 family
199 was acquired early during plant evolution. CMT enzymes are divergent DNMT1 related
200 enzymes that likely appeared in the very first plant lineages. Its conservation in green
201 algae is more contrasted as many partial CMT enzymes as well as misannotated
202 DNMT1 transcripts were found in early reports³⁵. Accordingly, we could not find *bona*
203 *fide* CMT transcripts in green algae species of the MMETSP data base and instead
204 found two partial CMT transcripts in *Polytomella parva* and *Chlamydomonas*
205 *chlamydogama* that lack BAH, PHD and chromodomains (supplementary Figure 1,
206 supplementary table S2). Interestingly, we found a DNMT1 related enzyme in three
207 haptophytes species (*Gephyrocapsa oceanica*, *Isochrysis.sp-CCMP1324* and
208 *Coccolithus pelagicus*) of the MMETSP database that cluster with annotated CMTs
209 found in the coccolithophore *Emiliana huxleyi* (Figure 1a, supplementary Figure 1).
210 Haptophytes enzymes distantly relate to the conserved green algae CMTs (hCMT2)
211 enzymes (Figure 1a, supplementary Figure S1). We found that the enzymes of
212 *Gephyrocapsa oceanica* (CAMPEP_0188208858), *Isochrysis-CCMP1324*
213 (CAMPEP_0188844028) and *Emiliana huxleyi* (jgi_458278) have DNMT1-like
214 structures with a Replication Foci Domain (RFD) followed by a BAH (in *Emiliana*
215 *huxleyi* only) and a conserved DNMT domain (Figure 1b, supplementary table S2). We
216 also found two putative DNMT1-like enzymes in the transcriptomic database of
217 *Raphidophyceae* brown microalgae *Heterosigma akashiwo* and *Chatonella subsala*.
218 They are composed of a conserved DNMT domain and a plant homeodomain (PHD)
219 (Figure 1b, supplementary figure 1, Figure 2, supplementary table S2). Raphidophycea
220 enzymes poorly define a DNMT family by themselves and distantly relate to the
221 DNMT1 enzymes (Figure 1a, Supplementary Figure 1). We could not find compelling
222 evidence for *bona fide* DNMT1s in any other microalgae. Our data support the
223 hypothesis that the DNMT1 maintenance machinery is absent from the main
224 microeukaryote lineages that therefore rely on DNMT1-independent mechanisms for
225 the maintenance of their DNA methylation pattern as previously suggested¹¹.

226 DNMT4 enzymes have been identified in fungi and are involved in the MIP or
227 RIP processes. Closely related to DIM2 enzymes in fungi³⁴; two DNMT4 enzymes were
228 also found in the pennate diatom *F. cylindrus* and the centric diatom *T. pseudonana*¹⁰.
229 Accordingly, we identified a monophyletic diatom enzyme family that relate to the
230 fungal DNMT4 gene family (Figure 1a, Supplementary Figure 1). Diatoms and
231 RID/DMTA enzymes as well as DNMT1 proteins form a moderately supported
232 monophyletic clade. A total of 31 diatoms, pennate and centric species expresses or
233 encodes at least one DNMT4 related transcript (supplementary table S4). This
234 indicates that the DNMT4 family was acquired early during diatoms evolution. Most
235 diatoms DNMT4 enzymes are composed of a single DNMT domain as in *T.*
236 *pseudonana*, which contrasts with fungal enzymes (Figure 1b – supplementary table
237 S2). However, 9 diatom's DNMT4 proteins possess an additional N-ter chromodomain
238 as observed in *Thalassiosira miniscula* (Fig. 1b, supplementary table S2 and S4). We
239 found the microalgal DNMT4 family is restricted to diatoms. The DNMT4 family is thus
240 only observed in two very distantly related microeukaryote lineages questioning the
241 evolutionary origin of this gene family in eukaryotes.

242 Our data indicate that the DNMT3 family is not particularly abundant in
243 microalgae (Figure 2, supplementary table S3). DNMT3 is absent in most
244 stramenopiles except for diatoms; in which both genomic and transcriptomic data
245 strongly support its presence (supplementary table S4). DNMT3 seems absent in
246 haptophytes (Figure 2, supplementary table S3). Only one transcript from the
247 cryptomonad *Goniomonas pacifica* could be annotated as DNMT3. In addition, we
248 could not identify DNMT3 enzymes in any green algae of the MMETSP database, but
249 it is present in red algae as found in the genomes of *Cyanidioschyzon merolae* and
250 *Galdieria sulphuraria* (Figure 2, supplementary table S3). We also report few additional
251 DNMT3 transcripts in dinoflagellates as previously described²³ (Figure 2,
252 supplementary table S3). Upon alignment, dinoflagellates DNMT3 enzymes (including
253 former annotated enzymes²³) and *Goniomonas pacifica* DNMT3s are closely related
254 to red algal genes but diverged compared to other DNMT3s while diatoms form their
255 own DNMT3 family (Supplementary figure 1). This suggest that the DNMT3 family was
256 iteratively lost and acquired several times during microalgae evolution. As observed in
257 *P. tricornutum*, DNMT3 enzymes found in microalgae all lack chromatin associated

258 domains (Figure 1b, supplementary table S2). This contrasts with mammalian
259 DNMT3s³⁶ that interact with histone post-translational modifications.

260 We found that the DNMT6 enzymes are among the most widespread DNMTs in
261 microeukaryotes. We found a DNMT6 transcript in the MMETSP transcriptomes of 3
262 *Tetraselmis* green algae and 7 dinoflagellates (Figure 2 – supplementary table S3). In
263 addition, DNMT6 is extensively found in stramenopiles as in *Dictyochophyceae*,
264 *Crysophyceae* and *Pelagophyceae* (Figure 2, supplementary table S3). In diatoms,
265 DNMT6 is very abundant (Supplementary table S4). DNMT6 is also present in the non-
266 photosynthetic labyrinthulomycetes *Aplanochytrium stocchinoi* and probably in
267 *Aplanochytrium keurgelense* (Figure 2, supplementary table S3). In addition, our data
268 also strongly support the presence of DNMT6 orthologue in the major *Chromalveolata*
269 lineage of *Rhizaria* (Figure 2, supplementary table S3) as suggested in previous
270 reports²³. DNMT6 enzymes are for most part homogeneous and do not contain
271 chromatin associated signatures, as in *P. tricornutum* DNMT6 and DNMT3 enzymes
272 (Figure 1b, supplementary table S2). Monophyletic relationships within the DNMT6
273 family and between micro-eukaryotes could not be resolved (supplementary figure 1).

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

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

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

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

295 To get a quantitative single base resolution of DNA methylation loss in
296 DNMT5:KOs, we performed whole genome bisulfite sequencing on the M23, M25 and
297 the reference Pt18.6 line. To assess highly methylated regions, we first filtered
298 cytosines by coverage and methylation levels considering a 5X coverage threshold
299 and 60% methylation levels in all lines. In the reference Pt18.6, we identified 10349
300 methylated CGs (0.3% of all analyzed CGs), 129 methylated CHHs (3/10000) and 12
301 methylated CHGs ($<5/10^6$). About 76% of methylated cytosines are found within TEs
302 confirming that repeated sequences are the main targets of CG methylation in *P.*
303 *tricornutum* genome (Figure 5a). In comparison only 9% of methylated cytosines
304 overlap with genes (Figure 5a). As a result, 876/3406 (25%) of transposable elements
305 and 170/9416 (1,8%) of genes are marked by at least one methylated CG in Pt18.6.
306 Among methylated TEs, 711 (81%) were also found methylated in the reference strain
307 in earlier reports²⁴ and reciprocally (supplementary Figure 3). Gene methylation was
308 less consistent between already published reports²⁴ as 230 (63%) of known methylated
309 genes are not found methylated in the present study (supplementary Figure 3), which
310 might be due to differences in culture conditions and/or coverage depths and methods
311 used to profile DNA methylation. In DNMT5:KOs, a maximum of 2 cytosines in any
312 context are found methylated. Methylated cytosines identified in the reference strain
313 show background levels of DNA methylation in DNMT5:KOs (Figure 5 a). The knockout
314 of DNMT5 thus generated a global loss of all cytosine methylation in *P. tricornutum*.

315 We asked whether the extensive loss of DNA methylation in DNMT5:KOs could
316 define differentially methylated regions (DMRs) and computed DMRs between
317 DNMT5:KOs and WT lines using the bins built-in DMRcaller³⁸ tools (material and
318 method). DMRs corresponding to hypo methylated regions in DNMT5:KOs cover
319 ~0.8% of *P. tricornutum* genome. In addition, 96% of DMRs found in M23 and M25 are
320 shared while about 4% of DMRs are found in one mutant only, showing the consistent
321 loss of DNA methylation in independent cell lines upon knockout of DNMT5
322 (supplementary Figure S3). We found that ~83% of DMRs overlap with transposable

323 elements and their regulatory regions (~5%) (Figure 5 b). Reciprocally, 584 (~15%)
324 annotated transposable elements are found within DMRs (+/-500bp) in DNMT5:KOs
325 lines (Figure 5 c, Supplementary table 7). Among them 544 showed at least one highly
326 methylated CG (Figure 5c). We found that 70% of all methylated cytosines in the
327 reference strain concentrate in TEs that overlap with DMRs and 90% of all methylation
328 found at TEs are found within DMR-TEs (supplementary figure S3). Outside DMRs,
329 methylated TEs show less than 2 methylated cytosines while within DMRs TEs show
330 extensive methylation (supplementary Figure 3). Our DMR analysis thus identified a
331 significant loss of DNA methylation at transposons that also concentrate most of
332 densely methylated cytosines in *P. tricornutum*. There is a low overlap of DMRs with
333 genes or their regulatory regions and only ~0.5% of all protein coding genes (30% of
334 genes with at least one highly methylated CGs) are found within DMRs (Figure 4c).
335 Nonetheless, genes within DMRs concentrate more than 80% of methylation found in
336 genes in the reference strain. As for TEs, genes identified in hypoDMRs are also
337 among the most methylated genes in the reference strain. We then asked whether
338 DMRs associate with known regions marked by histone posttranslational
339 modifications. Interestingly, between 80% and 90% of all DMRs overlap with known
340 regions marked by H3K27me3, H3K9me3 or H3K9me2 defined in the reference Pt18.6
341 line³⁷ (Figure 5d). We found that 70% of DMRs are found within regions co-marked by
342 all three repressive histone marks in association or not with active marks. Our DMR
343 analysis thus retrieved the denser and highly heterochromatic regions of *P. tricornutum*
344 genome that are reported targets of CG methylation in the reference strain. Our data
345 are consistent with a global loss of DNA methylation in DNMT5:KOs at TE-rich DNA
346 methylated-H3K27me3, H3K9me2 and H3K9me3 rich regions of *P. tricornutum*
347 genome. Of note, one hypermethylated region was found in the mutant M23 in the
348 promoter region of the gene Phatr3_J46344. Consistent with the low level of CHH and
349 CHG methylation in the reference strain we found no CHG and CHH hypo nor hyper
350 DMRs.

351 The control of transposable elements by the DNA methyltransferases family is
352 a key unifying feature within eukaryotes². We hence monitored the transcriptional
353 effect of the loss of DNMT5 in M23/M25 by whole RNA high throughput sequencing
354 (Material and Methods). To identify the most significant changes in mRNA levels, we
355 focused our analysis on genes showing a significant 2 folds induction or reduction of

356 expression in mutants compared to the reference line (LFC > |1| and an FDR < 0.01).
357 We also analyzed expression of genes overlapping with TE annotations (TE-genes) as
358 previously described³⁹. In both DNMT5:KOs, we observed a global overexpression of
359 genes and TE-genes and comparatively fewer downregulation (Figure 6a). TE-genes
360 show higher level of overexpression than genes that show milder inductions
361 suggesting that primary targets of DNA methylation in the reference strains are also
362 the first to respond to the loss of DNA methylation in DNMT5:KOs. Gene
363 overexpression is inconsistent between lines (Figure 6b). At the transcriptional level,
364 M23 shows up to 1133 up regulated genes and M25 shows 393 upregulated genes.
365 Only 219 genes have consistent upregulation in both mutants. Transposable element
366 showed more overlap. We found that 339 TE-genes with up-regulation in at least one
367 mutant are consistently up-regulated in both (Figure 6c). We performed GO term
368 enrichment using TopGO tools⁴⁰ and found that overexpressed genes in both mutants
369 are enriched for GOs associated with protein folding responses as well as nucleotide
370 phosphate metabolism and nucleotide binding activity. This is typified by the
371 overexpression of chaperone DnaJ domain containing proteins and Hsp90 like proteins
372 (Supplementary tables S10 and S11). The recent advances in high through output
373 RNA sequencing technologies led to the recent identification of gene co-regulatory
374 networks in *P. tricornutum*⁴¹. Genes with overexpression in DNMT5:KOs are enriched
375 within the “salmon” and “steelblue” coregulatory gene modules. The ‘salmon’ module
376 contains genes associated with chloroplast biology as well as amino acid related
377 enzymes which might highlight metabolic defects. The ‘steelBlue’ regulatory module
378 associates with protein translation as well as nucleotide metabolism and is enriched in
379 genes involved in purine and pyrimidine cellular pathways⁴¹. Accordingly, some
380 upregulated proteins are involved in ribosome biosynthesis and RNA processing. The
381 chaperone like enzymes found upregulated in DNMT5:KOs are part of both the
382 ‘salmon’ and ‘steelblue’ pathway (supplementary table S12). Only 36 genes showed
383 consistent down-regulation in both lines in a DNA methylation independent manner
384 (Supplementary figure 4). Among those genes we found back DNMT5a, that shows a
385 3-fold loss in both mutants confirming qPCR analysis and suggesting that premature
386 stop codons triggered mRNA degradation (Supplementary table S13). We could not
387 find compelling GO enrichment nor coregulation network enrichment for
388 downregulated genes in DNMT5:KOs. A handful of TE genes show loss of expression

389 in DNMT5:KOs that is not associated with DNA methylation loss (Supplementary figure
390 4).

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

415 **Discussion**

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

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

430 In our phylogeny, the RID/DMTA and the diatoms DNMT4 enzymes are clearly
431 related as shown by Huff and Zilberman¹¹ and Pungler and Li¹⁰. Within fungi, RID
432 proteins were found closely related to DIM2 enzymes³⁴ for which we have not found
433 any related orthologue in diatoms. We make the hypothesis that the DNMT4 family is
434 closely related to DNMT1 enzymes and arose by convergence in the diatom and fungi
435 lineages. The function of DNMT4 in diatoms is unknown. Whether RID or MIP process
436 occur in any DNMT4 containing diatom also remain an open question. The presence
437 of chromodomains that are known to bind histone post translational modifications as
438 in CMT enzymes^{1,42} nonetheless suggests that diatoms DNMT4 might be functional as
439 either *de novo* or maintenance enzymes.

440 DNMT6 enzymes are very well detected in diatoms and dinoflagellates. Our
441 study independently reports the same DNMT6 enzymes found in *Bigelowella natans*
442 and *Aplanochytrium stochhinoi* by earlier works although not specified by the authors²³.
443 As it is reported in trypanosomes¹⁰, we suggest that DNMT6 likely emerged prior to the
444 *Chromalveolata* radiation. It is the only known DNMT enzyme in Trypanosomes that
445 does not show extensive DNA methylation pattern in any context⁴³. DNMT6 is not
446 active in *Leishmania* species⁴³. Its presence in several lineages, hence, does not
447 predict DNA methylation *per se* and must be further investigated.

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

453 subfamilies apart of their DNMT domain. It was established that DNMT5s rely on the
454 ATPase activity of their SNF2 like domains to methylate hemi-methylated cytosines¹⁴.
455 In that regards DNMT5c that are also dinoflagellate specific, lacks ATPases. The
456 DNMT5 family also likely diverged in diatom species as DNMT5b enzymes specially in
457 centric diatoms. DNMT5b might be multifunctional related DNMT5 enzymes as it is
458 suggested by the presence of N-ter HAND domains, which are found in chromatin
459 remodelers⁴⁴, as well as TUDOR domains that are found in histone modifying enzymes
460 and readers⁴⁵ and small RNA interacting proteins^{46,47}. In the genomes of *Fragilariopsis*
461 *cylindrus* species as well as in *Synedra* both DNMT5a and b are found. This supports
462 a scenario in which both DNMT5a and b were acquired early in diatoms and might be
463 paralogues. In other diatoms, DNMT5a is however phylogenetically restricted. It is
464 almost exclusively found in annotated genomes of pennate diatoms. *Corethron*
465 *pennatum* is the only centric diatom with a DNMT5a transcript. Besides transcriptomic
466 evidence, the absence of DNMT5a in centric diatoms is further supported by genomic
467 data in 5 centric species in which at least another DNMT could be detected, especially
468 DNMT5b (table S4). The diatoms DNMT5a enzymes might originate from a pennate
469 specific horizontal gene transfer with another stramenopiles encoding DNMT5a and or
470 within diatoms themselves.

471 The independent accession of DNMT5s in ochrophyte algae is supported by our
472 phylogenies in which diatoms and pelagophytes/dictyochophytes enzymes are not
473 monophyletic. DNMT5 enzymes are thus related gene families that show lineage
474 specific divergence. The origin of the DNMT5c is also intriguing since it is both highly
475 divergent compared to other DNMT5s and dinoflagellate specific. In DNMT5 enzymes,
476 we noticed that the DNMT motifs I; involved in the binding of S-Adenosyl-methionine;
477 and IV; the catalytic site; are highly divergent compared to other DNMTs suggesting
478 that DNMT5 enzymes might require accessory proteins aside of their ATPase activity.
479 In addition, the lack of chromatin associated domains in DNMT3 proteins, but also
480 DNMT6 and DNMT4 proteins suggest that the link, if any, between DNA methylation
481 and histone modifications is more indirect than observed in plants and mammals (more
482 discussed below) and might also require the activity of accessory proteins such as
483 UHRF1 type⁴⁸ or DNMT3-like⁴⁹ enzymes that we have not investigated in this study.

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

486 should retain much attention. In this context, we investigated the role of DNMT5a which
487 is conserved in the pennate diatom *P. tricornutum*. As discussed above, the DNMT5a
488 enzyme of *P. tricornutum* is not shared with many other diatoms but is an orthologue
489 of the DNMT5a protein from *Cryptococcus neoformans* which is involved in the
490 maintenance of DNA methylation patterns^{11,14}. In laboratory conditions, *P. tricornutum*
491 propagates via mitotic division. Our study demonstrates that the ‘somatic’ (i.e. clonal)
492 loss of DNMT5 was sufficient alone to generate a global loss of CG methylation upon
493 repeated cellular divisions. For this reason, we state here that we monitored a loss due
494 to defective maintenance activity. TEs are major targets of DNA methylation in
495 numerous eukaryote lineages including dinoflagellates and diatoms anchoring these
496 features to the origin of eukaryotes^{23,24}. Our DMR analysis *de novo* identified regions
497 that show extensive methylation in the reference strain i.e TEs that were also reported
498 in earlier studies²⁴ (supplementary figure S3).

499 We further demonstrate that only 15% of TEs concentrate more than 70% of
500 DNA methylation in *P. tricornutum*. We found that most TEs associate with active
501 epigenetic marks alone. It strongly suggests that although TE methylation is indeed
502 observed, TE regulation in diatoms is for most part not directly dependent of DNA
503 methylation in general. Genes showed low overlap with both DMRs and previous
504 reports (supplementary figure 3) strongly suggesting that gene methylation is more
505 labile even within independent lines of *P. tricornutum*. In addition, the loss of DNA
506 methylation did not trigger extensive changes in gene expression but a reactivation of
507 transposable elements. This is a strong support for the hypothesis that 5mC was
508 recruited to keep TEs directly or indirectly at bay in the very first eukaryotic radiations.
509 Nonetheless, our data also show that not all TEs are responsive to the loss of DNA
510 methylation independently of their methylation levels. In *P. tricornutum*, chromatin
511 immunoprecipitation studies followed by high throughput sequencing demonstrated
512 that a subset of TEs are co-marked by DNA methylation as well as repressive histone
513 modifications such as H3K27me3 and H3K9me2/3 suggesting redundancy or
514 dependency within the epigenetic code of diatoms³⁷. Interestingly, we found that
515 upregulation is also associated with TEs in highly heterochromatic regions including
516 transposable elements co-marked by DNA methylation and H3K27me3. This
517 combination of repressive marks at TEs is remarkable. DNA methylation is known to
518 interact with histone modifications. The interdependency of DNA methylation and

519 H3K9 methylation is well documented in plants⁵⁰, fungi⁵¹ and mammals^{52,53}. In
520 *Arabidopsis thaliana*, upon loss of CG methylation in *met-1* (DNMT1) mutants,
521 H3K27me3 is relocated at specific hypomethylated TEs⁵⁴. Similar observations were
522 made in *ddm1* mutants defective for heterochromatic cytosine methylation⁵⁵. In
523 mammalian systems, compensatory relocation mechanisms of H3K27me3, but also
524 H3K9me3, occur upon extensive DNA hypomethylation to maintain repression in a loci
525 specific but also TE family specific manner⁵⁶. We make the hypothesis that, aside
526 direct repressive roles, the loss of DNA methylation also triggered genome wide
527 changes within the epigenetic landscape of *P. tricornutum* in turn controlling the
528 expression of a subclass of TEs. It was found that a subset of transposons are
529 responsive to nitrogen depletion in conjunction with hypomethylation in *P.*
530 *tricornutum*⁵⁷. It is thus possible that TE expression also relies on condition specific
531 transcription factor binding that are not fully recapitulated in our growth conditions.

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

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

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

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

565

566

567 **Material and Methods**

568 **Phylogenetic analysis of DNMTs in microeukaryotes**

569 HMMER and RBH analysis

570 We performed an extensive scan of the MMETSP database by HMMER-search using
571 the model PF00145 to fetch any DNMT-like, including partial transcripts, sequence
572 within micro eukaryotes. We ran HMMER in a no-stringent fashion aiming to do not
573 miss positives DNMT sequences. We used eDAF approach to filter the expected high
574 number of false positives. It is worth noting that we initially use HMMER for screening
575 instead of the built-in module of eDAF due to the time complexity of the latter for
576 extensive searches (tens to hundreds of times slower than HMMER). Reciprocal
577 BLAST best hit analysis was performed as previously described³¹. The DNMT3
578 (*Phatr3_J47136*), DNMT4 (*Thaps3_11011*), DNMT5 (*Phatr3_EG02369*) and DNMT6
579 (*Phatr3_J47357*) orthologues found in *P. tricornutum* or *T. pseudonana* (for DNMT4)
580 were blasted on a phylogenetically optimized database that include MMETSP
581 transcriptomes. Upon reciprocal BLAST, putative DNMT sequence hits giving back the
582 corresponding enzyme (DNMT3, DNMT4, DNMT5 or DNMT6) at the threshold of e-

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

585 eDAF-guided domain architecture analysis

586 enhanced Domain Architecture Framework (eDAF) is a four module computational tool
587 for gene prediction, gene ontology and functional domain predictions³⁰. As previously
588 described for Polycomb and Trithorax enzymes³⁰, candidate DNMTs identified by RBH
589 and HMMER-search were analyzed using the DAMA-CLADE guided built-in functional
590 domain architecture. The domain architecture of representative enzymes used in this
591 study can be found in Supplementary table S2.

592 Phylogenetic tree analysis

593 The DNMT domain of candidate enzymes were aligned using ClustalΩ⁵⁹. The
594 alignment was manually curated and trimmed using trimAL (removing >25% gap
595 column) to align corresponding DNMT motifs in all gene families. A convergent
596 phylogenetic tree was then generated using the online CIPRES Science gateway
597 program⁶⁰ using MrBAYES built-in algorithm. Default parameters were used with the
598 following specifications for calculation of the posterior probability of partition:
599 sumt.burninfraction = 0.5, sump.burningfraction = 0.5.

600 **Cell cultures**

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

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

617 **CRISPR/Cas9 mediated gene extinction**

618 The CRISPR/Cas9 knock outs were performed as previously described⁶¹. Our strategy
619 consisted in the generation of short deletions and insertions to disrupt the open reading
620 frame of putative DNMTs of *P. tricornutum*. We introduced by biolistic the guide RNAs
621 independently of the Cas9 and ShBle plasmids, conferring resistance to Phleomycin,
622 into the reference strain Pt18.6 (referred hereafter as 'reference line' or 'wild-type'-
623 WT). Briefly, specific target guide RNAs were designed in the first exon of
624 Phatr3_EG02369 (DNMT5), Phatr3_J47357 (DNMT6) and Phatr3_J36137 (DNMT3)
625 using the PHYTO/CRISPR-EX⁶² software and cloned into the pU6::AOX-sgRNA
626 plasmid by PCR amplification. For PCR amplification, plasmid sequences were added
627 in 3' of the guide RNA sequence (minus -NGG) which are used in a PCR reaction with
628 the template pU6::AOX-sgRNA. Forward primer - sgRNA seq +
629 GTTTTAGAGCTAGAAATAGC. Reverse primer - sequence to add in 3' reverse
630 sgRNA seq + CGACTTTGAAGGTGTTTTTTG. This will amplify a new pU6::AOX-
631 (your_sgRNA). The PCR product is digested by the enzyme DPN1 (NEB) in order to
632 remove the template plasmid and cloned in TOPO10 *E. coli*. The sgRNA plasmid, the
633 pDEST-hCas9-HA and the ShBLE Phleomycin resistance gene cloned into the plasmid
634 pPHAT-eGFP were co-transformed by biolistic in the Pt18.6 'Wild Type' strain as
635 described in⁶¹. We also generated a line that was transformed with pPHAT-eGFP and
636 pDEST-hCas9-HA but no guide RNAs. This line is referred as the Cas9:Mock line.

637 **RNA and DNA extraction**

638 Total RNA extraction was performed by classical TRIZOL/Chloroform isolations and
639 precipitation by isopropanol. Frozen cell pellets were extracted at a time in a series of
640 3 technical extraction/duplicates and pooled. RNA was DNase treated using DNase I
641 (ThermoFisher) as per manufacturer's instructions. DNA extraction was performed
642 using the Invitrogen™ Easy-DNA™ gDNA Purification Kit following 'Protocol #3'
643 instructions provided by the manufacturer. Extracted nucleic acids were measured
644 using QUBIT fluorometer and NANODROP spectrometer. RNA and gDNA Integrity
645 were controlled by electrophoresis on 1% agarose gels.

646 **RT-QPCR and analysis**

647 qPCR primers were designed using the online PRIMER3 program v0.4.0 defining 110-
648 150 amplicon size and annealing temperature between 58°C and 62°C. Primer
649 specificity was checked by BLAST on *P. tricornutum* genome at ENSEMBL. For cDNA
650 synthesis, 1 µg of total RNA was reverse transcribed using the SuperScript™ III First-
651 Strand (Invitrogen) protocol. For quantitative reverse transcription polymerase chain
652 reaction (RT-qPCR) analysis, cDNA was amplified using SYBR Premix ExTaq (Takara,
653 Madison, WI, USA) according to manufacturer's instructions. CT values for genes of
654 interest were generated on a Roche lightcycler® 480 qpcr system. CT values were
655 normalized on housekeeping genes using the deltaCT method. Normalized CT values
656 for amplifications using multiple couple of primers targeting several cDNA regions of
657 the genes of interest were then averaged and used as RNA levels proxies.

658 **Dot blot**

659 gDNA samples were boiled at 95°C for 10 min for denaturation. Samples were
660 immediately placed on ice for 5 min, and 250-500 ng were loaded on regular
661 nitrocellulose membranes. DNA was then autocrosslinked in a UVC 500 crosslinker –
662 2 times at 1200uj (*100). The membranes were blocked for 1 hr in 5% PBST-BSA.
663 Membranes were probed for 1 hr at room temperature or overnight at 4°C with 1:1000
664 dilution of 5mC antibody (OptimAbtm Anti-5-Methylcytosine – BY-MECY 100). 5mC
665 signals were revealed using 1:5000 dilution of HRP conjugated antirabbit IgG
666 secondary antibody for 1 hr at room temperature followed by chemo luminescence.
667 Loading was measured using methylene blue staining.

668 **RNA sequencing and Bisulfite sequencing**

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

677 **Bisulfite sequencing analysis**

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

682 **R version and tools**

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

686 **DMR calling**

687 Differentially methylated regions were called using the DMRcaller Rpackage³⁸. Given
688 the low level of correlation of DNA methylation observed in *P. tricornutum*^{11,24} and low
689 sequencing coverage in Pt18.6 (supplementary Figure S3), only cytosines with $\geq 5X$
690 coverage in all three lines were kept for further analysis (supplementary Figure S3)
691 and the bins strategy was favored over other built-in DMRcaller tools. DMRs were
692 defined as 100bp regions with at least an average 20% loss/gain of DNA methylation
693 in either one of the DNMT5:KOs compared to the reference strain. The 'Score test'
694 method was used to calculate statistical significance and threshold was set at p.value
695 < 0.01 . In addition, to separate isolated differentially methylated cytosines from regions
696 with significant loss of DNA methylations, each hypoDMRs must contain at least 4
697 methylated cytosines in the reference strain.

698 **Overlap with histone post translational modifications and genomic annotations.**

699 Percentage overlaps between DMRs as well as the overlap of gene and TEs
700 coordinates with histone posttranslational modifications and DMRs were calculated
701 using the genomation R package computation⁶³ and the 'annotateWithFeature' and
702 'getMembers' functions. DMRs were extended 500bp in 5' and 3' to account for
703 promoter and regulatory region marking. For RNAseq analysis, we analyzed the
704 expression of TEgenes previously defined³⁹. To define TE genes in DMRs we
705 crosscheck overlapping TEgenes annotations with *bona fide* TEs in DMRs using
706 'annotatewithFeature' function

707 **RNAseq analysis**

708 150bp pair-end sequenced reads were aligned on the reference genome of *P.*
709 *tricornutum* (Phatr3) using STAR (v2.5.3a). Expression levels were quantified using
710 HTseq v 0.7.2. Differentially expressed genes were analyzed using DESeq2 v1.19.37.
711 FDR values are corrected P. values using the Benjamin and Hochberg method.

712 **Legend**

713 **Figure 1**

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

724 **Figure 2**

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

732

733 **Figure 3**

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

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

742 **Figure 4**

743 Homozygous mutations generated by CRISPR/Cas9 in M23 and M25 lines at two
744 independent target sequences. In M25, the mutation consists in 16 base pair out of
745 frame deletion around CRISPR/Cas9 cutting sites that generates a loss of amino acids
746 28 to 34 leading to a premature STOP codon at amino acid 280. M23 has a 11 base
747 pair out of frame deletion that generates a loss of amino acids 58 to 60/61 followed by
748 a premature STOP codon at amino acid position 179-180 from ATG

749

750 **Figure 5**

751

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

766 **Figure 6**

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

770

771 **Figure 7**

772 **A.** Association between annotated transposable elements and regions targeted by
773 histone post-translational modifications in the reference strain and regions with hypo
774 methylation in DNMT5:KO M23. For each category, the total number of TEs is
775 indicated. Proportion of TEs with consistent overexpression in both DNMT5:KOs is
776 showed in pink. Red circles indicate regions overlapping with hypoDMRs in M23. We
777 labelled combination of epigenetic landscape for which significant changes in
778 expression are pictured in the following panel **B.** For epigenetic categories defined in
779 **A.** We show distribution of expression of TE-genes as a log₂ of fold change in RNA
780 levels in DNMT5:KOs compared to the WT strain. Only fold change levels of TE-genes
781 with a false discovery rate (FDR) <0.01 are used. We compared levels of expression
782 between the different epigenetic categories within each mutant. Distributions sharing
783 the same letters are statistically different based on the Mann-Whitney U-test at p.value
784 <0.01. Within categories and between mutants M23 showed lower induction level in
785 the 'C' (H3K27me₃ + H3K9me₂ +H3K9me₃) and 'E' (hypoDMRs+H3K27me₃)
786 categories.

787 **Legend of supplemental information**

788 **Figure S1**

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

793 **Figure S2**

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

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

809 **Figure S3**

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

821 **Figure S4**

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

826

827 **Supplementary table S1**

828 List of putative DNMTs fetched by HMMER in the MMETSP database (“HMMER hits”)
829 and reciprocal best hits search in the phylogenetically optimized database described
830 in ³¹. When already reported in other studies we added references. We also added
831 enzymes found in reference database and literature in the tree for comparison
832 purposes. References – MMETSP⁶⁴ Dorrell et al.³¹; De Mendoza et al. 2018²³;
833 GENOSCOPE (courtesy of GENOSCOPE sequencing project); Bewick et al. 2016³⁵

834 **Supplementary table S2**

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

837 **Supplementary table S3**

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

840 **Supplementary table S4**

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

849 **Supplementary table S5 and S6**

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

853 **Supplementary table S7**

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

858 **Supplementary table S8**

859 Summary of RNAseq analysis.

860 **Supplementary table S9**

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

863 **Supplementary table 10**

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

867 **Supplementary table S11**

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

871 **Supplementary table S12**

872 Summary and list of upregulated genes in DNMT5:KOs.

873 **Supplementary table S13**

874 Summary and list of downregulated genes in DNMT5:KOs.

875 **Acknowledgements**

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

880 **Author Contributions**

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

887

888

889

890

891

892
893
894
895
896
897
898
899
900
901
902
903
904

References

- 905 1. Kato, M., Miura, A., Bender, J., Jacobsen, S. E. & Kakutani, T. Role of CG and
906 non-CG methylation in immobilization of transposons in Arabidopsis. *Curr. Biol.*
907 (2003) doi:10.1016/S0960-9822(03)00106-4.
- 908 2. Zilberman, D. The evolving functions of DNA methylation. *Current Opinion in*
909 *Plant Biology* (2008) doi:10.1016/j.pbi.2008.07.004.
- 910 3. Barlow, D. P. & Bartolomei, M. S. Genomic imprinting in mammals. *Cold Spring*
911 *Harb. Perspect. Biol.* (2014) doi:10.1101/cshperspect.a018382.
- 912 4. Galupa, R. & Heard, E. X-Chromosome Inactivation: A Crossroads Between
913 Chromosome Architecture and Gene Regulation. *Annu. Rev. Genet.* (2018)
914 doi:10.1146/annurev-genet-120116-024611.
- 915 5. Bestor, T. H. DNA methylation: evolution of a bacterial immune function into a
916 regulator of gene expression and genome structure in higher eukaryotes.
917 *Philosophical transactions of the Royal Society of London. Series B, Biological*
918 *sciences* (1990) doi:10.1098/rstb.1990.0002.

- 919 6. Kumar, S. *et al.* The DNA (cytosine-5) methyltransferases. *Nucleic Acids*
920 *Research* (1994) doi:10.1093/nar/22.1.1.
- 921 7. Cheng, X., Kumar, S., Klimasauskas, S. & Roberts, R. J. Crystal structure of
922 the HhaI DNA methyltransferase. in *Cold Spring Harbor Symposia on*
923 *Quantitative Biology* (1993). doi:10.1101/SQB.1993.058.01.039.
- 924 8. Zemach, A. & Zilberman, D. Evolution of eukaryotic DNA methylation and the
925 pursuit of safer sex. *Current Biology* (2010) doi:10.1016/j.cub.2010.07.007.
- 926 9. Zemach, A., McDaniel, I. E., Silva, P. & Zilberman, D. Genome-wide
927 evolutionary analysis of eukaryotic DNA methylation. *Science* (80-.). (2010)
928 doi:10.1126/science.1186366.
- 929 10. Ponger, L. & Li, W. H. Evolutionary diversification of DNA methyltransferases in
930 eukaryotic genomes. *Mol. Biol. Evol.* (2005) doi:10.1093/molbev/msi098.
- 931 11. Huff, J. T. & Zilberman, D. Dnmt1-independent CG methylation contributes to
932 nucleosome positioning in diverse eukaryotes. *Cell* (2014)
933 doi:10.1016/j.cell.2014.01.029.
- 934 12. Greenberg, M. V. C. & Bourc'his, D. The diverse roles of DNA methylation in
935 mammalian development and disease. *Nature Reviews Molecular Cell Biology*
936 (2019) doi:10.1038/s41580-019-0159-6.
- 937 13. Kouzminova, E. & Selker, E. U. Dim-2 encodes a DNA methyltransferase
938 responsible for all known cytosine methylation in *Neurospora*. *EMBO J.* (2001)
939 doi:10.1093/emboj/20.15.4309.
- 940 14. Dumesic, P. A., Stoddard, C. I., Catania, S., Narlikar, G. J. & Madhani, H. D.
941 ATP Hydrolysis by the SNF2 Domain of Dnmt5 Is Coupled to Both Specific
942 Recognition and Modification of Hemimethylated DNA. *Mol. Cell* (2020)
943 doi:10.1016/j.molcel.2020.04.029.
- 944 15. Gladyshev, E. Repeat-Induced Point Mutation and Other Genome Defense
945 Mechanisms in Fungi. *Microbiol. Spectr.* (2017)
946 doi:10.1128/microbiolspec.funk-0042-2017.
- 947 16. Galagan, J. E. & Selker, E. U. RIP: The evolutionary cost of genome defense.
948 *Trends in Genetics* (2004) doi:10.1016/j.tig.2004.07.007.

- 949 17. Yang, K. *et al.* The DmtA methyltransferase contributes to *Aspergillus flavus*
950 conidiation, sclerotial production, aflatoxin biosynthesis and virulence. *Sci.*
951 *Rep.* (2016) doi:10.1038/srep23259.
- 952 18. Malagnac, F. *et al.* A gene essential for de novo methylation and development
953 in ascobolus reveals a novel type of eukaryotic DNA methyltransferase
954 structure. *Cell* (1997) doi:10.1016/S0092-8674(00)80410-9.
- 955 19. Bewick, A. J., Vogel, K. J., Moore, A. J. & Schmitz, R. J. Evolution of DNA
956 methylation across insects. *Mol. Biol. Evol.* (2017)
957 doi:10.1093/molbev/msw264.
- 958 20. Feng, S. *et al.* Conservation and divergence of methylation patterning in plants
959 and animals. *Proc. Natl. Acad. Sci. U. S. A.* (2010)
960 doi:10.1073/pnas.1002720107.
- 961 21. Vanyushin, B. F. & Ashapkin, V. V. *DNA methylation in plants. DNA*
962 *Methylation in Plants* (2011). doi:10.1146/annurev.arplant.49.1.223.
- 963 22. Zhong, X. *et al.* Molecular mechanism of action of plant DRM de novo DNA
964 methyltransferases. *Cell* (2014) doi:10.1016/j.cell.2014.03.056.
- 965 23. De Mendoza, A. *et al.* Recurrent acquisition of cytosine methyltransferases into
966 eukaryotic retrotransposons. *Nat. Commun.* (2018) doi:10.1038/s41467-018-
967 03724-9.
- 968 24. Veluchamy, A. *et al.* Insights into the role of DNA methylation in diatoms by
969 genome-wide profiling in *Phaeodactylum tricornutum*. *Nat. Commun.* (2013)
970 doi:10.1038/ncomms3091.
- 971 25. Fan, X. *et al.* Single-base methylome profiling of the giant kelp *Saccharina*
972 *japonica* reveals significant differences in DNA methylation to microalgae and
973 plants. *New Phytol.* (2020) doi:10.1111/nph.16125.
- 974 26. Armbrust, E. V. The life of diatoms in the world's oceans. *Nature* (2009)
975 doi:10.1038/nature08057.
- 976 27. Malviya, S. *et al.* Insights into global diatom distribution and diversity in the
977 world's ocean. *Proc. Natl. Acad. Sci. U. S. A.* (2016)
978 doi:10.1073/pnas.1509523113.

- 979 28. Traller, J. C. *et al.* Genome and methylome of the oleaginous diatom *Cyclotella*
980 *cryptica* reveal genetic flexibility toward a high lipid phenotype. *Biotechnol.*
981 *Biofuels* (2016) doi:10.1186/s13068-016-0670-3.
- 982 29. Lister, R. *et al.* Human DNA methylomes at base resolution show widespread
983 epigenomic differences. *Nature* (2009) doi:10.1038/nature08514.
- 984 30. Zhao, X. *et al.* Probing the Diversity of Polycomb and Trithorax Proteins in
985 Cultured and Environmentally Sampled Microalgae. *Front. Mar. Sci.* (2020)
986 doi:10.3389/fmars.2020.00189.
- 987 31. Dorrell, R. G. *et al.* Chimeric origins of ochrophytes and haptophytes revealed
988 through an ancient plastid proteome. *Elife* (2017) doi:10.7554/eLife.23717.
- 989 32. Bernardes, J. S., Vieira, F. R. J., Zaverucha, G. & Carbone, A. A multi-objective
990 optimization approach accurately resolves protein domain architectures.
991 *Bioinformatics* (2016) doi:10.1093/bioinformatics/btv582.
- 992 33. Bernardes, J., Zaverucha, G., Vaquero, C. & Carbone, A. Improvement in
993 Protein Domain Identification Is Reached by Breaking Consensus, with the
994 Agreement of Many Profiles and Domain Co-occurrence. *PLoS Comput. Biol.*
995 (2016) doi:10.1371/journal.pcbi.1005038.
- 996 34. Bewick, A. J. *et al.* Diversity of cytosine methylation across the fungal tree of
997 life. *Nat. Ecol. Evol.* (2019) doi:10.1038/s41559-019-0810-9.
- 998 35. Bewick, A. J. *et al.* The evolution of CHROMOMETHYLASES and gene body
999 DNA methylation in plants. *Genome Biol.* **18**, (2017).
- 1000 36. Jurkowska, R. Z., Jurkowski, T. P. & Jeltsch, A. Structure and Function of
1001 Mammalian DNA Methyltransferases. *ChemBioChem* (2011)
1002 doi:10.1002/cbic.201000195.
- 1003 37. Veluchamy, A. *et al.* An integrative analysis of post-translational histone
1004 modifications in the marine diatom *Phaeodactylum tricornutum*. *Genome Biol.*
1005 (2015) doi:10.1186/s13059-015-0671-8.
- 1006 38. Catoni, M., Tsang, J. M. F., Greco, A. P. & Zabet, N. R. DMRcaller: A versatile
1007 R/Bioconductor package for detection and visualization of differentially
1008 methylated regions in CpG and non-CpG contexts. *Nucleic Acids Res.* (2018)

- 1009 doi:10.1093/nar/gky602.
- 1010 39. Rastogi, A. *et al.* Integrative analysis of large scale transcriptome data draws a
1011 comprehensive landscape of *Phaeodactylum tricornutum* genome and
1012 evolutionary origin of diatoms. *Sci. Rep.* (2018) doi:10.1038/s41598-018-
1013 23106-x.
- 1014 40. Alexa, A. & Rahnenführer, J. Gene set enrichment analysis with topGO.
1015 *Bioconductor Improv.* (2007).
- 1016 41. Ait-Mohamed, O. *et al.* PhaeoNet: A Holistic RNAseq-Based Portrait of
1017 Transcriptional Coordination in the Model Diatom *Phaeodactylum tricornutum*.
1018 *Front. Plant Sci.* (2020) doi:10.3389/fpls.2020.590949.
- 1019 42. Stroud, H. *et al.* Non-CG methylation patterns shape the epigenetic landscape
1020 in *Arabidopsis*. *Nat. Struct. Mol. Biol.* (2014) doi:10.1038/nsmb.2735.
- 1021 43. Cuypers, B. *et al.* C-5 DNA methyltransferase 6 does not generate detectable
1022 DNA methylation in *Leishmania*. *bioRxiv* (2019) doi:10.1101/747063.
- 1023 44. Grüne, T. *et al.* Crystal structure and functional analysis of a nucleosome
1024 recognition module of the remodeling factor ISWI. *Mol. Cell* (2003)
1025 doi:10.1016/S1097-2765(03)00273-9.
- 1026 45. Lu, R. & Wang, G. G. Tudor: A versatile family of histone methylation 'readers'.
1027 *Trends in Biochemical Sciences* (2013) doi:10.1016/j.tibs.2013.08.002.
- 1028 46. Pek, J. W., Anand, A. & Kai, T. Tudor domain proteins in development. *Dev.*
1029 (2012) doi:10.1242/dev.073304.
- 1030 47. Tóth, K. F., Pezic, D., Stuwe, E. & Webster, A. The pirna pathway guards the
1031 germline genome against transposable elements. in *Advances in Experimental*
1032 *Medicine and Biology* (2016). doi:10.1007/978-94-017-7417-8_4.
- 1033 48. Bostick, M. *et al.* UHRF1 plays a role in maintaining DNA methylation in
1034 mammalian cells. *Science* (80-.). (2007) doi:10.1126/science.1147939.
- 1035 49. Ooi, S. K. T. *et al.* DNMT3L connects unmethylated lysine 4 of histone H3 to de
1036 novo methylation of DNA. *Nature* (2007) doi:10.1038/nature05987.
- 1037 50. Jackson, J. P., Lindroth, A. M., Cao, X. & Jacobsen, S. E. Control of CpNpG

- 1038 DNA methylation by the KRYPTONITE histone H3 methyltransferase. *Nature*
1039 (2002) doi:10.1038/nature731.
- 1040 51. Honda, S. & Selker, E. U. Direct Interaction between DNA Methyltransferase
1041 DIM-2 and HP1 Is Required for DNA Methylation in *Neurospora crassa*. *Mol.*
1042 *Cell. Biol.* (2008) doi:10.1128/mcb.00823-08.
- 1043 52. Estève, P. O. *et al.* Direct interaction between DNMT1 and G9a coordinates
1044 DNA and histone methylation during replication. *Genes Dev.* (2006)
1045 doi:10.1101/gad.1463706.
- 1046 53. Ikegami, K. *et al.* Genome-wide and locus-specific DNA hypomethylation in
1047 G9a deficient mouse embryonic stem cells. *Genes to Cells* (2007)
1048 doi:10.1111/j.1365-2443.2006.01029.x.
- 1049 54. Mathieu, O., Probst, A. V. & Paszkowski, J. Distinct regulation of histone H3
1050 methylation at lysines 27 and 9 by CpG methylation in *Arabidopsis*. *EMBO J.*
1051 (2005) doi:10.1038/sj.emboj.7600743.
- 1052 55. Rougée, M. *et al.* Polycomb mutant partially suppresses DNA hypomethylation-
1053 associated phenotypes in *Arabidopsis*. *Life Sci. Alliance* (2021)
1054 doi:10.26508/LSA.202000848.
- 1055 56. Walter, M., Teissandier, A., Pérez-Palacios, R. & Bourc'His, D. An epigenetic
1056 switch ensures transposon repression upon dynamic loss of DNA methylation
1057 in embryonic stem cells. *Elife* (2016) doi:10.7554/eLife.11418.001.
- 1058 57. Maumus, F. *et al.* Potential impact of stress activated retrotransposons on
1059 genome evolution in a marine diatom. *BMC Genomics* (2009)
1060 doi:10.1186/1471-2164-10-624.
- 1061 58. Rogato, A. *et al.* The diversity of small non-coding RNAs in the diatom
1062 *Phaeodactylum tricornutum*. *BMC Genomics* (2014) doi:10.1186/1471-2164-
1063 15-698.
- 1064 59. Madeira, F. *et al.* The EMBL-EBI search and sequence analysis tools APIs in
1065 2019. *Nucleic Acids Res.* (2019) doi:10.1093/nar/gkz268.
- 1066 60. Miller, M. A., Pfeiffer, W. & Schwartz, T. Creating the CIPRES Science
1067 Gateway for inference of large phylogenetic trees. in *2010 Gateway Computing*

- 1068 *Environments Workshop, GCE 2010* (2010). doi:10.1109/GCE.2010.5676129.
- 1069 61. Nymark, M., Sharma, A. K., Sparstad, T., Bones, A. M. & Winge, P. A
1070 CRISPR/Cas9 system adapted for gene editing in marine algae. *Sci. Rep.*
1071 (2016) doi:10.1038/srep24951.
- 1072 62. Rastogi, A., Murik, O., Bowler, C. & Tirichine, L. PhytoCRISP-Ex: A web-based
1073 and stand-alone application to find specific target sequences for CRISPR/CAS
1074 editing. *BMC Bioinformatics* (2016) doi:10.1186/s12859-016-1143-1.
- 1075 63. Akalin, A., Franke, V., Vlahoviček, K., Mason, C. E. & Schübeler, D.
1076 Genomation: A toolkit to summarize, annotate and visualize genomic intervals.
1077 *Bioinformatics* (2015) doi:10.1093/bioinformatics/btu775.
- 1078 64. Keeling, P. J. *et al.* The Marine Microbial Eukaryote Transcriptome Sequencing
1079 Project (MMETSP): Illuminating the Functional Diversity of Eukaryotic Life in
1080 the Oceans through Transcriptome Sequencing. *PLoS Biol.* (2014)
1081 doi:10.1371/journal.pbio.1001889.
- 1082