Published in final edited form as:
Nat Genet. 2018 March ; 50(3): 452–459. doi:10.1038/s41588-018-0061-8.

Evolutionary analysis indicates that DNA alkylation damage is a byproduct of cytosine DNA methyltransferase activity

Silvana Rošić[#1,2], **Rachel Amouroux**[#1,2], **Cristina E. Requena**[#1,2], **Ana Gomes**[1,2], **Max Emperle**[4], **Toni Beltran**[1,2], **Jayant K. Rane**[1,2], **Sarah Linnett**[1,2], **Murray E. Selkirk**[3], **Philipp H. Schiffer**[5], **Allison J. Bancroft**[6], **Richard K. Grencis**[6], **Albert Jeltsch**[4], **Petra Hajkova**[1,2,+], and **Peter Sarkies**[1,2,+]

[1]MRC London Institute of Medical Sciences, Du Cane Road, London W12 0NN, United Kingdom

[2]Institute of Clinical Sciences, Imperial College London, Du Cane Road, London W12 0NN, United Kingdom

[3]Department of Life Sciences, Imperial College London, South Kensington Campus, SW7 2AZ, United Kingdom

[4]Institute of Biochemistry, Universitaet Stuttgart, Germany

[5]Department of Ecology and Evolution, University College London, United Kingdom

[6]School of Biological Sciences and Wellcome Trust Centre for Cell Matrix Research, FBMH, MAHSC, University of Manchester, Oxford Road, Manchester, M13 9PT, UK

[#] These authors contributed equally to this work.

Abstract

Methylation at the 5 position of cytosine in DNA (5meC), is a key epigenetic mark in eukaryotes. Once introduced, 5meC can be maintained through DNA replication due to the activity of “maintenance” DNA methyltransferases.. Despite their ancient origin, DNA methylation pathways differ widely across metazoans, such that 5meC is either confined to transcribed genes or lost altogether in several lineages. Here we use comparative epigenomics to investigate the evolution of DNA methylation. Although the model nematode *C. elegans* has lost DNA methylation, more basal nematodes retain cytosine DNA methylation, targeted to repeat loci. Unexpectedly, we find that DNA methylation coevolves with the DNA alkylation repair enzyme ALKB2 across eukaryotes. We further show that DNA methyltransferases introduce the toxic lesion 3meC into

[+]Correspondence: Petra Hajkova; petra.hajkova@lms.mrc.ac.uk, Peter Sarkies; psarkies@imperial.ac.uk.

Author contributions

Conception of study: PS, PH; Design of experiments: PS, PH, AJ; DNA extraction and Bisulfite sequencing: SR, PS; Bioinformatic and computational analyses: PS; 3meC analysis by LC/MS RA,CER,SL, PS; ESC CRISPR deletion and analysis AG, JR, PS; DNMT3A analysis in vitro ME, AJ; Genome assembly TB, PHS; Nematode culture SR, MES, RKG, AJB; Data analyses PS, PH, AJ; Manuscript preparation PS, PH, AJ.

Competing Interests Statement

The authors declare that they have no competing interests

DNA both *in vitro* and *in vivo*. Alkylation damage is thus intrinsically associated with DNA methyltransferase activity, and this may promote the loss of DNA methylation in many species.

DNA methylation is an important regulatory mechanism in eukaryotes, with important functions including transposable element silencing and gene regulation1. 5meC acts as an epigenetic modification, which, once introduced by de novo methyltransferases (DNMT3a/b in mammals), can be maintained through cell division due to the activity of maintenance methyltransferases (DNMT1 in mammals)2. Both de novo and maintenance methylation are conserved in many species across eukaryotes including animals, plants and fungi3,4. Nevertheless, DNA methylation pathways evolve rapidly in multiple lineages. Levels of DNA methylation vary widely, with many insects displaying sparse DNA methylation confined to a subset of transcribed genes5–9. Moreover, in many species, including the key model organisms *Drosophila melanogaster*, *Caenorhabditis elegans* and *Saccharomyces cerevisiae*, cytosine DNA methylation has been lost altogether5,10. The factors driving such rapid evolution of DNA methylation pathways and their targets are unclear. Here we investigate the evolution of DNA methylation in the nematode phylum and more widely across eukaryotes. We discover that DNA methylation coevolves with DNA repair pathways and with the ALKB2 alkylation repair system in particular. We explain this by identifying a hitherto unknown off-target effect of DNA methyltransferases, whereby they introduce alkylation damage into DNA. Indeed, we show that DNA methyltransferases are the major endogenous source of the alkylation 3meC lesion in cells. We further hypothesise that this toxic activity may act to promote the loss of DNA methylation altogether in multiple lineages.

Results

DNA methylation is conserved in basal nematodes

In order to study the factors driving the evolution of DNA methylation we searched for cytosine DNA methyltransferases (DNMTs) in nematode genomes across the phylum (Figure 1A). We first used the Pfam core domain to identify potential cytosine DNA methyltransferases and then used phylogenetic analysis to group these with known eukaryotic DNMTs. All identified nematode DNMTs were homologues of DNMT1, 2 or 3. DNMT2, which predominantly methylates tRNA3,4,11, was the most widespread amongst nematodes but has been lost independently in some lineages including *Caenorhabditis* whilst being conserved in the closely related parasitic nematode *Nippostrongylus brasiliensis* (Clade V; Figure 1B). In agreement with previous analyses of individual species12,13 we found cytosine DNA methyltransferases DNMT1 and DNMT3 are retained in early-branching lineages, confirming that they are ancestral to nematodes. Cytosine DNA methyltransferases DNMT1 and 3 were most likely lost completely in the common ancestor of the Rhabditida group that contains *C. elegans* (Clades III-V; Figure 1B). Interestingly, amongst nematodes retaining cytosine DNA methyltransferases some nematodes possess both DNMT1 and DNMT3 (*Romanomermis culicivorax*), while some species possess only DNMT1 (*Plectus sambesii*) or DNMT3 (*Trichuris muris* and *Trichinella spiralis*) (Figure 1B,C). Importantly, in species where DNMT3 is the sole identified DNMT (*T. spiralis* or *T. muris)* this protein has not adopted any additional domains from DNMT1 (Figure 1D, E).

To investigate the impact of the presence of various combinations of cytosine methyltransferases we measured the abundance of cytosine methylation (5meC) in genomic DNA using ultrasensitive Liquid Chromatography/Mass spectrometry (LC/MS). 5meC was clearly detectable in all species containing DNMT1 or DNMT3 (Figure 2A). We did not detect any 5meC in C*. briggsae*, which does not have DNMTs, and only detected very low levels in *N. brasiliensis*, which only has DNMT2. Notably, *R. culicivorax*, which has both DNMT1 and DNMT3, contains higher levels of genomic 5meC than the other nematodes.

Nematode DNA methylation is enriched at transposable elements

To investigate the targeting of DNMT1 and DNMT3 to different genomic regions we undertook whole genome bisulfite sequencing (Supplementary Table 1). In agreement with our LC/MS analysis (Figure 2A), we detected significant levels of DNA methylation above the Bisulfite non-conversion rate as estimated based on the inclusion of an unmethylated spike-in in the bisulphite reaction (p<1e-100, Fisher's exact test; Figure 2D) in all nematodes with DNMT1 or 3; this methylation was significantly enriched at CG sites over non-CG sites (p<1e-100, Fisher's Exact test; Supplementary Fig 1A). We did not observe significant differences in non-CG methylation between nematodes with DNMT1 or DNMT3 only (Supplementary Figure 1A). Although we did detect trace amounts of 5meC in *N. brasiliensis* possessing DNMT2 only, our bisulfite analysis did not reveal any significant enrichment of 5meC above the non-conversion rate (Fig 2D; Supplementary Fig 1A) suggesting at most a very low and non-specific activity of DNMT2 on cytosine within DNA, as found in *D. melanogaster*10. Of note, the DNMT1-only methylome and the DNMT3-only methylomes showed different preferences for the nucleotide following the methylation (CG) site, normalized to the abundance of each trinucleotide in the genome. Comparison with bisulfite sequencing data from mouse ES cells lacking either DNMT1 or DNMT314 and from the arthropod *Bombyx mori*, which only has DNMT18, revealed that the trinucleotide preferences of DNMT1 were highly similar between nematodes, mammalian cells and *B. mori*. In contrast the DNMT3 preferences were different between nematodes and mammalian cells, suggesting differential conservation of DNA interactions for these two types of DNA methyltransferases (Figure 2B,C).

We next annotated methylation sites genome-wide. All nematodes with DNMT1 or DNMT3 showed significant enrichment of CG methylation above the genome-wide level (p<1e-5, Fisher's Exact Test) normalized to CG content for at least one category of repetitive elements (Fig 2D). In contrast, we did not observe enrichment in overall methylation at genes. Notably, this observation cannot be explained by different trinucleotide composition within repeats, as trinucleotide content was similar across different repeat types (Supplementary Fig 1B, C). *C. briggsae* (no DNMTs) and *N. brasiliensis* (DNMT2-only) showed no such enrichment (Figure 2D). Interestingly *P. sambesii* (DNMT1 only) showed marked enrichment for repeats over the genome-wide background (Fig 2D). DNA methylation could be found across the entire body of many repetitive elements in all species (examples in Figure 2E-H and Supplementary Figure 2,3) and genome-wide, elements with high levels of methylation were enriched for at least one category of repeats (Figure 3A-D).

We next examined DNA methylation in protein-coding genes. Analysis of the DNA methylation level across genes indicated that there were no notable populations of genes with higher levels of DNA methylation in *P. sambesii*, (DNMT1 only), *T. spiralis* (DNMT3 only) or *T. muris* (DNMT3 only) (Fig3E, Supplementary Figure 2,3). The few genes that showed appreciable DNA methylation in these species were likely misannotated repeats, as genes with homology to repeats had higher CG methylation than genes without (Supplementary Fig 4). *T. spiralis* has previously been reported to show gene body methylation12, however, this study did not normalize for CG content. As CG density is markedly higher in the coding regions of all nematodes examined, this likely accounts for the discrepancy with our findings (Supplementary Fig 5).

In *R. culicivorax*, (DNMT1 and DNMT3) there was a bimodal distribution of DNA methylation across genes, with a small population of genes showing elevated DNA methylation (Figure 3E). This finding is potentially reminiscent of gene body methylation in other invertebrates5–9.. However, Gene Ontology (GO) analysis of the top 50 methylated genes with GO annotations showed that ~14% were annotated as nucleic acid integration (enrichment p=1e-55 compared to all genes, Chi squared test with BH multiple test correction; Supplementary tables 2&3), thus even in *R. culicivorax* at least some genes with high levels of methylation may be either misannotated TEs or genes with TE insertions.

Altogether, our analysis of DNA methylation across nematodes indicates that methylation of repeats is its most widely conserved function and was likely to have been present in the common ancestor of nematodes. Methylation in the bodies of transcribed protein-coding genes has been lost altogether in the lineage leading to *T. spiralis* and *T. muris*, in *P. sambesii* and exists only in a minority of genes in *R. culicivorax*, thus is not a conserved feature of DNA methylation in nematodes.

It has been argued that gene body methylation is a universal feature of DNMT1 and 3 activity but that repeat-targeted cytosine methylation evolved independently in plants and vertebrates5,6.. Our data is in accordance with a more nuanced view that the functions of DNA methylation evolve rapidly5,15 and recognising that repeat-targeted DNA methylation is found in invertebrates16,17.. Overall, the rapid evolution both of DNA methylation mechanisms and its targets in nematodes adds to our growing picture of the complex evolution of epigenetic mechanisms in metazoans5,15,18, in which the ancestral metazoans had a rich set of epigenetic mechanisms, which have subsequently been lost independently in many descendent organisms.

DNA methylation coevolves with DNA alkylation damage repair across eukaryotes

What drives the rapid evolution of cytosine DNA methylation pathways in metazoans? One approach to this question is to identify genes coevolving with DNMTs, which may indicate pathways that are linked to the presence or absence of DNA methylation. We took the genomes in Ensembl metazoa (Release 28) and identified 133 human proteins that coevolve with DNMT1 or 3 ($p<0.01$ Fisher's exact test after multiple test correction, Supplementary Table 4). Surprisingly, we found that the most strongly enriched GO term was for DNA repair (Figure 4A; Supplementary Table 5). In particular, we noted the presence of alkylation repair enzymes amongst this set (Supplementary Figure 6, Supplementary Table 4),

including the enzymes ALKB2 and its paralogue ALKB3 (a mammalian-specific duplication; hitherto ALKB2/3). ALKB2/3 are members of the Fe^{2+} dependent oxygenase family of DNA repair enzymes homologous to *E. coli* ALKB19.. Whilst *E. coli* ALKB repairs a wide range of alkylated adducts including protein, RNA and DNA, in eukaryotes the family has diversified with different ALKB family enzymes specialised for repair of particular substrates. Mammalian ALKB2/3 are the only members of the ALKB family that repair alkylation damage in DNA19,20 and are the only members that coevolve with DNA methyltransferases DNMT1 and DNMT3 (Figure 4C, Supplementary Figure 7). To verify independently the association between DNA methylation and ALKB2/3, we carried out phylogenetic profiling of ALKB2&3 and DNMTs across the eukaryotic genomes in the Ensembl database (fungi, protozoa and metazoa) and tested for coevolution between ALKB2/3 and DNA methylation. Importantly, in this analysis we corrected for the overrepresentation of several closely related species in Ensembl (for example the Drosophila genus, in which there are 12 species represented in Ensembl all of which have no ALKB2/3 and no DNMT1 or 3, or mammals, all of which have ALKB2/3 and DNMT1 and 3) by ensuring that only one member from each lineage with the same profile of ALKB2/3 and DNMTs was included in the analysis (Figure 4B; Supplementary Figures 8-11, see Supplementary Tables 6-8 for the list of all species considered for the analysis). All three groups showed statistically significant co- occurrence between ALKB2/3 and the presence of at least one cytosine DNA methyltransferase (DNMT1 and 3) ($p<0.001$ fungi, $p<0.005$ metazoa, $p<0.01$ protazoa using Fisher's Exact Test; Figure 4B). Additionally in some fungi where ALKB2/3 is present but DNMT1 is absent, DNMT5, which acts on CG sequences15, is conserved (Supplementary Figures 8-11).

We note that there are some potentially interesting exceptions to the general coevolution between ALKB2/3 and DNMTs, particularly in arthropods, where several species have lost ALKB2/3 whilst retaining DNA methyltransferases. To investigate this further we compared genome-wide methylation levels across arthropods using previously published data from 18 insects9, the crustaceans *Parhylale hawaiensis*17, *Daphnia pulex* and *Daphnia magna*21 and the desert locust *Schistocerca gregaria*16.. This showed that species retaining ALKB2/3 have >10-fold higher median levels of DNA methylation than species which have lost ALKB2/3; this is true both in coding sequences and genome-wide ($p<0.01$, Wilcox Unpaired Test; Supplementary Figure 12A and B; Supplementary Table 9).

DNA methyltransferases introduce 3-methyl cytosine (3meC) alkylation damage into DNA

Overall, our analysis confirms robust and widespread coevolution between ALKB2/3 and DNMTs across eukaryotes. On the basis of this observation, we wondered about a possible mechanistic link between DNA methylation and the presence of alkylation DNA damage. The preferred substrates for ALKB2/3 within DNA are 1meA and 3meC22,23.. We thus wondered whether the activity of cytosine DNMTs might be associated with the generation of 3meC in addition to the major activity of these enzymes to produce 5meC. In order to test this we used synthetic nucleoside standards to develop an ultrasensitive mass spectroscopy (LC/MS) approach that enabled us to specifically distinguish and quantify 3meC and 5meC in DNA (Online Methods; Figure 5A, B and Supplementary Figure 13A). To further verify this detection method, we treated a plasmid with the mutagen MMS, which amongst other

lesions, is known to introduce 3meC into DNA. The LC/MS analysis confirmed that we could detect a robust induction of 3meC but no induction of 5meC (SupplementaryFigure 13B, C).

To further examine the possible association between DNMTs and 3meC we tested whether cytosine DNA methyltransferase activity might be sufficient to produce DNA alkylation damage *in vitro*. We carried out *in vitro* methyltransferase reactions using the recombinant catalytic domain of DNMT3a. The subsequent LC/MS analysis revealed a robust production of 5meC but also a clear evidence for 3meC induction (Figure 5C, D and Supplementary Figure 13D, E). The induction of 3meC is far less abundant and occurs in a ratio 1: 2850 3meC:5meC, i.e. 3meC = ~0.035% of 5meC (Figure 5C, D). To verify that this result required the catalytic activity of DNMT3a we expressed and purified the F646A point mutant of the catalytic domain of DNMT3a that shows reduced ability to bind the cofactor SAM (Supplementary Figure 14). Consistent with previous results24 we found that this enzyme has markedly reduced catalytic activity in introducing 5meC (Figure 5C). Importantly, this mutation also completely eliminated 3meC formation, demonstrating that catalytic activity is essential for DNMT3a to promote 3meC introduction (Figure 5D). Together these results suggest that DNMTs can use SAM to promote the introduction of 3meC at a low rate in addition to their usual 5meC product. Notably, the bacterial methyltransferase mSssI also introduced 3meC *in vitro* (Supplementary Figure 13C), suggesting that the introduction of 3meC is a general property of cytosine methyltransferases. It is possible that generation of 3meC involves a direct catalytic activity of the enzyme; alternatively DNMTs may promote this indirectly by flipping the base out from the double helix25 and positioning it near to the SAM.

To test whether DNMTs can promote introduction of 3meC also *in vivo* we used our LC/MS method to examine 3meC levels in mouse embryonic stem cells (ESCs) carrying DNMT1, DNMT3a and DNMT3b deletion (TKO)26. In wild type mouse ESCs we detected a clear signal for 3meC. Notably, the measured 3meC level was around 10-fold lower than the level measured in vitro (Fig5 C-F), consistent with the existence of endogenous DNA repair mechanisms capable of removing 3meC (Figure 5G and see below). Contrastingly, we were not able to detect any 3meC in TKO cells (Figure 5E); (p=0.0017 ANOVA; Figure 5F). As an independent validation, dot blots using an antibody specific for 3meC showed similar data (Supplementary Figure 14A, B). We thus conclude that the presence of active DNMT1 and DNMT3a/b is clearly associated with increased levels of 3meC in genomic DNA.

Mammalian ALKB2/3 have been shown to repair 3meC *in vitro* and in cultured mouse cells20,22,23. In order to test whether 3meC induced by DNMTs activity is processed by ALKB2 in ES cells we used the CRISPR-CAS9 system to target deletions to the 1st exon of ALKB2 in both WT and TKO cells (Supplementary Figure 15A). We obtained clones with homozygous deletions in both alleles of ALKB2, which showed a reduction in ALKB2 protein in both WT and TKO cells (Supplementary Figure 15B, C). Moreover, these clones showed increased sensitivity to the mutagen MMS relative to their parent line (p=0.042, ANOVA test for ALKB deletion; Supplementary figure 15D) consistent with disruption of ALKB2 function in repairing alkylation DNA damage. We next analysed 3meC levels and found that the loss of ALKB2 leads to a ~15% increase in steady-state 3meC levels (p=0.02,

ANOVA test for ALKB2 deletion; Figure 5F) confirming that ALKB2 is implicated in the removal of 3meC. Importantly, in TKO cells even the lack of ALKB2 did not raise the level of 3meC above the limit of our LC/MS quantification (Figure 5F). Overall, this data is consistent with a role of ALKB2 in removing 3meC associated with the activity of DNMTs *in vivo*.

The presence of 5meC in DNA is known to be mutagenic due to deamination of 5meC to T, resulting in depletion of CG dinucleotides over evolutionary time27,28.. 5meC to thymine deamination results in a G-T mismatch. However, alkylation damage such as 3meC poses a much more severe threat as 3meC blocks DNA polymerases involved in normal DNA replication29,30.. Thus our finding that 3meC is produced by DNMTs indicates that DNMT activity may directly cause replication stress in cells. On the basis of the average GC composition of the mouse genome we calculated that the level of 3meC we observe *in vivo* corresponds to ~5 modified cytosines in every 10^6 base pairs. The most common form of endogenous DNA damage known is the formation of abasic sites through cytosine deamination and subsequent uracil excision as well as spontaneous depurination20. This form of DNA damage has a profound effect in shaping nucleotide frequencies through evolutionary processes28. Abasic sites have been measured in cultured cells and tissues with estimates ranging from 1-20 nucleotides per 1e6 base pairs31. Our results show that 3meC introduced by the off-target activity of DNMTs exists at similar levels to abasic sites and is thus one of the most abundant forms of spontaneous DNA damage yet known in cells.

Discussion

Our study documents that DNA methylation is a rapidly evolving epigenetic system. Our findings show that although *C.elegans* and other nematodes lost DNA methylation system, other nematode species contain combinations of DNMTs homologous to the mammalian DNMT1 and DNMT3 enzymes that install genomic DNA methylation in these species. We further show that, at least in nematodes, DNA methylation is primarily targeted to repetitive elements in the genome.

Crucially, our evolutionary analysis of DNA methylation has revealed an unexpected coevolution between DNA methylation and DNA repair systems. Our data shows that DNA methyltransferase activity is associated with the generation of 3meC both *in vitro* and *in vivo* and that ALKB2 demethylase is required to process this type of alkylation damage. We suggest that the relatively high level of endogenous DNA damage introduced by this off-target activity of DNMTs explains why ALKB2/3 is generally needed in organisms with 5meC (Figure 5G). Notably, even in the presence of ALKB2/3, 3meC introduction by DNMTs is likely to pose a threat to genome stability by causing DNA polymerases to stall thus leading to the appearance of ds DNA breaks. In line with this possibility, members of the BRCA complex and Rad18 both of which are important in DNA double strand break repair32 coevolve with DNMTs (Supplementary Figure 6; Supplementary Table 5).

Although the future investigation into the relationship between DNA methylation and DNA repair may reveal additional mechanistic links, our data shows that the propensity of cytosine DNMTs to induce alkylation damage may be an important factor explaining the

Europe PMC Funders Author Manuscripts

frequent independent losses of DNA methylation across different animal groups. Last, our data provides an important example of how analysis of the evolutionary relationships between proteins can enable identification of novel biochemical mechanisms.

Online Methods

Nematode Collection and DNA isolation

Romanomermis culicivorax adults were a gift from C. Kraus (University Köln) and derived from the culture of E. Platzer (University of California Riverside). *Trichinella spiralis* animals were prepared according to standard methodology. *Trichuris muris* adults were collected using fine forceps from the ceca SCID of mice orally infected 42 days previously with 400 embryonated eggs.

Plectus sambesii animals were grown on low salt agar with semiliquid HB101 at 25°C. Adults were isolated from mixed stage cultures by sorting on a COPAS large particle sorter.

Analysis of DNA methylation sequencing

We assembled a draft *P. sambesii* genome from Illumina Short Read sequencing (See below). Other genomes were taken from Wormbase (*C. briggsae*, WS240), Wormbase Parasite (*N. brasiliensis*, *T. spiralis*, *T. muris* WBPS4) or Nembase (*R. culicivorax*). Libraries for bisulfite sequencing were prepared using Pico Methyl-Seq kit (Zymo Research). Bisulfite sequencing reads were mapped using Bismark, using the bowtie2 option. To obtain the methylation levels for different CG contexts and for different categories of genomic annotation (Figure 2) we used the Bismark methylation extractor module to convert bismark alignments into genome-wide coverage files reporting the methylation status. As the Bismark methylation extractor can only operate on a small number of contigs, prior to alignments we had to artificially condense all genomes except that of *C. briggsae* (which is already assembled into 6 chromosomes) into 10 “pseudo-contigs” without disrupting the sequence of the contigs themselves. Subsequently, we selected cytosines covered by at least 10 reads using a custom perl script for further analysis. We converted genome coordinates from pseudo-contigs back to the original contigs and annotated individual CpG sites according to gene predictions, either our own (*P. sambesii*) or taken from Wormbase (WBPS4; WS245) and repeats annotated using RepeatMasker using the parameters –no-low, --no-is, --species animalia, using Bedtools34. We then obtained percentage methylation by summing the methylated reads and the unmethylated reads across all CG sites within different regions.

Statistical enrichment of CG methylation was calculated using the Fisher’s exact test comparing the number of methylated sites and the number of unmethylated sites in both of the genomic regions of interest (i.e. genes vs entire genome).

In order to analyse the distribution across genes and TEs in more detail (Figure 2D and Figure 3) we again used Bismark to align DNA methylation sequencing data to contigs directly to avoid artefacts potentially caused by joining contigs together in the middle of repeats or genes. We then used MethylExtract to obtain site-specific methylation information and converted the output to bed files using a custom perl script. Bedtools was used to

annotate CG sites as above and the mean methylation across individual features (e.g. repeat, gene etc) was calculated by averaging across the fractional methylation at each site with >10 reads coverage within the feature. Features with >=5 CGs covered were used to draw Figures 2D and Figures 3. All statistical analyses and graphics were performed in the R environment.

Identification of DNMTs in nematode genomes

We searched the predicted proteins from nematode genomes for Cytosine methyl-transferase domains using pfam hmm-search with the Cytosine-5-methyltransferase domain. All proteins with matches to this domain were extracted. We then used blast to compare these against human DNMT1, 2 and 3 to annotate potential methyltransferases.. Any proteins that did not match to DNMT1, 2 or 3 were tested against the Uniprot database- this revealed them to be bacterial contaminants and they were removed from further analysis. We verified these annotations by phylogenetic analysis: nematode DNMTs along with selected DNMTs from other metazoans were aligned using MUSCLE, and these alignments were used to construct a phylogenetic tree according to the workflow in Phylogeny.fr. Domains within the nematode DNMTs were identified using pfam searches with the seeds for PWWP, BAH and the cytosine DNA methyltransferase domain. We could not find clear evidence for the CXXC domain in any of the nematode DNMTs, but this could be because of poor assembly of the genome in the N terminus of the protein in *R. culicivorax* and *P. sambesii* both of which fall near to the boundary of contigs.

Coevolution analysis

We used Blast using ALKB1-8 to analyse the conservation of ALKB proteins across the nematodes. The e-value of the best hit was tabulated. To identify co-evolving proteins across Ensembl Metazoa we downloaded each predicted proteome from Ensembl (release 28). We ran blastp using the human proteome as the query sequence and each predicted proteome as a database, retaining the best blast hit. Proteins with a blast hit log10 e-value less than -20 were given a score of 1 and those greater than -20 given a score of 0 to build a binary conservation matrix. We then used a Fisher's exact test to identify proteins with a significant tendency to be lost or gained with DNMT3 or DNMT1 using the Benjamani and Hofberg multiple test correction. Gene ontology information for all the human Uniprot database was downloaded using BiomaRt and significantly enriched categories were identified using a Fisher's Exact test following multiple test correction. To test further for coevolution between presence of ALKB2/3 and DNMTs we used a modified phylogenetic profiling method. We first used reciprocal blast to test for the presence of ALKB2/3 (retaining any hit that blasted reciprocally to ALKB2 or 3, including examples where the best blast hit for ALKB2 blasted back to ALKB3 and vice versa), DNMT1 and DNMT3 in all metazoan, fungal and protozoan genomes downloaded from Ensembl. To ensure we retained data only for phylogenetically independent loss events we constructed phylogenetic trees for these groups using the references detailed in the supplemental information (Supplementary Note 2). Finally we mapped loss of ALKB2, DNMT1 and DNMT3 and collated these for each group before testing for co-occurrence of ALKB2 and one or more of DNMT1 or DNMT3 using Fisher's Exact Test.

Analysis of DNA methylation levels across arthropods

For all species except *S. gregaria* we obtained estimates of mCG/CG genome-wide and at coding sequences directly from the relevant references9,17,21. For *S. gregaria* the published reference16 did not report a genome wide mCG/CG estimate as only coding sequences have been sequenced fully in this organism, thus we used the FastMC algorithm9 to estimate genome-wide mCG/CG directly from raw sequencing data and calculated the coding sequence mCG/CG methylation level direct from the reference. We searched for conservation of ALKB2 in these species using the reciprocal blast method described above.

Dot blot analysis of methylation in genomic DNA samples

DNA was extracted using the Qiagen DNA blood/Tissue isolation kit and redissolved in distilled water. DNA was diluted 50:50 with freshly prepared 0.2M NaOH and heated for 5 minutes at 95°C to denature. 2ul DNA was then spotted onto a nitrocellulose membrane and air-dried before cross-linking with a Stratalinker. The membrane was blocked with 5% milk in Tris buffered saline. Anti-3meC (Active motif) used at a 1/5000 dilution or anti-5meC (Clone 33D10, Abcam or Active Motif) at dilution 1/2500 was added for an overnight incubation in 1% milk in TBS with 0.1% Tween-20. The membrane was washed and exposed to appropriate secondary antibody for 2hrs at room temperature before developing with ECL.

A positive control for 3meC was prepared by incubating polydIdC in the presence of 20mM MMS (Sigma) for 4 hours at 37°C. Excess MMS was quenched by addition of 0.2M NaOH before dot blot analysis.

Positive and negative controls for 5meC, PCR products from the APC promoter made either with 5meCTP or CTP, were purchased from Active motif.

Liquid chromatography-mass spectrometry

N^3-methyl-2'-deoxycytidine (3meC) standards were purchased from ChemGenes ; 2'-deoxycytidine (dC) and 2'-deoxyguanosine (dG) were purchased from Berry & Associates; C^5-methyl-2'-deoxycytidine (5meC) was purchased from CarboSynth. Genomic DNA or synthetic oligonucleotides were digested to nucleosides for a minimum of 9 hours at 37°C using a digestion enzymatic mix (kind gift from NEB). All samples and standard curve points were spiked with a similar amount of isotope-labelled synthetic nucleosides: 100 fmol of dC* and dG* purchased from Silantes, 5 fmol of 5meC* obtained from T. Carell (Center for Integrated Protein Science at the Department of Chemistry, Ludwig-Maximilians-Universität München, München, Germany). The nucleosides were separated on an Agilent RRHD Eclipse Plus C18 2.1 × 100 mm 1.8u column by using the HPLC 1290 system (Agilent) and analysed using Agilent 6490 triple quadrupole mass spectrometer. Quantification was carried out in multiple reaction monitoring mode (MRM) by monitoring specific transition pair of *m/z* 250.1/134.1 for dC, 290.1/174.1 for dG, 264.1/148.1 for 5mC and 242.2/95.1 for 3meC. To calculate the concentrations of individual nucleosides (for dC, dG and 5meC), standard curves representing the ratio of the peak response of known amounts of synthetic nucleosides and the peak response of the isotope-labelled nucleosides were generated and used to convert the peak-area values to corresponding concentrations.

Europe PMC Funders Author Manuscripts

For 3meC, the concentrations were calculated directly using a standard curve with light nucleosides. The threshold for peak detection is a signal-to-noise (calculated with a peak-to-peak method) above 10, and the limit of quantification (LOQ) was 25 amol for 5mC and 50 amol for 3mC. Final measurements were normalized by dividing by the dG level measured for the same sample. The detectable limit was calculated by dividing the minimum detected value by the dG level for each sample.

DNA methylation in vitro

Unmethylated plasmid was prepared from DAM/DCM – *E. coli* cells. For mSSSI-methylation we used a pUC19 plasmid and after purification by MaxiPrep (Qiagen) we treated with mSSSI (NEB) for 1hr at 37°C. To induce alkylation damage we exposed unmethylated pUC19 plasmid to 20mM MMS (Sigma) for 1 hour at 25°C before purification. DNMT3A and mutants thereof was expressed and purified from *E. coli* cells as described previously35. The reaction mixture was incubated for 2hrs at 37°C and DNA was purified using phenol-chloroform extraction and analysed using LC-MS as described above.

Plectus genome sequencing and assembly

We assembled and annotated a genome for *P. sambesii* using Illumina high-throughput sequencing data and using the methods documented in the supplemental material section (Supplementary Note 1). The final genome had a span of 186Mbp and an N50 of 4039bp comparing well with other nematode genomes used in this study. The genome has been deposited in NCBI (PRJNA390260).

Generation and validation of ALKB2 deletion mutants

We obtained plasmids containing GFP-tagged CRISPR-CAS9 and guide RNAs targeting the first protein-coding exon of ALKB2 from Sigma. We used Lipofectamine transfection to introduce this plasmid into mouse ES cells and after recovery for 18 hours at 37°C, we used FACS to sort GFP positive cells into individual wells of a 96 well plate. We screened the resultant clones for ALKB2 using PCR across the targeted exon searching for apparent size shifts. We then used Sanger sequencing of the PCR products to select clones showing indels in both alleles. We confirmed ALKB2 protein reduction using Western blot with anti-ALKBH2, a mouse monoclonal antibody (C-9; Santa Cruz) using a Rabbit anti-mouse HRP-conjugated secondary antibody (Abcam). To test sensitivity to MMS, cells were treated with 200mM MMS for one hour before the MMS was washed out. We then sorted single cells using FACs and counted colonies formed after 5 days, comparing to a control treated with 0mM MMS for each line.

Data availability

Bisulfilte sequencing data has been deposited to the Gene Expression Omnibus (GEO, www.ncbi.nlm.nih.gov/geo/), with accession number GSE104339. Mouse, *B. mori* and *P. hawaiensis* methylation data is available from GEO (GSE18315, *B. mori*; GSE61457, Mouse ESCs; GSE82141, *P. hawaiensis*). The *P. sambesii* genome assembly has been deposited to NCBI (PRJNA390260). Other nematode genomes are available from Wormbase and Wormbase ParaSite (www.sanger.ac.uk/science/tools/wormbase). Metazoa, fungal and

protist genomes are available from Ensembl (http://ensemblgenomes.org/). The genome of *P. hawaiensis* is available from NCBI (https://www.ncbi.nlm.nih.gov/genome/?term=hawaiensis)

Code accessibility statement

DNA methyltransferase annotation: hmmer (version 3.1) freely available from hmmer.org/; blast+ (version 2.2.30) freely available from https://blast.ncbi.nlm.nih.gov; Phylogenetic tree construction: MUSCLE v3.8.31 for alignment, Gblocks 0.91b for curation and PhyML 3.1 for maximum likelihood phylogeny, all provided via www.Phylogeny.fr. Bisulfite alignment and mapping bismark version 0.14.2 freely available from https://www.bioinformatics.babraham.ac.uk/projects/bismark/; bowtie2 (version 2.1.0) freely available from bowtie-bio.sourceforge.net/bowtie2/; Methylextract version 1.9 freely available from https://github.com/bioinfoUGR/methylextract?files=1. Bedtools (version 2.19.0) was used for data integration; freely available from http://bedtools.readthedocs.io/en/latest/.

Coevolution analysis: blast+ version 2.2.30

All statistical analysis was carried out using R (version 3.1.0); freely available from https://www.r-project.org/.

Custom perl scripts (Perl version 5.16) used for intermediate processing DNA methylation data are available on request.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

We thank H. Leitch and M. Borkowska for invaluable help with mouse ES cell culture. We would like to thank M. Merkenschlager, L. Aragon, J. Sale and B. Lehner for helpful comments on the manuscript, M. Blaxter for advice on nematode genomics and M. Berriman for access to the *N. brasiliensis* draft genome. PS is funded by an Imperial College Research Fellowship. Work in the Sarkies and Hajkova laboratories is funded by the Medical Research Council. P.H is a recipient of the ERC CoG grant “dynamicmodifications” and a member of the EMBO Young Investigator Programme. RKG and AJB are funded by Wellcome Trust grant 083620Z and centre grant 203128/Z/16/Z. PHS is funded by the ERC in a grant to Max Telford (ERC-2012-AdG 322790).

References

1. Law JA, Jacobsen SE. Establishing, maintaining andmodifying DNA methylation patterns in plants and animals. Nat Rev Genetics. 2010; 11:204–220. [PubMed: 20142834]
2. Holliday R. Epigenetics: A Historical Overview. Epigenetics. 2014; 1:76–80.
3. Ponger L. Evolutionary Diversification of DNA Methyltransferases in Eukaryotic Genomes. Mol Biol Evol. 2005; 22:1119–1128. [PubMed: 15689527]
4. Jurkowski TP, Jeltsch A. On the evolutionary origin of eukaryotic DNA methyltransferases and Dnmt2. PLoS ONE. 2011; 6:e28104. [PubMed: 22140515]
5. Zemach A, McDaniel IE, Silva P, Zilberman D. Genome-Wide Evolutionary Analysis of Eukaryotic DNA Methylation. Science. 2010; 328:916–919. [PubMed: 20395474]
6. Feng S, et al. Conservation and divergence of methylation patterning in plants and animals. Proc Nat Acad Sci. 2010; 107:8689–8694. [PubMed: 20395551]

7. Lyko F, et al. The honey bee epigenomes: differential methylation of brain DNA in queens and workers. Plos Biol. 2010; 8:e1000506. [PubMed: 21072239]
8. Xiang H, et al. Single base-resolution methylome of the silkworm reveals a sparse epigenomic map. Nat Biotechnol. 2010; 28:516–520. [PubMed: 20436463]
9. Bewick AJ, Vogel KJ, Moore AJ, Schmitz RJ. Evolution of DNA Methylation across Insects. Mol Biol Evol. 2016; msw264. doi: 10.1093/molbev/msw264
10. Raddatz G, et al. Dnmt2-dependent methylomes lack defined DNA methylation patterns. Proc Nat Acad Sci. 2013; 110:8627–8631. [PubMed: 23641003]
11. Goll MG. Methylation of tRNAAsp by the DNA Methyltransferase Homolog Dnmt2. Science. 2006; 311:395–398. [PubMed: 16424344]
12. Gao F, et al. Differential DNA methylation in discretedevelopmental stages of the parasitic nematodeTrichinella spiralis. Genome Biol. 2012; 13:R100. [PubMed: 23075480]
13. Schiffer PH, et al. The genome of Romanomermis culicivorax: revealing fundamental changes in the core developmental genetic toolkit in Nematoda. BMC Genomics. 2013; 14:923. [PubMed: 24373391]
14. Li Z, et al. Distinct roles of DNMT1-dependent and DNMT1-independent methylation patterns in the genome of mouse embryonic stem cells. Genome Biol. 2015; 16:1. [PubMed: 25583448]
15. Huff JT, Zilberman D. Dnmt1-Independent CG Methylation Contributes to Nucleosome Positioning in Diverse Eukaryotes. Cell. 2014; 156:1286–1297. [PubMed: 24630728]
16. Falckenhayn C, et al. Characterization of genome methylation patterns in the desert locust Schistocerca gregaria. J Exp Biol. 2013; 216:1423–1429. [PubMed: 23264491]
17. Kao D, et al. The genome of the crustacean Parhyale hawaiensis, a model for animal development, regeneration, immunity and lignocellulose digestion. eLife. 2016; 5
18. Sarkies P, et al. Ancient and novel small RNA pathways compensate for the loss of piRNAs in multiple independent nematode lineages. Plos Biol. 2015; 13:e1002061. [PubMed: 25668728]
19. Ougland R, Rognes T, Klungland A, Larsen E. Non-homologous functions of the AlkB homologs. J Mol Cell Biol. 2015; 7:494–504. [PubMed: 26003568]
20. Sedgwick B. Repairing DNA-methylation damage. Nat Rev Mol Cell Biol. 2004; 5:148–157. [PubMed: 15040447]
21. Strepetkaitė D, et al. Analysis of DNA Methylation and Hydroxymethylation in the Genome of Crustacean Daphnia pulex. Genes. 2016; 7:1.
22. Ringvoll J, et al. Repair deficient mice reveal mABH2 as the primary oxidative demethylase for repairing 1meA and 3meC lesions in DNA. The EMBO Journal. 2006; 25:2189–2198. [PubMed: 16642038]
23. Nay SL, Lee D-H, Bates SE, O'Connor TR. Alkbh2 protects against lethality and mutation in primary mouse embryonic fibroblasts. DNA Repair (Amst). 2012; 11:502–510. [PubMed: 22429847]
24. Gowher H, et al. Mutational Analysis of the Catalytic Domain of the Murine Dnmt3a DNA-(cytosine C5)-methyltransferase. J Mol Biol. 2006; 357:928–941. [PubMed: 16472822]
25. Klimasauskas S, Kumar S, Roberts RJ, Cheng X. HhaI methyltransferase flips its target base out of the DNA helix. Cell. 1994; 76:357–369. [PubMed: 8293469]
26. Tsumura A, et al. Maintenance of self-renewal ability of mouse embryonic stem cells in the absence of DNA methyltransferases Dnmt1, Dnmt3a and Dnmt3b. Genes to Cells. 2006; 11:805–814. [PubMed: 16824199]
27. Sved J, Bird A. The expected equilibrium of the CpG dinucleotide in vertebrate genomes under a mutation model. 1990; 87:4692–4696.
28. Alexandrov LB, et al. Clock-like mutational processes in human somatic cells. Nature. 2015; 47:1402–1407.
29. Drabløs F, et al. Alkylation damage in DNA and RNA—repair mechanisms and medical significance. DNA Repair (Amst). 2004; 3:1389–1407. [PubMed: 15380096]
30. Furrer A, van Loon B. Handling the 3-methylcytosine lesion by six human DNA polymerases members of the B-, X- and Y-families. Nucleic Acids Res. 2013; 42:553–566. [PubMed: 24097443]

31. Chastain PD, et al. Abasic sites preferentially form at regions undergoing DNA replication. FASEB J. 2010; 24:3674–3680. [PubMed: 20511393]
32. Shrivastav M, De Haro LP, Nickoloff JA. Regulation of DNA double-strand break repair pathway choice. Cell Res. 2008; 18:134–147. [PubMed: 18157161]
33. Blaxter ML, et al. A molecular evolutionary framework for the phylum Nematoda. Nature. 1998; 392:71–75. [PubMed: 9510248]
34. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics. 2010; 26:841–842. [PubMed: 20110278]
35. Emperle M, Rajavelu A, Reinhardt R, Jurkowska RZ, Jeltsch A. Cooperative DNA binding and protein/DNA fiber formation increases the activity of the Dnmt3a DNA methyltransferase. J Biol Chem. 2014; 289:29602–29613. [PubMed: 25147181]

Europe PMC Funders Author Manuscripts

Europe PMC Funders Author Manuscripts

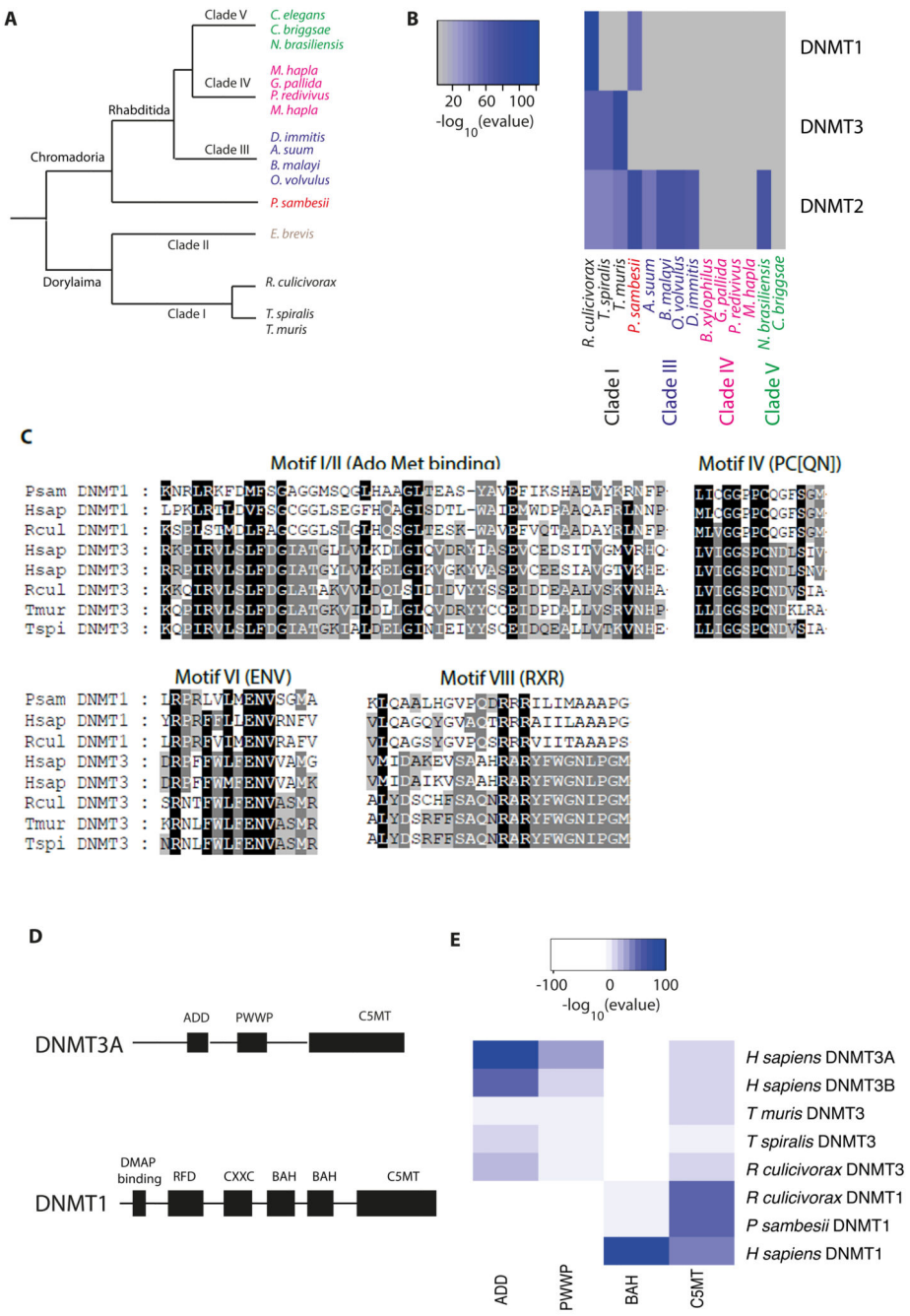

Figure 1.

Analysis of DNA methyltransferases in nematodes. **A** shows a cladogram of nematodes including the species profiled in this study. Clade nomenclature and phylogenetic positions are taken from Blaxter et al., 199833. **B** presence of DNMTs in nematodes as assessed by reciprocal blast. Grey indicates no best reciprocal blast hit. **C** Multiple sequence alignment showing key motifs important for DNMT activity in nematode DNMTs along with human DNMT1 and DNMT3 for comparison. **D** Domains within DNMT1 and DNMT3 from human. **E** conservation of domains within nematode DNMTs as assessed by comparison to

the pfam seed of each domain. Domains from Figure 1D that are in the Pfam database are shown and found in at least one nematode DNMT are shown. The N-terminal regions of DNMT1 from *R. culicivorax* and *P. sambesii* are missing precluding definitive assessment of CXXC presence or absence; this is probably due to incomplete genome assembly (see Online Methods for more information)

Europe PMC Funders Author Manuscripts

Europe PMC Funders Author Manuscripts

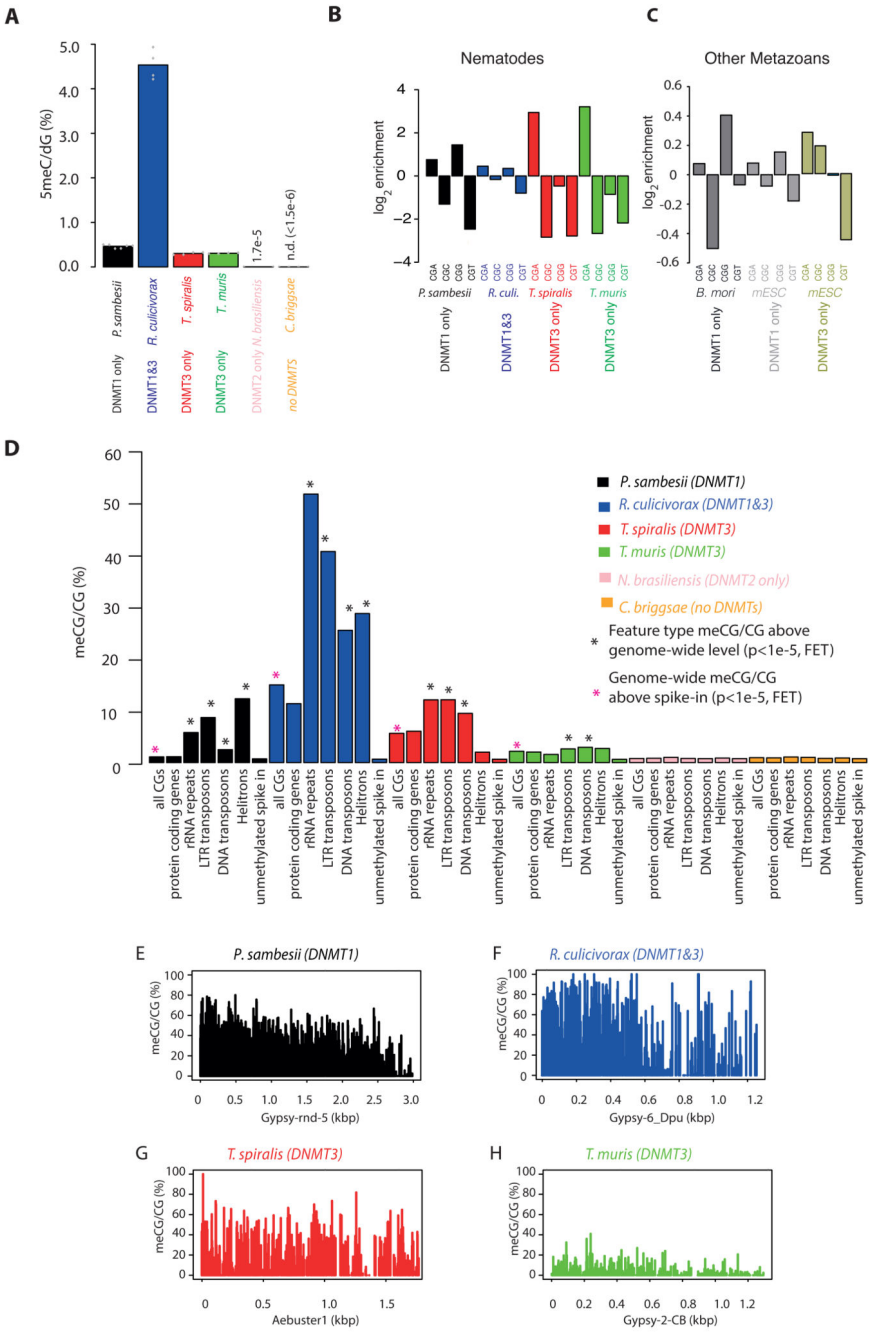

Figure 2.

Genome-wide DNA methylation analysis of nematodes. **A** shows quantification of 5meC in DNA by LC/MS for different nematode species. Bar lines are at the mean and the standard deviation is shown by error bars. Each point overlayed shows the mean of two technical repeats, for n independent DNA extractions (n=6, *P. sambesii*; n=4, *R. culicivorax*, T. spiralis, T. muris, C. briggsae; n=2, *N. brasiliensis*). **B** and **C** show the overall fraction of sites with >10% methylation for each of the specified CG containing trinucleotides. Total number of CGs analysed is in Supplemental table 3. **D** shows the average methylation of

each CpG within different annotated regions, compared to unmethylated spike-in. The total number of CGs analysed is in Supplementary Note 3. **E-H** shows individual examples of repeat element consensus sequences with high levels of DNA methylation.

Europe PMC Funders Author Manuscripts

Europe PMC Funders Author Manuscripts

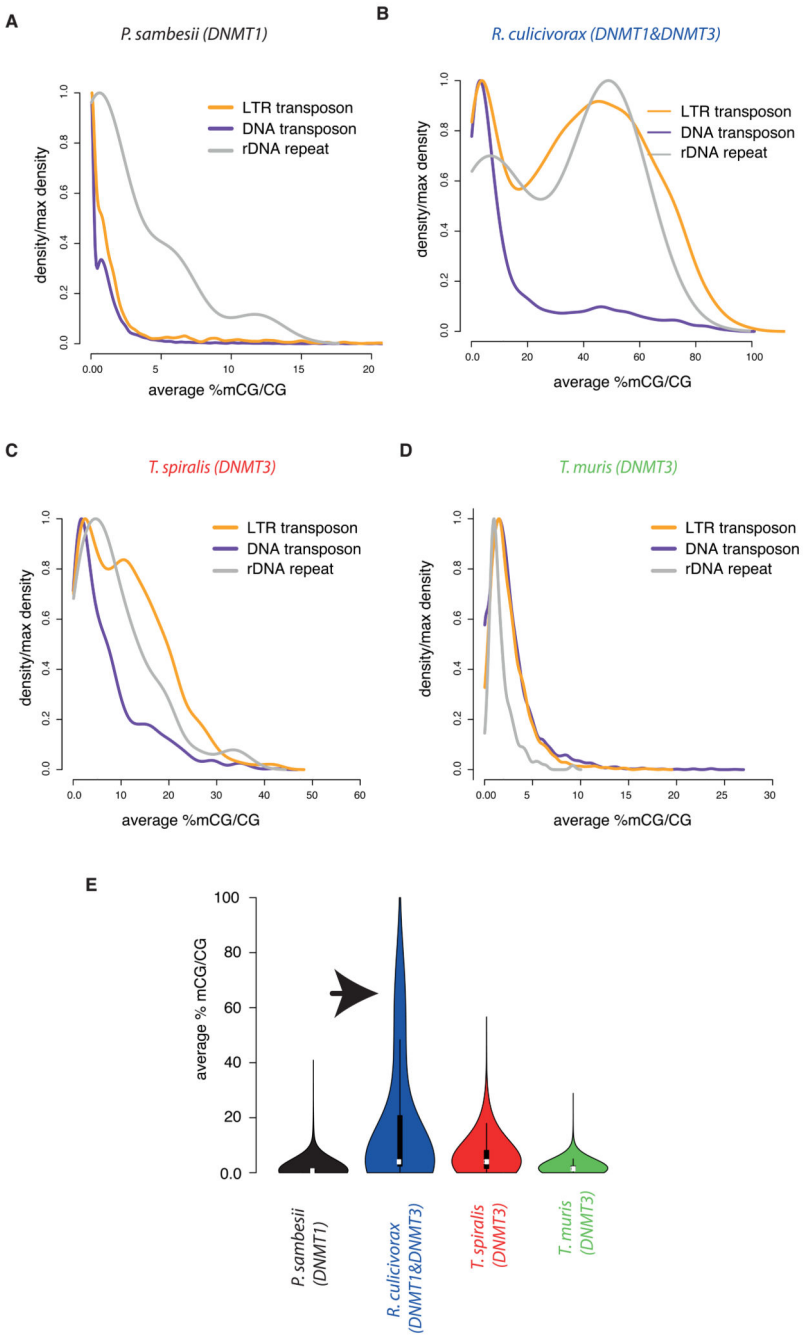

Figure 3.

A-D show histograms of methylation levels averaged across the body of different genomic features. **E** shows violin plots of methylation levels of all genes across the different species with the subset of genes carrying high levels of DNA methylation observed in *R. culicivorax* indicated by an arrow. The dot is at the median, box shows interquartile range, and the whiskers extend to the greatest point that is no more than 1.5 times the interquartile range.

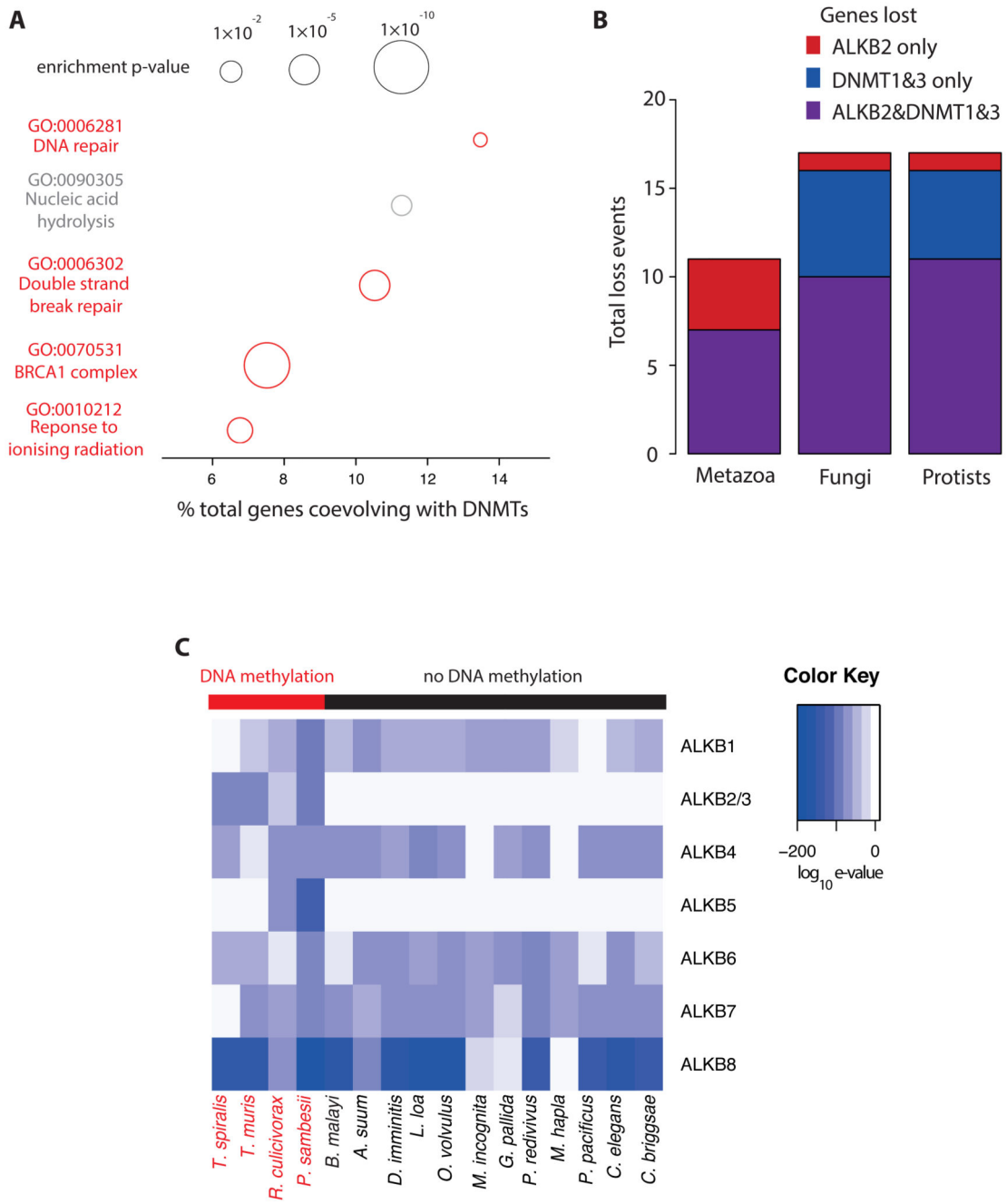

Figure 4.

A shows the top 5 statistically significantly enriched GO terms within the set of genes that coevolve with DNMTs across metazoans. Terms associated with DNA repair are shown in red, others in grey. **B** shows the conservation of ALKB2 along with DNMTs across different taxonomic groups. Losses amongst N independent branches are shown, where N=20, metazoa; N=31, Fungi; N=23, Protists. **C** shows the conservation of different members of the ALKB family in nematodes with and without DNA methylation.

Europe PMC Funders Author Manuscripts

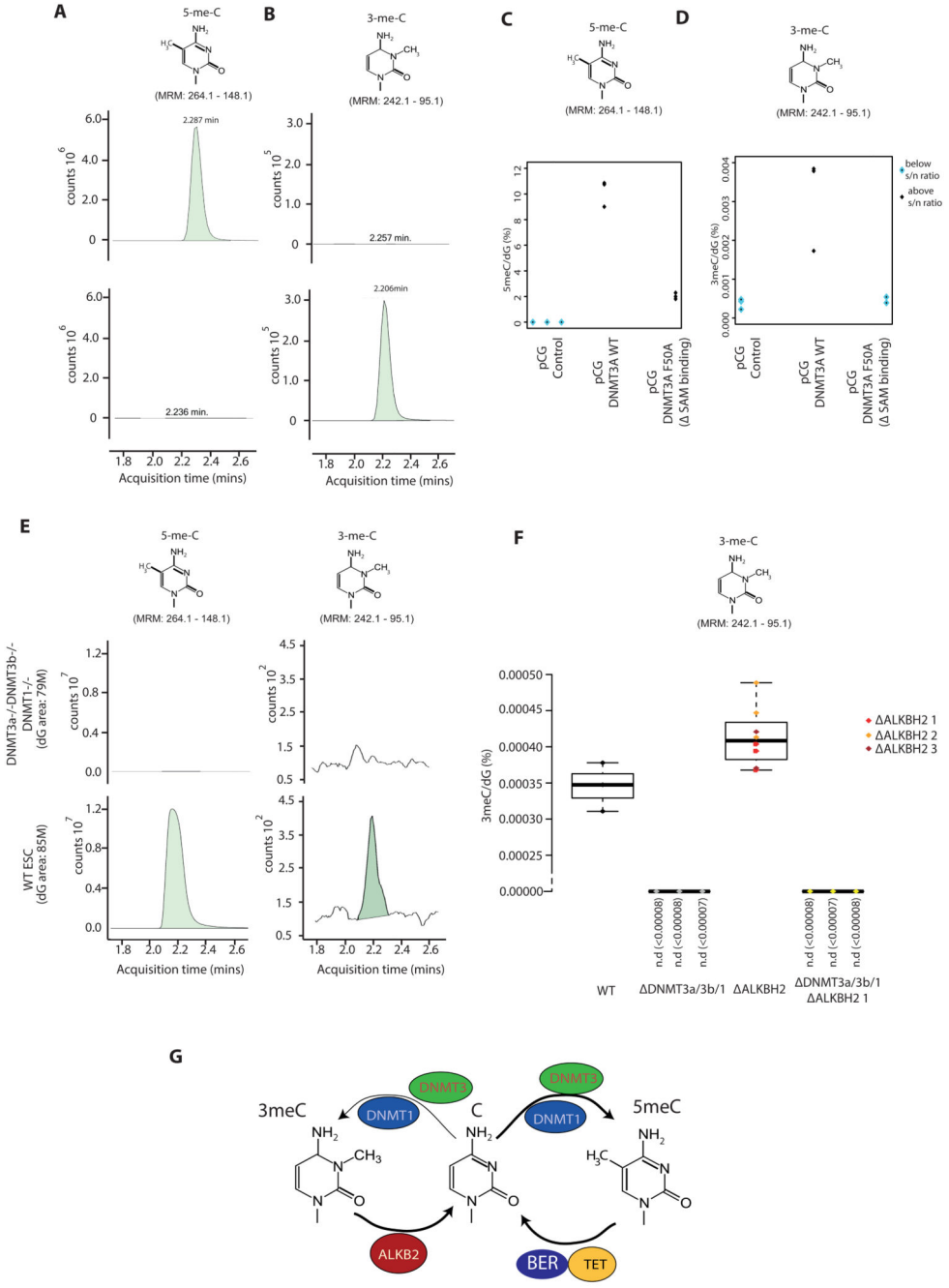

Figure 5.

DNA alkylation damage in DNA associated with DNMT activity. **A, B** Validation of method to detect 3meC specifically in the presence of 5meC using LC/MS. **C, D** LC/MS measurement of 3meC introduced by the catalytic domain of DNMT3a *in vitro* compared to 3meC induction by the F646A mutant, which does not bind the SAM cofactor. Each of the 3 individual points for each sample shows the mean of two technical replicates for an independent in vitro reaction. Measurements below the signal to noise ratio are shown in cyan. **E** Example LC/MS traces for 3meC and 5meC for ESCs with or without DNA

Europe PMC Funders Author Manuscripts

methyltransferases. Screenshots of the LC/MS analysis are shown. Colours for peaks are automatically assigned by the software on the basis of the peak settings**. F** LC/MS analysis of 3meC in mouse ES cells with and without DNMTs and ALKBH2 (the Ensembl gene name of the mouse ALKB2 orthologue). The boxplots show interquartile range of 3meC normalized to dG, with a line at the median, and whiskers extending to the furthest point within 95% of the range. Each of the 3 points for each cell line shows the mean of two technical replicates for independent DNA extractions. **G** model for how DNMTs influence methylation on different positions of cytosine.