Bacterial N4-methylcytosine as an epigenetic mark in eukaryotic DNA

Fernando Rodriguez

Marine Biological Laboratory <https://orcid.org/0000-0003-4044-8734>

Irina Yushenova

Marine Biological Laboratory <https://orcid.org/0000-0001-6291-6215>

Daniel DiCorpo

Marine Biological Laboratory

Irina Arkhipova (✉ iarkhipova@mbi.edu)

Marine Biological Laboratory <https://orcid.org/0000-0002-4805-1339>

Article

Keywords: non-canonical modifications, amino-methyltransferase, horizontal gene transfer, transposable elements, retrotransposons, epigenetic silencing

Posted Date: August 16th, 2021

DOI: <https://doi.org/10.21203/rs.3.rs-360382/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.
[Read Full License](#)

Version of Record: A version of this preprint was published at Nature Communications on February 28th, 2022. See the published version at <https://doi.org/10.1038/s41467-022-28471-w>.

Bacterial N4-methylcytosine as an epigenetic mark in eukaryotic DNA

Fernando Rodriguez^{1,3}, Irina A. Yushenova^{1,3}, Daniel DiCorpo^{1,2}, Irina R. Arkhipova^{1*}

¹Josephine Bay Paul Center for Comparative Molecular Biology and Evolution, Marine Biological Laboratory, Woods Hole, MA 02543, USA

²Present address: Department of Biostatistics, Boston University School of Public Health, Boston, MA 02118, USA

³These authors contributed equally to this work

*Correspondence: iarkhipova@mbi.edu (I.A.)

Keywords: non-canonical modifications; amino-methyltransferase; horizontal gene transfer; transposable elements; retrotransposons; epigenetic silencing

In eukaryotes, 5-methylcytosine is the predominant DNA base modification, followed by N6-methyladenine. However, N4-methylcytosine (4mC) is confined to bacteria. Here we report that 4mC can serve as an epigenetic mark in eukaryotes. Bdelloid rotifers, freshwater invertebrates with transposon-poor genomes that are rich in foreign genes, lack C5-methyltransferases but encode an amino-methyltransferase, N4CMT, captured from bacteria >60 Mya. N4CMT introduces 4mC into DNA, and its chromodomain shapes the “histone-read-DNA-write” architecture together with a “DNA-read-histone-write” SETDB1/eggless H3K9me3 histone methyltransferase variant preferentially binding 4mC-DNA, to maintain 4mC and silent chromatin at transposons and tandem repeats. Our results bring the third base modification into the eukaryotic repertoire, demonstrate how non-native DNA methyl groups can reshape complex epigenetic systems to suppress transposon proliferation, and establish horizontal gene transfer as the source of regulatory innovation in eukaryotes.

Modification of nucleobases without changes in the underlying genetic code offers unmatched opportunities for “writing” extra information onto DNA, the primary carrier of hereditary material. Covalent association of modifying groups with DNA provides advantages over more easily removable carriers of epigenetic information, such as RNA or proteins, for potential transmission across cell divisions and generations. In bacteria and archaea, DNA modifications are first and foremost associated with restriction-modification (R-M) systems acting to discriminate and destroy the invading foreign DNA, although multiple “orphan” methyltransferases (MTases) may perform regulatory functions ^{1,2}. Eukaryotes mostly use base modifications for regulatory purposes, with the predominant form of epigenetic modification in eukaryotic genomes being C5-methylcytosine (5mC) with its derivatives ^{3,4}. Often called “the fifth base”, 5mC plays an important role in genome defense against mobile genetic elements, and is often associated with transcriptional silencing, establishment of the closed chromatin configuration, and repressive histone modifications ⁵. The 5mC mark is introduced by C5-MTases, DNMT1 and DNMT3, thought to have originated from bacterial C5-MTases in early eukaryotes *via* fusions with additional domains interacting with proteins and DNA ⁶, while DNMT2 acts primarily on tRNA ^{7,8}. Recently, another modified base, N6-methyladenine (6mA), gained attention as a novel form of epigenetic modification in diverse eukaryotes ⁹⁻¹¹. In 6mA, the methyl group is added to the exocyclic amino group of adenines by amino-MTases, some of which are related to RNA-modifying MTases ^{10,12}. However, the third type of DNA methylation naturally occurring in bacteria, the N4-methylcytosine (4mC), has not been convincingly demonstrated in eukaryotes

¹³, and earlier claims of 4mC existence in eukaryotes could neither provide confirmation by orthogonal methods nor identify the corresponding enzymatic component. Here, we combine multiple lines of evidence to establish the first case of 4mC modification in eukaryotic genomes, investigate its recruitment as an epigenetic mark, and characterize the underlying enzymatic machinery, revealing how a horizontally transferred gene can become established in a complex regulatory network and maintained by selection over tens of millions of years of evolution.

Results

A bacterial amino-MTase in bdelloid rotifers. Rotifers of the class Bdelloidea are tiny freshwater invertebrates a fraction of a millimeter long, characterized by clonal reproduction, eutely, direct development, syncytial tissues, and paleotetraploid genome structure ¹⁴. They are known for an unmatched ability to incorporate foreign genes into genomic DNA, largely preserving their functionality ¹⁵. In sequenced bdelloids, 8-12% of coding sequences are of non-metazoan, mostly bacterial, origin ¹⁶⁻¹⁸. Surprisingly, we found that one such bacterial gene in the sequenced bdelloid *Adineta vaga* ¹⁶ is represented by an allelic pair of MTases containing the *N6_N4_MTase* domain (PF01555), which is closely related to amino-MTases of bacterial R-M systems acting on the exocyclic amino group of adenines and cytosines (Fig. 1a). Its orthologs, sharing the same four conserved intron positions, are present in sequenced representatives of each major family of the class Bdelloidea, dating back 40-60 Mya, but are absent from sequenced members of the sister class Monogononta or from any other sequenced eukaryotes (Fig. 1e,f). Both classes, however, encode each of the three MTase types previously implicated in adding 6mA marks to eukaryotic DNA: METTL4-like (PF05063: *MT-A70*), N6AMT1-like (PF05175: *MTS*) and N6AMT2-like (PF10237: *N6-adenineMlase*) ^{12,19-22} (Fig. 1b,f). Notably, none of the sequenced rotifers encode the most common eukaryotic C5-MTases *Dnmt1* or *Dnmt3*, harboring only the tRNA-modifying *Dnmt2/Trdmt*.

The *A. vaga* *N6_N4_MTase* belongs to the permuted type, in which the catalytic domain is located N-terminally to the S-adenosylmethionine (AdoMet) binding domain ²³ (Fig. 1a). Its evolutionary history (Fig. 1e) differs dramatically from that of 5mC- or N6A-MTases ⁶. The small non-permuted pan-eukaryotic MTases N6AMT1 and N6AMT2 (Fig. 1b), variably annotated either as N(6)-adenine MTases or as small N5-glutamine (HemK-like) and lysine (EEF1A) MTases, respectively, have been implicated in N6A methylation based on knockout/knockdown data ^{20,24}, but do not carry N- or C-terminal extensions, and modify proteins rather than DNA in functional assays ²⁵⁻²⁹, suggesting that *in vivo* perturbations may have indirect effects. The presumptive N6A-MTase METTL4_Av has a conserved N-terminal domain (KOG2356:

transcriptional activator, adenine-specific DNA methyltransferase) present in METTL4-like ORFs of most eukaryotes, including *A. vaga* (Fig. 1b, top); this permuted MTase, found in all bdelloids, may have persisted in eukaryotes throughout their evolutionary history (Fig. 1f, Table S1). In contrast, the bdelloid N6_N4_MTase has no eukaryotic homologs, and can be aligned only with permuted bacterial N4C- and N6A-MTases (Type II, subtype β), which cluster in accordance with their target recognition domains (TRD) recognizing specific targets compiled in REBASE^{1,23,30} (Fig. 1e; Supplementary Data File S1). Interestingly, the bdelloid lineage clusters with phage MTases of unknown target specificity, and its closest bacterial homologs are N4C-MTases recognizing TCGA and CCSGG. Thus, we tentatively assigned it to N4C-MTases and named it N4CMT, since it harbors the catalytic SPPY motif shared with most bacterial N4C-MTases and differing from bacterial N6A-MTases (DPPY), eukaryotic N6AMT1 (NPPY), N6AMT2 (DPPY/F) or METTL4-like enzymes (DPPW, also seen in METTL3/IME4-like m⁶A-RNA MTases)^{12,23,31} (Table S2).

Presence of 4mC and 6mA marks in genomic DNA. We next sought to find out whether recruitment of a horizontally transferred bacterial MTase resulted in establishment of bacterial epigenetic marks in bdelloid genomic DNA (gDNA). A strong indication that N4CMT could interact with chromatin to add 4mC to DNA comes from the presence of a eukaryotic chromodomain from the HP1/chromobox subfamily of methylated lysine-binding Royal family of structural folds³² at the C-terminus of the bacterial *N6_N4_MTase* moiety in sequenced bdelloids (Fig. 1a).

Fig. 1. Putative DNA methyltransferases and modified bases in bdelloid rotifers. **a-b**, Domain structure of putative N4C (a) and N6A (b) bdelloid amino-MTases. PFAM/KOG domains are indicated; conserved catalytic motifs and S-adenosylmethionine (AdoMet) binding sites are flagged; numbers correspond to aa positions in *A. vaga*. See Supplementary Data File S1 for gene IDs and aa sequences. **c**, Immuno-dot-blot analysis of membrane-immobilized gDNA from *A. vaga* Av-ref (746 ng), AvL1 (500 ng), *E. coli* C2925 *dam*-/ *dcm*- (550 ng) and *E. coli* M28 (2 μ g) probed with anti-4mC (top panel) and anti-6mA (bottom panel) antibodies. **d**, Summary of gDNA digestion (+) with restriction enzymes differing by methylation sensitivity: *Mbol* (blocked by *dam* methylation); *DpnI* (cleaves only *dam* methylated DNA); *Sau3AI* (not sensitive to *dam* or *dcm* methylation); *McrBC* (cleaves at any methylated cytosine). **e**, Neighbor-joining phylogram of permuted MTases of Type II, subtype β , displaying clustering by recognition sequences (obtained from REBASE). Clustering is not intended to uncover phylogenetic relationships in bacteria. Red arrow indicates acquisition of a chromodomain (CHD) by the bdelloid N4CMT. The source alignment is provided in Supplementary Data File S1. **f**, Phyletic distribution patterns of putative DNA methyltransferases implicated in 4mC, 6mA and 5mC addition. A consensus cladogram of metazoan phyla is shown on the left. **g**, *Adineta vaga* (isolate AvL1) under polychromatic polarization microscope. Photo credit: M. Shribak, I. Yushenova. Scale bar, 50 μ m.

To detect 4mC/6mA marks in bdelloid genomes, we extracted gDNA from the *A. vago* laboratory reference strain (hereafter Av-ref) ¹⁶ fed with methyl-free *E. coli* (Table S3) and performed immuno-dot-blotting with anti-4mC and anti-6mA antibodies (Methods). We also extracted gDNA from the natural *A. vago* isolate L1 (hereafter AvL1; Fig. 1g), which was caught in the wild and identified as *A. vago* through morphological criteria and mtDNA phylogeny, but represents a distinct morphospecies within the *A. vago* species complex, as its gDNA is only 88% identical to Av-ref ³³. Fig. 1c shows that gDNA from Av-ref and AvL1 reacts positively with

both antibodies, suggesting the presence of 4mC and 6mA marks. Control DNAs isolated from the *dam-/dcm-*, *DH5α* and *Top10 E. coli* strains, or from *E. coli* M28 strain used as food (Table S3), did not react with anti-4mC antibodies (Fig. 1c), and neither we observed any cross-reactivity of the anti-4mC antibody with 5mC-containing human DNA (data not shown). Also consistent with the presence of modified cytosines were the results of treatment of total *A. vago* gDNA with the McrBC endonuclease, which cleaves at methylated cytosines (5mC, 5hmC, 4mC)^{34,35} (Fig. 1d; see also Fig. 3b below). Together with the absence of C5-MTases, similarity of N4CMT to bacterial N4C-MTases (Fig. 1e) and the lack of 5mC deamination signatures in gDNA from observed/expected CpG ratios (Extended Data Fig. 1a), our data support the hypothesis that cytosines in bdelloids are modified at the N4- rather than C5-positions. Still, signals in gDNA may originate from residual methylated bacterial DNA from sources other than food. Thus, we sought to examine distribution of 4mC marks over annotated genomic features in *bona fide* eukaryotic contigs.

Genome-wide analysis of 4mC and 6mA by DIP-seq. We exploited immunoreactivity of bdelloid DNA with anti-4mC and anti-6mA antibodies to study genome-wide distribution of these methylation marks by DIP-seq (DNA immunoprecipitation followed by sequencing, also called MeDIP-seq; see Methods). After read mapping to Av-ref, the MACS peak-calling tool identified 1008 and 1735 DIP-seq peaks (p-value < 1e-5) for 4mC and 6mA, respectively, which were broadly distributed throughout the assembly. To uncover biologically relevant patterns behind peak distribution, we compared peak coverage densities for 4mC and 6mA near annotated genomic features, such as gene coding sequences (CDS) and transposable elements (TEs), using different window and bin sizes. We visualized distribution of 4mC and 6mA sites near TEs by aligning TEs at the 5' end (profiles) or aligning TE bodies from 5' to 3' end at a fixed distance (metaprofiles), and plotting the tag occupancy, which shows the relative number of tags (peaks, base modifications) against the total number of TEs for each bin size within a pre-determined upstream and downstream window. Fig. 2a shows representative results for 5-kb window size. About one-half of 4mC peaks (468 out of 1008) and a quarter of 6mA peaks (430 out of 1735) are close to TEs, and 4mC shows elevated peak coverage density near TE insertions in comparison with 6mA (Fig. 2a, left), suggesting that TE insertions could be an important 4mC modification target. For gene annotations, modification density is much lower and appears inversed in comparison with TEs, with more 6mA depositions (1261 out of 1735) than 4mC (398 out of 1008) (Fig. 2a, center).

Fig. 2. Genome-wide distribution of 4mC and 6mA methylation in *A. vago*. **a**, Distribution of DIP-seq 4mC and 6mA sites around genes, TEs (metaprofiles) and 5'-end TE profiles in Av-ref, showing peak coverage in 25-bp bins within ± 2.5 kb of each feature. In metaprofiles, the body size feature, representing genes or TEs, is automated and normalized (0-100% length). TE 5'-profile shows 4mC and 6mA sites near 5' boundaries, aligning transposons at the 5' end. **b**, IPD ratios in AvL1 SMRT-seq data at four representative 4mC and 6mA modification sites. Purple and orange bars, Watson and Crick strands. **c**, Numbers of SMRT-seq 4mC and 6mA modified bases in 5-bp and 25-bp bin sizes within ± 0.5 kb and ± 2.5 kb of 5' TE boundaries. **d**, MEME-ChIP motif analysis of regions around SMRT-seq 4mC and 6mA sites. Windows of ± 5 , ± 10 and ± 20 bp were extracted and searched for significant motif enrichment. The p-value generated by MEME-ChIP is shown under each motif. **e**, Methylation fraction distribution at modified sites detected by SMRT-seq. Most 4mC sites are fully methylated (fraction=1); average methylation level of 6mA sites is 0.74. **f**, PacBio read coverage distribution by base modification sites. Minimal threshold coverage limit applied for calling 4mC and 6mA methylated sites to calculate methylation fraction per site in (e) is shown by a dashed line. **g**, Average numbers of 4mC and 6mA base modifications in protein-coding genes, TEs and tandem repeats. Average is calculated as the total number of modified sites divided by total number of annotations (unique IDs) in each feature and divided (normalized) by the genome fraction covered by such annotation in the genome (genes, 0.533; TE, 0.021; TR, 0.0084). **h**, Distribution of SMRT-seq 4mC and 6mA sites within genic features (CDS, intron, 5' UTR, 3' UTR, 5'-promoter region) and intergenic regions by average feature size (bp). **i**, DNA methylation density (mean number of SMRT counts) vs TE copy integrity (full, medium and short, as indicated).

To find out whether peaks are distributed non-randomly, we examined the statistical significance of genomic correlations between peak distribution and annotated genomic features. Since functional interactions often depend on spatial proximity between the reference feature and the density of query features relative to it, we used spatial correlations as a proxy for functional analysis (Table S4). Genometric correlation analysis of annotated Av-ref scaffolds shows that 4mC peaks and TEs are non-uniformly distributed (p-value: $1.29\text{E-}13$, Kolmogorov-Smirnov test), and that the query features (4mC peaks) are closer than expected to the reference features (TEs) (Jaccard and permutation test). In contrast, we find that 6mA peaks and TEs are more uniformly (randomly) distributed (p-value: 0.036, Kolmogorov-Smirnov test), and that 6mA peaks tend to be further away from TEs (permutation test). When gene annotations are used as reference points, both 4mC and 6mA modifications are uniformly distributed, but for 6mA peaks the distance from genes is consistently small, while for 4mC peaks the distance from genes tends to be larger (Jaccard and permutation test).

The presence and distribution of 4mC and 6mA DNA modifications in AvL1 strain was similarly interrogated by DIP-seq. We generated DIP-seq reads and mapped them to the AvL1 assembly (Methods). After peak calling with MACS, we identified 1473 and 1385 peaks (p-value $< 1\text{e-}05$) for 4mC and 6mA, respectively. To further understand methylation patterns in AvL1, we performed initial gene and TE annotations with fully automated training methods for gene prediction, using genomic and RNA-Seq data (Braker2; see Methods) (Table S5). AvL1 repeat library was constructed *de novo*, curated, and used to annotate TEs (Methods). Initial analysis showed that 4mC-DIP-seq and 6mA-DIP-seq peaks were more frequently deposited close to TEs than to genes, with peak coverage increasing over TEs. Extended Data Fig. 2a,b show the

peak analysis for a 5-kb window size, in which 1097 4mC peaks (out of 1473) and 1042 6mA peaks (out of 1385) are close to TEs, while 863 4mC and 813 6mA peaks are close to genes (excluding TEs). Genometric correlation analysis on AvL1 showed that both modification peaks, 4mC and 6mA, have a small absolute positive correlation (Table S4) and are closer than expected to TEs as reference features than to gene models (Jaccard and permutation test). In sum, DIP-seq data in both Av-ref and Av-L1 suggest preferential localization of 4mC over TEs.

Modification analysis at single-base resolution by SMRT-seq. While immuno-dot-blot and differential gDNA digestion suggested the presence of 4mC in bdelloid gDNA, it was not possible to fully eliminate gDNA from commensal bacteria, even using methyl-free *E. coli* food strains and applying starvation/antibiotic treatments prior to DNA extraction (Methods). Hence, we chose not to use mass-spectrometry (MS) as a method to confirm the presence of 4mC in bdelloids, especially considering reports that unknown MS-peaks can comigrate with 4mC¹³. Further, low resolution of the DIP-seq method limits the power of correlation analyses to the length of DNA fragments used for antibody binding (250-450 bp), not to mention residual IgG binding to non-modified fragments inherent to the method³⁶. Thus, we chose to examine genome-wide distribution of modified bases by single-molecule real-time (SMRT) sequencing, which provides single-nucleotide resolution and allows validation of rotifer contigs (Methods).

SMRT-based detection exploits the kinetic signatures of polymerase passage through modified vs non-modified bases and is quantified in terms of inter-pulse duration (IPD) ratios. It is best suited for detection of 4mC and 6mA, characterized by strong kinetic signatures, which require ~10-fold lower coverage than 5mC detection (Pacific Biosciences Methylome Analysis Technical Note) and is widely used in bacterial methylome analyses^{30,37}. We obtained PacBio reads (15 SMRT cells, totaling 9.87 Gb) from gDNA extracted from AvL1 eggs and analyzed the kinetic profiles with SMRT® Portal (Methods). Prior to quantification of modified bases, we bioinformatically removed residual bacterial contigs (Methods), which, as expected, show high methylation density.

SMRT-analysis detected 4mC modifications on 21,016 cytosines (0.0643% of the total cytosines in the assembly) and 6mA modifications on 17,886 adenines (0.0236% of total adenines) using a minimum cutoff PacBio coverage defined in Fig. 2f (see Table S6 for comparison of 10x and 20x coverage levels). As with DIP-seq, SMRT-seq shows broad distribution of both modifications across the AvL1 assembly. Comparison of DIP-seq and SMRT-seq modification patterns shows a considerable overlap, with 36% of 4mC peaks and 32% of 6mA peaks overlapping with 4mC and 6mA identified by SMRT analysis, respectively, showing

that many peaks are conserved between eggs and adults. This overlap is quite substantial, given the modest percentage of modified bases, and is comparable to the overlap reported for 6mA and 5mC using the same orthogonal methods in mouse ESCs²², even with developmental differences between eggs and adults not seen in protostomes, at least for 5mC³⁸.

In contrast to the predominantly symmetric patterns of 5mC deposition at CpG doublets in eukaryotes, AvL1 shows mostly asymmetric patterns of methylation for both 4mC and 6mA, i.e. only one strand is usually modified (Fig. 2b shows typical examples). At 4mC sites, CpG and CpA dinucleotides are the most prevalent, making up 74% of modified doublets. For better identification of sequence preferences, we extracted different sequence windows (5, 10 and 20 bp) upstream and downstream from 4mC sites and searched for significant motif enrichment with MEME-ChIP (Methods) (Fig. 2d). For 4mC, three motifs with CG or CA dinucleotides were most significantly enriched (from $p=2.8e-593$ to $p=1.4e-513$). For 6mA, a similar approach yielded three significantly enriched short motifs (from $p=7.3e-656$ to $p=4.3e-420$) and increasing the motif length yielded GA embedded in an A-rich region ($p=2.4e-1243$). The dinucleotide GA is the most prevalent at 6mA sites, and the most common triplets AGG or GAA, when combined, compose 34% of all 6mA triplets. These findings parallel 6mA motif preferences in most metazoans but differ from unicellular eukaryotes and early-diverging fungi, in which the symmetric 6mA methylation targets ApT dinucleotides (Table S1).

In addition to measuring coverage at each 4mC and 6mA site, the SMRT-analysis pipeline reports different methylation levels (fraction), referring to the proportion of times a given nucleotide is identified as methylated (1 equals 100% methylation). Notably, most of the 4mC methylation corresponds to high-fraction sites (0.5-1), dominating over low-fraction sites (0.1-0.5) at a ratio 71:1, with 58% of 4mC sites being fully methylated (Fig. 2e). Methylation at 6mA sites appears more dynamic, although the highly methylated (0.8-1) and moderately methylated (0.5-0.8) sites still dominate over low-fraction sites (0.1-0.5), which constitute only 12% of 6mA sites.

We plotted the density of 4mC and 6mA in AvL1 (DIP-seq and SMRT-seq) across annotated features (genes, TEs, tandem repeats) (Fig. 2c,g; Extended Data Fig. 2a-d). The 4mC and 6mA tag densities reach similar levels for each annotation type (Fig. 2g). The 4mC density appears higher near TE 5'-ends (Fig. 2c), as was also seen in Av-ref DIP-seq showing increased deposition of 4mC peaks close to TE 5' ends (Fig. 2a, right). Nevertheless, a considerable number of 6mA sites (DIP-seq and SMRT-seq) is found near TEs (Fig. 2g,i; Extended Data Fig. 2b,d).

Methylation density in tandem repeats (TR) deserves a special mention. Fig. 2g shows that the average counts of 4mC and 6mA sites in TRs are elevated in comparison with TEs and genes. According to TRF annotation, only a small fraction (0.84%) of the AvL1 assembly is composed of TRs. Inspection of SMRT-seq modification data identified two repeats with very high density of methylated sites, located mainly on contigs 1882 and 785 adjacent to large *Athena* retroelements³⁹. Such extra-high modification density, approaching that in bacterial contigs, mostly accounts for over-representation of modified bases in TRs, leaving other TRs virtually unmethylated. In subsequent experiments, we took advantage of the high methylation susceptibility of these repeats (see below).

In genes, the PacBio methylation tag density is much lower than that in TEs and TRs (Fig. 2g). Still, genic regions cover slightly over one-half of the AvL1 genome, attracting a sizeable fraction of 4mC and 6mA modifications (52% of 4mC and 54% of 6mA). To correlate methyl marks with gene structure, we examined 4mC and 6mA distribution using more refined features: gene bodies, promoters within 2 kb upstream of the transcription start site (TS), and intergenic regions which may include TEs and TRs, with gene bodies further subdivided into CDS (exons excluding 5' and 3' UTRs), introns, 5' and 3' UTRs (Fig. 2h). Altogether, base modifications are found in all features (CDS, promoters and intergenic regions), however when the density per average feature size is compared, CDS regions appear denser than introns (Fig. 2h), which is reminiscent of 5mC patterns in mammals⁴⁰.

In AvL1, DIP-seq peaks show enrichment with 4mC and 6mA within TE bodies (Extended Data Fig. 2b). The PacBio 4mC sites display a trend for enrichment near the 5' TE boundaries, while 6mA sites show a local depletion (Fig. 2c), which is visible even though TE promoters are located near TE 5'-ends but not necessarily at the boundary, and is not due to a local change in base or dinucleotide composition (Extended Data Fig. 1b). Moreover, 4mC and 6mA marks are primarily found over full-length or nearly full-length TE copies and are practically absent from shorter TE fragments spanning less than one-half of TE consensi, suggesting that active TE copies are preferentially targeted (Fig. 2,i; see legend). The lack of 4mC and 6mA marks in shorter TE copies together with concentration of 4mC near 5' TE boundaries suggest that their deposition is associated with transcriptional activity.

To visualize 4mC and 6mA densities in TRs, TEs and genes on representative contigs, we built the corresponding Circos plots (Extended Data Fig. 3a-d), in which the PacBio modification layer is plotted as modification fraction (from 0 to 1) for each modified base. In agreement with Fig. 2e, highly methylated 4mC sites dominate in most locations, while 6mA sites are distributed over a much wider methylated fraction range and across a wider feature range. Importantly,

higher densities of modified bases are not correlated with areas of higher PacBio read coverage, indicating that over-representation of methyl marks over TEs and TRs is not due to excess coverage in these regions (e.g. mtDNA at 800x coverage displays very few such marks) (Extended Data Fig. 3e). Extended Data Fig. 3c,d shows that long copies of *Vesta* and *Athena* retrotransposons attract both methyl marks, but short copies do not.

N4CMT acts as 4mC-methyltransferase in E. coli. While the presence of METTL4-like (or N6AMT-like) MTases in bdelloids likely ensures deposition of 6mA marks onto gDNA, the existence of N4CMT *per se* cannot be taken as evidence of its N4C-MTase activity, since the *N6_N4_MTase* domain repeatedly evolved 6mA or 4mC specificities⁴¹. However, it is not possible to disrupt N4CMT function *in vivo*, as the tools for genetic manipulation in bdelloids are yet to be developed. We therefore sought to investigate the activity of the recombinant N4CMT protein in a heterologous system. To this end, N4CMT was PCR-amplified from *A. vaga* cDNA to obtain intronless versions (Methods; Table S7). Amplicons were cloned into pET29b expression vector with the N-terminal S-tag and the C-terminal 6xHis-tag and expressed in *E. coli*. We examined two *A. vaga* allozymes A and B, which differ by six amino acids (aa): three in the *N6_N4_MTase* domain and three in the chromodomain-containing C-terminus (Table S8; Extended Data Fig. 4a). We also tested two inter-allelic recombinants swapping the rightmost substitution near the C-terminal His-tag, which may have arisen during rotifer cultivation or PCR amplification, and two 3'-truncated derivatives with the C-terminal chromodomain removed.

To assess N4CMT activity *in vivo*, its expression was induced by adding IPTG to the Rosetta 2(DE3) *E. coli* transformed with plasmid-borne N4CMT, and gDNA was extracted 4h post-induction (Methods). Fig. 3a shows the immuno-dot-blot of membrane-immobilized gDNAs probed with anti-4mC and anti-6mA antibodies, with 4mC signal observed from full-length N4CMT allozymes in the absence of signal from the untransformed host strain. As expected in the *dam+* background, 6mA methylation was detected in all samples, thereby serving as an internal DNA control. Not surprisingly, removal of the chromodomain, which leaves a core MTase equal in length to its bacterial counterparts, did not reduce its activity, and even led to elevated signal intensity due to better solubility of the 33-kDa vs 45-kDa enzyme (Fig. 3a, N4CMT-ΔCbx). The N4CMT_A allozyme mostly showed weaker activity, suggesting that substitutions in the presumed TRD region of the *N6_N4_MTase* domain affect protein solubility or interaction with target DNA. These findings were corroborated by digestion of corresponding gDNAs with the McrBC endonuclease, which cleaves DNA at modified cytosines. Indeed, DNAs extracted from Rosetta 2(DE3) transformed with six N4CMT-expressing plasmids, as well as the

control human DNA, were readily digested with McrBC, while DNA from the untransformed *dcm*-strain was not (Fig. 3b).

Finally, to ensure that the observed activity is directly attributable to N4CMT, we created N4CMT mutants in which the catalytic SPPY motif was replaced with APPA (Table S8). Fig. 3c shows that 4mC addition is abolished after substitution of the catalytic Ser and Tyr residues with Ala, indicating that N4CMT is responsible for adding N4-methyl groups to cytosines in dsDNA, with SPPY as the catalytic motif, thereby justifying our initial N4CMT designation.

***In vitro* activity and substrate specificity of N4CMT.** Next, we sought to find out whether recombinant N4CMT displays *in vitro* activity, as do other bacterial MTases. To this end, N4CMT was expressed in *E. coli*, partially purified by immobilized metal-affinity chromatography (Extended Data Fig. 4b,c), and used to methylate *E. coli* gDNA for 4 h at 25°C in 1x *M.Bam*HI buffer (NEB) supplemented with 80 µM S-adenosylmethionine as a donor of methyl groups (Methods). Further, to check if pre-existing N6A- and C5-methyl groups could modulate the efficiency of N4C methylation, we used as substrates gDNA from five *E. coli* strains differing by genetic backgrounds with regard to methylation: Rosetta 2(DE3) and BL21(DE3) (both *dam*+ *dcm*-), derived from *E. coli* B, and three *E. coli* K12 derivatives (methyl-positive M28 (*dam*+ *dcm*+) and methyl-negative ER2738 (*dam*- *dcm*- *EcoK1*-) and ER2925 (*dam*- *dcm*-)) (Table S3). After incubation, samples and control DNAs were spotted on two identical membranes and probed with anti-4mC and anti-6mA antibodies, respectively; the latter served as an internal control and agreed with expectations from the genetic background of each strain (Fig. 3d; Extended Data Fig. 4e). Interestingly, *dam*+ *dcm*- *E. coli* B derivatives displayed stronger signal than *dam*- *dcm*- and *dam*+ *dcm*+ strains, suggesting that pre-existing 6mA marks might facilitate 4mC addition, and that the presence of 5mC in the carbon ring of the cytosine may interfere with 4mC addition at the neighboring amino group. Activity in *E. coli* strains is summarized in Table S9.

We also checked N4CMT for *in vitro* activity on dsDNA substrates (Fig. 3h). The positive control (*M.Bam*HI-methylated pUC19) was readily detected with anti-N4mC antibodies. However, unmethylated pUC19 and pBluescript SK+, grown in the *dam*- *dcm*- background, acquired only barely detectable 4mC marks upon N4CMT treatment. A more favorable *in vitro* substrate was N4CMT itself, as inspection of AvL1 PacBio data revealed 4mC marks over its ORF, indicating that it serves as its own substrate *in vivo*. Indeed, PCR-amplified N4CMT_A and B fragments yielded 4mC signal *in vitro*, hinting at the possibility of self-regulation *in vivo*.

Fig. 3. N4CMT activity and substrate preferences *in vivo* and *in vitro*. **a**, Immuno-dot blot of total DNA extracted from *E. coli* Rosetta 2(DE3) strain transformed with recombinant N4CMT variants. DNA from non-transformed Rosetta 2(DE3) was used as a control. DNA was extracted after 4 h of IPTG-induced N4CMT expression; 400 ng of each DNA was spotted in one dot. Methyl marks are indicated on the right. **b**, Methylcytosine-sensitive digestion of *E. coli* DNA. Total DNA from Rosetta 2(DE3) strain, either transformed with recombinant N4CMTs (as shown on the top) or untransformed, was extracted 4 h post-induction; DNA (6.3 μg) was treated with McrBC. DNA from HepG2 liver cell line (*H. sapiens*) served as a positive control (500 ng). **c**, Immuno-dot blot of total *E. coli* DNA showing the role of the SPY motif in N4CMT methylation. Same designations as in (A). **d**, Immuno-dot blot of total *E. coli* DNA extracted from different strains and treated with N4CMT. Rosetta 2(DE3) and *dam-/dcm-*, 950 ng per dot; BL21-AI and M28, 500 ng per dot. **e**, Immuno-dot blot with anti-4mC antibody for sAvL1-451 substrate treated with N4CMT allozymes and their catalytic site mutants *in vitro*. **f**, Nucleotide sequence conservation in *A. vanga* tandem repeats with high density of 4mC modifications. Interspecific conservation was determined from two *A. vanga* isolates plus the sibling species *A. ricciae*. Two bottom sequences show inserts m97 and m119 which confer N4CMT substrate properties to the pUC vector; conserved motifs are highlighted in red. **g**, Diagram of the *A. vanga* 4mC-rich 460-bp tandem repeat region and its substrate derivatives. Red arrows, modified cytosines. Conserved motifs are in purple. The dsDNA fragments used in activity assays are depicted on the bottom, with pluses and minuses summarizing N4CMT activity on the corresponding substrates. **h-i**, Immuno-dot blots with anti-4mC antibody for different substrates treated with recombinant

N4CMT allozymes. N4CNT_A methylation is weak but visible at higher exposures. Positive control, 100 ng *M. Bam*HI-treated pUC19. In (h), 500 ng of each DNA was loaded per dot, except for sAvL1-30 (6 µg). In (i), 1 µg of pUC19 plasmids (grown in *dam-/dcm-* *E. coli* strain) was loaded per dot. Linear dsDNA fragments (sAvL1-200, sAvL1-209, sAvL1-451) were equalized to 600 ng.

The inability of empty plasmids to serve as efficient substrates may be explained by the lack of cytosines in a rotifer sequence context favored by N4CMT. We performed *in vitro* assays on the ~460-bp tandem repeat from AvL1 DNA with high density of DNA modifications (see above; Fig. 3f,g), reasoning that it would serve as an efficient substrate *in vitro*. The PCR primers, spanning 451 bp of the repeat, amplified 1, 2 and 3 repeat units, which were separately used as substrates. We also annealed two complementary oligonucleotides to form a 30-bp dsDNA fragment containing the cytosines modified with 100% efficiency in SMRT-seq data (Fig. 3g; Table S7). The 451-bp fragment indeed served as an efficient substrate *in vitro* (Fig. 3h,i), although increasing the number of repeat units did not improve methylation efficiency (Extended Data Fig. 4d). However, the short 30-bp fragment failed to yield detectable signal, even when used in large amounts (Fig. 3h). Another short G-rich substrate, made of annealed GT-rich repeat-containing complementary oligonucleotides (Table S7), also failed to acquire 4mC (not shown). In agreement with *in vivo* results, the SPPY→APPA catalytic mutants were unable to add 4mC to the 451-bp fragment, reconfirming the identity of catalytic residues *in vitro* and ruling out any co-purifying MTases (Fig. 3e). To test for 6mA addition, we used PCR-generated 405- and 589-bp fragments marked by 6mA in AvL1 SMRT-seq data (Table S7). No 6mA was acquired upon incubation with N4CMT, suggesting that it lacks 6mA MTase activity (not shown).

We further dissected the 451-bp fragment into sub-fragments of 127 bp and 357 bp, to check whether they would serve as substrates (Fig. 3g). While the 357-bp fragment was readily modified by N4CMT, the 127-bp fragment was not, perhaps due to minimal length requirements (Extended Data Fig. 4d,e). Alternatively, it may be that N4CMT, which underwent horizontal transfer from a prokaryote relatively recently on an evolutionary time scale in comparison with METTL4- or DNMT-like MTases, retains target specificity in its TRD and prefers one fragment over the other. In support of this idea, we identified a partially homologous 750-bp AvL1 tandem repeat, similarly associated with an *Athena* retroelement. Aligning it with a 481-bp tandem repeat from another AvL1 contig with high density of modifications, we defined a bipartite motif common to these repeat units (Fig. 3f,g). We further searched Av-ref and the sibling species *A. ricciae* for sequences homologous to the AvL1 460-bp repeat. In Av-ref, we identified 5 contigs with 174-, 482- and 660-bp tandem repeat units partially homologous to AvL1 repeat, all adjacent to *Athena-W1*; in *A. ricciae*, two contigs carried 6 and 11 units of a 490-bp tandem repeat 74% identical to AvL1, with the bipartite motif (Extended Data Fig. 4f).

Dividing the 451-bp fragment into two approximately equal parts (200 and 209 bp) yielded no detectable signals (Fig. 3,i), possibly due to insufficient fragment size. Since the positive sAvL1-357 bp fragment contains only one part of the bipartite motif, we tested a similarly sized PCR fragment (sAvL1-371 on Fig. 3g) containing the other part of the motif. However, it was not methylated by N4CMT, suggesting that this part is not essential *in vitro* (Extended Data Fig. 4e), although it may play a role *in vivo*. Most importantly, insertion of short 97-bp or 119-bp fragments with the bipartite motif (Fig. 3f,g) into pUC19, initially unable to act as a substrate, converted it into an efficient *in vitro* substrate (Fig. 3,i). Thus, an MTase of bacterial origin shows preference for certain recognition sequences, which might have served as targets in the distant evolutionary past.

Base modifications and histone modifications. In the context of eukaryotic chromosomal DNA environment in *A. vaga*, any intrinsic target preferences of N4CMT manifested *in vitro*, while apparently yielding higher 4mC densities in a subset of tandem repeats, should not necessarily be required for 4mC deposition in other genomic regions, which may instead be facilitated by the N4CMT C-terminal chromodomain of the chromobox type (CBX)³². CBX is expected to recognize methylated lysine residues K9 and K27, the best-studied heterochromatic marks embedded in the ARKS motif at the N-terminus of histone H3, which are typically associated with transcriptionally silent chromatin and frequently overlap, both in terms of antibody cross-reactivity and similar function in TE repression⁴²⁻⁴⁵. To associate DNA methylation marks with specific histone modifications, we performed chromatin immunoprecipitation followed by deep sequencing (ChIP-seq) on *A. vaga* chromatin with anti-H3K9me3 and anti-H3K27me3 antibodies (Methods). For contrasting comparison with active chromatin, we used the anti-H3K4me3 antibody, which recognizes the histone modification typically associated with active transcription start (TS) sites^{42,46}. After validating the antibodies by immuno-dot-blotting (Methods), we profiled the distribution of these three H3 modifications in Av-ref and AvL1 by ChIP-seq. We found that H3K9me3, a mark for constitutive heterochromatin, often co-localizes with H3K27me3 known to characterize facultative heterochromatin, but not with H3K4me3, which marks active genes (Table S10). As expected, host genes display significant H3K4me3 enrichment, which typically covers 1-2 kb around the TS and shows a characteristic bimodal peak in both strains (Fig. 4a,c top). In contrast, H3K9me3 and H3K27me3 enrichment is observed mostly over TEs and covers the entire TE body, often extending upstream and downstream from a TE insertion, which may be indicative of spreading (Fig. 4b,d top).

Fig. 4. Base modifications and histone modifications in *A. vanga* strains Av-ref (a, b, e) and AvL1 (c, d, f). **a, c**, Profiles and heatmaps for gene regions with transcription start (TS) and transcription termination (TT) sites. Profiles (top) show relative fold enrichment of H3K4me3, H3K9me3 and H3K27me3. Heatmaps (bottom) represent H3K4me3, H3K9me3 and H3K27me3 ChIP-seq reads, where each row corresponds to gene regions with ± 3 kb from TS and TT boundaries. **b, d**, Profiles and heatmaps for TE annotations delimited by 5' and 3' boundaries. Profiles (top) show relative fold enrichment of H3K4me3, H3K9me3 and H3K27me3. Heatmaps (bottom) represent H3K4me3, H3K9me3 and H3K27me3 ChIP-seq reads, where each row corresponds to ± 3 kb from 5' or 3' TE boundary. **e, f**, Intersection of 4mC and 6mA DIP-seq peaks with ChIP-seq peaks for histone modification marks. Profiles of 4mC and 6mA peaks intersecting with H3K4me3, H3K9me3 and H3K27me3 modification tags are shown for Av-ref (e) and AvL1 (f). The signal is shown over a scaled window ± 3 kb from the peak; Y-axis, relative fold enrichment. Asterisk in (e) denotes artefactual signal from 9 contigs with anomalous coverage. **g**, DNA base methylation counts near histone marks. Y-axis, mean number of counts (SMRT-seq 4mC and 6mA) detected at H3K4me3, H3K9me3 and H3K27me3 ChIP-seq peaks. Counts are taken around each peak in a ± 500 bp window. **h**, Circos plot illustrating DIP/ChIP peaks, methylation sites,

sequencing read coverage and gene/TE annotations in selected Av-ref and AvL1 contigs. Features are explained in the key; source details are in Methods. SMRT-seq DNA methylation marks are shown within the PacBio layer in AvL1 for 4mC (blue triangle) and 6mA (red square). Mark height in the ring shows methylation fraction (0 to 1). Green line, RNA-seq coverage; purple line, small RNA coverage in Av-ref.

To explore association of 4mC and 6mA with active or repressive histone marks, we used ChIP-seq data for the euchromatic mark (H3K4me3) and two heterochromatic marks (H3K9me3 and H3K27me3) as a proxy for active and silent chromatin, respectively. The relatively low resolution of DIP-seq precludes genome-wide extrapolations in Av-ref, allowing only initial comparisons. For 6mA-DIP-seq peaks, 13.6% intersected with regions bearing euchromatic histone modifications (H3K4me3), while only 4.4% overlapped with heterochromatic histone modifications (H3K9me3 and H3K27me3 combined). For 4mC-DIP-seq peaks, 6.5% intersected with regions bearing heterochromatic histone modifications (H3K9me3 and H3K27me3), but only a minor fraction (1.5%) overlapped with H3K4-marked regions. Following normalization and aggregation of aligned reads in ChIP-seq datasets, the analysis reveals that DIP-seq peaks (4mC and 6mA) show pronounced superposition with the boundaries of H3K9me3 and H3K27me3 covered regions, however little if any overlap is seen with H3K4me3 (Fig. 4e).

In AvL1, for 4mC-DIP-seq peaks, 42.3% intersected with regions bearing heterochromatic histone modifications (H3K9me3 and H3K27me3 combined), but only 6.6% overlapped with H3K4me3-marked regions. Similarly, for 6mA-DIP-seq peaks, 42.9% overlapped with heterochromatic histone modifications (H3K9me3 and H3K27me3 combined), but only 6.3% intersected with regions bearing euchromatic H3K4me3 modifications. After normalization of aligned reads in the ChIP-seq dataset, we confirmed that DIP-seq peaks (4mC and 6mA) are strongly correlated with H3K9me3 and H3K27me3 heterochromatic peaks (Fig. 4f). Although initial analysis yielded some correlation of H3K4me3 reads with DIP-seq peaks, after cluster analysis (deepTools option `-kmeans` with `-outFileSortedRegions`) we found that the signal originates from only 9 contigs, is not correlated with H3K4me3 peaks called by MACS2, and likely stems from atypically high coverage in H3K4me3 reads. As seen in plotted examples for three of these contigs, no H3K4 methylation marks are visible (As1882 and As785, Extended Data Fig. 3a), at least in the vicinity of DIP-seq marked regions (As1218, Fig. 4h). Thus, the presence of DNA methyl marks is preferentially associated with silent chromatin in both strains. A similar pattern is observed in AvL1 PacBio SMRT analysis, where the 4mC and 6mA marks are more frequently associated with inactive chromatin domains marked by H3K9me3 and especially H3K27me3 (Fig. 4g). On balance, these results support the view that, in addition to any intrinsic target specificity of N4CMT, its action in the genome may be directed by the CBX moiety, targeting MTase activity to chromatin regions with repressive histone marks.

Methylomes, transcriptomes and small RNAs in the chromatin context. To associate histone marks with transcriptionally active or repressed genes in *A. vaga*, we plotted our RNA-seq data for genes co-localizing with either active or repressive H3Kme3 histone marks (Methods). As expected, genes near H3K4me3 have significantly higher RPKM (reads per kilobase of transcript per million mapped reads) values (ANOVA p-val <0.01) than genes with heterochromatic histone marks (H3K9me3, H3K27me3) or no marks (Fig. 5a). AvL1 displays the same pattern (Extended Data Fig. 6a). Further, tentative designation of 6mA modification as an active epigenetic mark^{9,10} prompted us to explore its correlation with gene transcription. The *A. vaga* gene dataset, after removing TE-derived genes, was divided into two groups, with and without presence of 6mA peaks within a window size of ± 500 bp of each gene ID, and RPKM values were counted in both groups. We found that genes with 6mA depositions tend to have higher RPKM than genes without 6mA (t-test p-val: 2.2E-16, Fig. 5b bottom). For 4mC modifications, no significant differences in gene expression were seen between genes with or without 4mC marks (Fig. 5b top). A detailed analysis of 6mA distribution in genes and their promoters, which shows that only a subset of genes is affected, and rules out contribution of m⁶A in RNA, is given in the Supplementary Note.

A different picture was observed for TEs. The exceptionally low TE content and diversified small RNA (sRNA) silencing machinery in bdelloids, averaging 20 Piwi/Ago and 30 RdRP variants, implies tight controls on TE proliferation via efficient silencing^{16,47}. In *A. vaga*, virtually every active TE family displays coverage by pi-like RNAs, which is correlated with low transcriptional activity⁴⁸. We examined association of transcript levels of TE-related genes with DIP-seq peaks (Methods). While TE-related genes with or without 6mA did not show much difference in RPKM values, TE-related genes with 4mC marks showed a significant decrease when compared to those without 4mC (t-test p-val: 6.8E-8, Fig. 5c, top). Thus, in expressed TEs 4mC may be regarded as a repressive mark. Note that co-localization of 4mC and 6mA is compatible with repression, as 6mA is involved in an adversarial network preserving Polycomb silencing²².

A significant overlap between TEs and heterochromatin, defined by H3K9me3 and H3K27me3 depositions (Fig. 4b,d), is paralleled by an overlap between TEs and sRNAs aligned to Av-ref (Fig. 5d; Methods). We analyzed sRNA association with histone marks and with 4mC/6mA DIP-seq peaks. Relative fold excess of sRNA is evident at heterochromatic H3K9me3 and H3K27me3 peaks and extends into nearby regions; in contrast, sRNA enrichment is low within H3K4me3 peaks, which mark active genes (Fig. 5e). Comparison of sRNA vs DIP-seq peaks for 4mC/6mA shows enrichment within DIP-seq peaks, with 4mC peaks having higher

relative fold enrichment than 6mA (Fig. 5f). To estimate the proportion of DIP-seq peaks contributing to each sRNA profile, we clusterized the peaks with the k-means algorithm, sorting by sRNA coverage (deepTools option --kmeans --outFileSortedRegions), which showed that ~25% and ~15% of 4mC and 6mA peaks, respectively, display small RNA enrichment. The overlap between sRNA and DIP-seq peaks is localized to the peak area, while sRNA enrichment at heterochromatic ChIP-seq peaks extends further into adjacent sequences, which may indicate spreading. Although the exact pathways linking piRNAs to histone and DNA methylation layers remain to be defined, the highly diversified PIWI proteins may serve as connectors to both layers.

Fig. 5. Association of transcript and small RNA levels with histone and DNA methylation.
a, Box plot showing Av-ref gene expression levels (log₂RPKM) associated with co-localized H3K4me₃, H3K9me₃, H3K27me₃, and both H3K9-27me₃ marks or without histone marks. ANOVA analysis shows significant differences in expression, with genes associated with the H3K4me₃ mark displaying the highest RPKM (reads per kilobase per million mapped reads). **b-c**, RPKM values associated with DIP-seq 4mC and 6mA base modifications in *A. vaga*. TE genes (C) are derived from Av-ref automated gene models after positively intersecting with TE annotations. Box represents the first and third quartiles; line, the median. The p-values were calculated by a two-tailed Student's t-test, with asterisks indicating significant differences. **d-f**, Distribution of sRNA with respect to genes and TEs (d), H3K4, H3K9, H3K27 ChIP-seq peaks (e), and IP4mC, IP6mA DIP-seq peaks (f). Relative fold enrichment is shown as reads per genomic context (RPGC normalization). In gene and TE profiles, regions in the map comprise gene bodies (5 for TS and 3 for TT) or TE bodies (5 for the 5'-boundary and 3 for the 3'-boundary) with ± 3 kb flanks. Peak profiles are represented by peak body flanked by ± 3 kb.

Interpreting the 4mC marks. To identify possible readers of bacterial marks, we searched for candidate proteins capable of discriminating between methylated and unmethylated cytosines. All known DNA methyl groups protrude from the major groove of the B-form double helix and can be recognized as epigenetic marks. In eukaryotes, several protein domains can read 5mC (SRA/SAD/YDG; MBD/TAM; Kaiso) or 6mA (HARE-HTH; RAMA) modifications^{6,12}, usually in a favored sequence context. We used profile HMM searches to detect candidate methyl readers in *Adineta* genomes. No homologs were found for the SAD_SRA domain (PF02182), which recognizes hemi-methylated CpGs by embracing DNA and flipping out the methylated cytosine⁴⁹. However, we saw drastic expansion of MBD/TAM-containing proteins, which do not require base-flipping: a total of 14 different alleles (originating from three quartets, Q1-Q3, plus a segmental duplication) encode 7 major SETDB1 variants, as opposed to only one in monogonont rotifers or other invertebrates (Fig. 6a,b; Extended Data Fig. 8a; Supplementary Data File S2). These proteins share the same domain architecture, with the MBD sandwiched between the N-terminal triple-Tudor domains and the C-terminal pre-SET/ SET/post-SET domains, present in all SETDB1/eggless-like H3K9me₃ histone lysine MTases (KMTs) (Fig. 6a). All seven proteins are transcribed in each *Adineta* spp. (data not shown). Additional MBD/TAM domains of BAZ2A/TIP5-like remodelers, which form heterochromatin on rDNA and satellites⁵⁰, comprise only one *A. vaga* quartet (Extended Data Fig. 8c; Supplementary Data S3).

To find out whether other KMTs are similarly expanded, we performed an inventory of SET domain-containing proteins in *A. vaga*, especially those known to methylate H3K9/H3K27 (Supplementary Data File S3). In addition to seven pairs of SETDB1 homologs acting on H3K9, we detected two quartets of E(z)/EZH/mes-2-like orthologs (KOG1079, Transcriptional repressor Ezh1), which are known to methylate H3K27. More distantly related SET-domain proteins showed domain architectures characteristic of H3K4, H3K36 and H4K20 KMTs (Trx-G/Ash1/ Set1/MLL, SETD2, SETD8) and comprised either a quartet or a pair. Interestingly, we found six stand-alone SET-domain homology regions resembling H3K4/H3K36 KMTs

(PRDM9/7/set-17), which were not predicted in the annotated gene set, were not transcribed, and lacked additional domains (KRAB_A-box, SSXRD) characteristic of PRDM9/7 proteins involved in localizing meiotic recombination hotspots and in male-specific expression^{51,52}. Unexpectedly, we failed to identify two known KMT types acting on H3K9 or K9/K27: Su(var)3-9/SUV39H1/set-25/Clr4, a “histone read-write” architecture consisting of chromo- and SET-domains, which is important for constitutive heterochromatin formation⁵³; and G9a/EHMT2/KMT1C (ankyrin repeats plus SET), which initiates *de novo* methylation and silencing of repeats and developmentally regulated genes^{54,55}. These domain architectures may have been lost and/or replaced by expanded SETDB1-like variants.

We next sought to determine whether SETDB1 is similarly amplified in all bdelloids. Six species in the genus *Rotaria* from the family Philodinidae¹⁸ possess the same seven variants as do Adinetidae, indicating that SETDB1 amplification occurred prior to divergence of the major bdelloid families (Fig. 6b). An unusual SETDB1 divergence pattern is seen in the bdelloid *Didymodactylos carnosus*, which forms the deepest-branching sister clade to other known bdelloids⁴⁷ and lacks N4CMT. While in three cases Dcar_SETDB1 forms sister clades to variants from other bdelloids, preceding quartet formation, the Q1 quartet lacks Dcar_SETDB1 homologs, and an ortholog of Av_s314 shows a clear evidence of loss, detected as a small 170-aa C-terminal fragment (Supplementary Data File 3). This natural gene knockout is associated with an increase in LINE elements to the levels seen in monogononts, and agrees well with high concentration of 4mC observed over LINEs (Extended Data Fig. 7), but is not correlated with high copy number of Ago/Piwi proteins (Fig. 6c)⁴⁷.

The presence of SETDB1 orthologs in species lacking 5mC, such as *D. melanogaster* and *C. elegans*, questioned the role of MBD as a universal discriminator of 5mC marks in DNA⁶. Notably, the structure of human MBD1 shows its unique potential for recognizing 5mC in the major groove without encircling DNA, which makes MBD an ideal candidate for interacting with nucleosome-bound DNA without interference from core histones^{56,57}. Moreover, three of the seven SETDB1-like variants in bdelloids display two conserved arginine residues in the MBD involved in recognition of cytosines in the DNA backbone, potentially accounting for CpG preference (Extended Data Fig. 8a,b). However, they show extensive variation in the length of the antiparallel β 1- β 2 loop, which reaches across the major groove and interacts with one of the methyl groups. Since the overall structure is compatible with recognition of an asymmetrical DNA methyl group in the nucleosomal context, we sought to find out whether some of the seven SETDB1 variants may have adapted to recognizing a novel methyl mark in the major groove.

Fig. 6. Amplification of SETDB1 histone methyltransferases and preference for 4mC-methylated DNA. **a**, Domain architecture of bdelloid SETDB1 proteins. Square bracket marks cloned MBD domains; aa numbering is for Av_s314. **b**, Unrooted maximum likelihood phylogram of SETDB1 variants in bdelloids (blue), monogononts (green) and acanthocephalan (olive). Q1-Q3, quartets of homeologs formed by paleotetraploidy. Bottom clades include single copy SETDB1 in 3 protostome phyla. See Supplementary Data File 2 for aa sequences. Scale bar, aa substitutions per site. **c**, LINE retrotransposon content (% genome) and Piwi/Ago copy numbers in 6 monogonont (green) and 10 bdelloid (blue) species (Table S2). Standard deviation (% LINE) and copy counts are given for sequenced isolates (numbers in parentheses). **d**, Affinity of AvMBD for 4mC-methylated DNA in electrophoretic mobility shift assays. EMSA was performed using 0.05 nM ³²P-labeled sAvL1-451 DNA and 3.75 nM AvMBD_s314 protein. Unmethylated and 4mC-methylated by N4CMT_B sAvL1-451 fragments were used as competitor DNA. **e**, AvMBD_s314 DNA binding preference for methylated DNA. Y-axis, per cent unbound ³²P-labeled sAvL1-451 DNA for 3.75 nM AvMBD_s314 in the presence of unmethylated and 4mC-methylated sAvL1-451 competitor DNA (n=2, mean±SD). Asterisks, p<0.05. **f**, A simplified model of the self-reinforcing regulatory loop based on the ability of N4CMT and SETDB1 to cross-recognize methyl marks on histones (circle) and DNA (square), respectively. Shown are the relevant conserved

proteins/domains described in the text; shadows, multiple copies. Hypothetical pathways from piRNAs/ Piwi (dashed lines) are not defined. [X] and [?] are putative mediator complexes with poorly conserved components, of which only Nxt1 is identifiable in *A. vaga* (see Discussion); HP1 and KDM4 are well-conserved. Components involved in other types of histone/DNA modification are not shown for simplicity.

To this end, we synthesized seven recombinant plasmids carrying tagged versions of the corresponding MBD/TAM domains (Extended Data Fig. 9a). We tested these proteins in electrophoretic mobility shift assays (EMSA) with the 451-bp repeat fragment (sAvL1-451, see above), which was either unmethylated or 4mC-methylated by N4CMT *in vitro*, to ensure sufficient methylation density and favorable position of methyl marks. As MBD/TAM is a generic DNA-binding domain, most AvMBD's are capable of binding both unmethylated and methylated DNA fragments (Extended Data Fig. 9b,c). We chose AvMBD_s314 to test its binding preferences for 4mC-methylated DNA, since the loss of its ortholog in *D. carnosus* is associated with a notable increase in LINE retrotransposon content⁴⁷. We tested four AvMBD_s314 concentrations (2.38 nM, 3.23 nM, 3.75 nM, 4.14 nM) in EMSA with ³²P-labeled sAvL1-451 and four concentrations of the unlabeled sAvL1-451 competitor, which was either unmethylated or 4mC-methylated by N4CMT_B *in vitro*. This approach provides a more adequate comparison than measurement of dissociation constants (Kd) for two labeled probes, as *in vitro* methylation is variably efficient. We observed a clear preference of s314 for binding 4mC-methylated DNA, with $p < 0.05$ in a one-tailed Student's t-test in four independent experiments, when using $>10\times$ excess of non-labeled competing methylated or unmethylated DNA ($p = 0.044$ for $40\times$; $p = 0.018$ for $100\times$). Fig. 6d shows a representative EMSA gel for the 3.75 nM s314 protein concentration, which yielded 88.3% protein-bound DNA with 0.05 nM ³²P-labeled sAvL1-451 fragment. This protein concentration was tested twice, and the average change in the amount of unbound DNA over increasing concentrations of unlabeled competitor DNA is shown in Fig. 6e, demonstrating that upon increase of competitor concentration, the shift from DNA-protein complex to unbound DNA occurs faster for 4mC-modified DNA than for unmethylated DNA. While other SETDB1 version(s) may be similarly adapted to prefer 4mC, they could also form alternative connections with multiple variants of bdelloid Piwi/Ago proteins, which would be interesting to explore in further studies.

Discussion

Here we report the first case of 4mC occurrence in eukaryotic DNA, expanding the repertoire of methylated bases in Metazoa with a modification known so far only in bacteria. We confirm its presence in bdelloid rotifers by combining multiple lines of investigation, and accounting for artefacts inherent to each modification detection method^{36,58,59} and for bacterial contaminations.

In agreement with the absence of *Dnmt1/Dnmt3*-like MTases, we failed to detect 5mC in bdelloids, while 4mC and 6mA are readily detectable by orthogonal methods. We identified N4CMT, a horizontally transferred enzyme of bacterial origin, as responsible for addition of the characteristically bacterial 4mC marks to DNA. Expression of recombinant N4CMT in *E. coli* results in 4mC addition, as follows from immuno-dot-blot analysis and methyl-sensitive digests of DNA from N4CMT-expressing bacteria vs. methyl-free strains. Not surprisingly, the chromodomain moiety is not required for 4mC deposition either onto bacterial DNA *in vivo*, which is not packaged into chromatin, or onto preferred DNA substrates by recombinant N4CMT *in vitro*. However, in the context of eukaryotic chromatin, ChIP-seq and DIP-seq peak distributions reveal strong correlations between silent H3K9/27me3 chromatin marks and DNA methyl marks. Thus, N4CMT may contribute to epigenetic homeostasis, whereby deposition of repressive chromatin marks is ensured by passive preservation of 4mC marks *via* covalent linkage to DNA in the absence of active enzymatic demethylation, helping to maintain TE repression in eggs and whole animals. Over-representation of 4mC at the 5' TE boundaries near TE promoter regions may affect transcription factor binding near promoters and cause transcriptional interference, as previously seen for 5mC ⁶⁰.

While the lack of candidate 4mC erasers supports 4mC role in maintaining TE silencing, other important components of epigenetic systems are the reader proteins, which could interpret the 4mC mark to form a regulatory loop, as is the case for 5mC and 6mA ⁶¹. The N4CMT architecture is somewhat reminiscent of plant chromomethylases (CMT), “histone-read-DNA-write” enzymes with a C5-MTase-embedded chromodomain, which reads H3K9me marks and deposits similar marks at nearby non-CG’s. Together with another domain configuration, “DNA-read-histone-write” provided by KYP, an H3K9-KMT with the 5mC SRA reader domain, the CMT3-KYP pair forms a mutually reinforcing loop reading each other’s epigenetic marks ⁶¹. The crosstalk between mCpG and H3K9me in animals and plants is even more complex, requiring multiple protein factors ⁵. In N4CMT, a very simple “histone-read-DNA-write” architecture, with the chromodomain reading the repressive H3K9/27me3 marks and MTase writing the atypical 4mC marks onto DNA in the absence of an eraser, links histone and DNA layers through a reinforcement loop which feeds back onto silent chromatin via “DNA-read-histone-write” SETDB1-like KMTs to help maintain repressive marks on histone tails throughout cell divisions for continuous TE silencing (Fig. 6f). Association of 4mC with full-length TEs capable of transcription, as well as the overlapping 4mC and sRNA distribution patterns, further suggest that the loop is triggered by pi-like RNAs from transcribed TEs, which can initiate transcriptional silencing on nascent RNAs via Piwi and perhaps SFINX-like protein complexes ⁶²⁻⁶⁴, or may

directly affect methylation, as in mice ⁶⁵. In this scenario, epigenetic inheritance relies on overriding the normally occurring H3K9me erasure by KDM4/JMJD2 ⁶⁶, which is present in *A. vaga*. Our finding that an amplified *A. vaga* SETDB1-like variant, also present in other bdelloids, shows preference for binding 4mC-methylated DNA *in vitro* suggests that 4mC deposition stimulates more efficient binding of SETDB1 in the nucleosomal context, linking 4mC deposition by N4CMT to the re-establishment of H3K9 methylation that helps to preserve silent chromatin marks on TEs and other repeats.

Notably, bdelloids exhibit some of the lowest TE content among metazoans, while members of their sister class Monogononta, which lack cytosine methylation and encode a single SETDB1 copy, show reduced ability to contain TE proliferation, which can double their genome size ⁶⁷. Earlier, we found a drastic expansion of Ago/Piwi and RdRP proteins in bdelloids, which are very TE-poor, in contrast to the acanthocephalan *Pomphorhynchus laevis* (Rotifera) with 66% TE content and no expansion of Ago/Piwi ^{16,47}, underscoring the importance of RNA silencing pathways in TE control. Notably, the bdelloid *D. carnosus*, despite Ago/Piwi expansion, does not show the dearth of retrotransposons typical of other bdelloids, displaying an elevated content of LINE elements matching that of *Brachionus* and shifting the average bdelloid LINE content upwards ⁴⁷. Here, we find that *D. carnosus* lacks both N4CMT and the two SETDB1 variants which may have evolved to interact with the 4mC mark, suggesting that the genome defense system in *D. carnosus* is missing an important layer that prevents TE expansion. Elevated LINE content in this natural knock-out of the 4mC-preferring KMT variant highlights the importance of cross-talk between two genome defense layers for efficient TE control, as *Adineta* and *Rotaria* during their evolutionary history experienced a strong decrease in retrotransposon content (see Fig S3 c,h,i in ⁴⁷), which coincided with the emergence of N4CMT and the 4mC-binding SETDB1 variant.

Collectively, our findings help to unravel a fascinating evolutionary puzzle: How can a bacterial enzyme decorating DNA with non-metazoan modifications penetrate eukaryotic gene regulatory networks and become preserved by natural selection for tens of millions of years? Given the importance of similar processes at the dawn of eukaryotic evolution, when MTases were recruited to create the extant epigenetic systems, the bdelloid case spans a unique time interval in the evolutionary history, when its advantages have been fully manifested and validated by natural selection, but its resemblance to bacterial counterparts has not yet been completely erased. Losses of DNA methylation have occurred multiple times throughout the eukaryotic tree of life; however, *de novo* recruitment of a bacterial mark into an existing epigenetic system has not been observed in more recent metazoan history. A synthetic “DNA

read-write” 6mA system in cultured human cells, based on *E. coli* Dam MTase and bypassing chromatin states through artificial targeting, has been created⁶⁸, however such a “shortcut” would be unlikely to persist in living species over evolutionary time scales. Our system helps to discern selectively advantageous features in epigenetic control systems and emphasizes that addition of a DNA epigenetic layer to the histone layer demands enhanced inter-connection of components between layers for efficient operation. Finally, it demonstrates that horizontally transferred genes, contrary to the established view^{69,70}, can re-shape complex eukaryotic regulatory networks and can drive major evolutionary innovations in eukaryotes.

Additional discussion can be found in Supplementary Discussion.

Online content. Any methods, additional references, Nature Research reporting summaries, source data, statements of data and code availability, and associated accession numbers are available online.

Acknowledgments. We thank Dr. Iain Murray (NEB) for the kind gift of anti-4mC and anti-6mA antibodies. This work was supported by R01GM11917 from the U.S. National Institutes of Health to I.A.

Author contributions. FR was responsible for high-throughput genomic and transcriptomic data generation and analysis. IY performed protein expression, purification, and biochemical characterization. DD conducted pilot experiments at early stages of this work. IA conceived and designed the project, analyzed the data, and drafted the manuscript. All authors contributed to writing and editing the final version.

Ethics declarations. The authors declare no competing interests.

List of Supplementary Tables and Datasets

Table S1. Comparison of N6A methylation in *A. vaga* and other eukaryotes.

Table S2. Putative amino-MTase and demethylase orthologs in the phylum Rotifera.

Table S3. Properties of *E. coli* strains.

Table S4. Genometric correlations between DIP-seq methylation marks and gene and TE annotations in Av-ref and AvL1 assemblies.

Table S5. Genome assembly and gene annotation metrics.

Table S6. SMRT-seq base modification detection.

Table S7. Primers and oligonucleotides.

Table S8. N4CMT recombinant proteins.

Table S9. Summary of N4CMT action on *E. coli* genomic DNA.

Table S10. Summary of ChIP-seq peaks identified by MACS2 (diagonal values) and overlap of peaks within Av-ref and AvL1 assemblies.

Table S11. Gene ontology analysis of methylated and unmethylated genes (.XLSX file).

All methylated (AvL1 with 6mA SMRT-seq signature at TS and homologous genes in Av-ref) and unmethylated genes are categorized by molecular function based on GO_slim ontology annotation. Gene counts and percentages for each category are shown for Av-ref and AvL1 strains. Asterisks represent significant differences in the counts of categories between methylated and unmethylated genes (Fisher's exact test).

Table S12. Methylation analysis in under-annotated regions.

Genome regions showing significant methylation density for 4mC and/or 6mA base modifications but no *ab initio* annotations for gene models or transposons.

Supplementary Data File S1. Amino acid sequences of bdelloid N4CMT and Type II subtype β bacterial methyltransferases, with accession numbers from Genbank or REBASE (.fasta format).

Supplementary Data File S2. Amino acid sequences of SETDB1 proteins from the phylum Rotifera and three representative protostome phyla, with accession numbers from Genbank (.fasta format).

Supplementary Data File S3. MBD- and SET-domain containing proteins in *A. vaga* Av-ref, with scaffold numbers from Genbank and protein IDs from Genoscope (.fasta format).

Supplementary Data File S4. Contigs with residual bacterial sequence joined to AvL1 DNA (.gff format).

809 **Table S1.** Comparison of N6A methylation in *A. vaga* and other eukaryotes.
810

Phylum	Species	% 6mA/A	Symmetry	Motifs	Enzymes	Feature	Transcripts	Methods	References
Rotifera	<i>Adineta vaga</i>	0.024	no	AGG,GAA	METTL4?	genes, TE	active	SMRT,IP	This study
Ciliates	<i>Tetrahymena thermophila</i>	0.4-0.8	yes (part) ss, ds	AT GATC	TAMT1	Gene body, linker+1+2	Activate, weak corr.	SMRT	71 19
	<i>Oxytricha nova</i>	0.71-1.04		AT	MTA1	linker	mix up/dwn	MS,SMRT	72
Plants	<i>Chlamydomonas</i>	0.3-0.5		AT		TS	Activate	IP,RE,exo	73
	<i>Arabidopsis</i>	.006-0.138	no	AGA,ACC		Genes, TE	gen ↑, TE ↓	MS,SM,IP	74
	<i>Oryza sativa</i>	0.2	no	GAGG	Ddm1 AlkB	Low at TS, up at TT	prom silent body active	IP,MS, SMRT	75 76
Oomycetes	Phytophthora	0.04-0.05		AT	DAMT2 ^a	TS bi,TE	Lowly exp.	IP,MS	24
Fungi	Early-DF	0.2-2.8	yes	AT	PF02384 ^b	genes	active	SM,IP,MS	77
	Dikarya	0.048-0.21	no	AV				SMRT	77
Ctenophora	Mnemiopsis	0.01-0.025	n/a		METTL4			ELISA	78
Ecdysozoa	<i>Caenorhabditis elegans</i>	0.01-0.4	no	AGAA GAGG	DAMT1 AlkB			SMRT,MS, IP	21
	<i>Drosophila</i>	0.001-0.07	n/a		Tet/AlkB	TE	silence	MS,IP	79
	<i>Aedes aegypti</i>	0.00001	n/a		METTL4 Tet				80
	<i>Bombyx mori</i>		no	ACAA	METTL4	Low TS, TT	silence	IP-seq	81
Vertebrates	Zebrafish	.002-.1 emb	no	AG		TE	activate	MS,IP	82
	<i>Xenopus laevis</i>	0.00009	no	AG		Low at TS			83
	<i>Sus scrofa</i>	.05-.17 emb	n/a	-				MS	82
	Mouse ES	.0006-.007	no				silence		84
	Brain			AAGA					85
	ESC			AGGA	METTL4	noncoding		MS,IP	22
	Rat	0.00001	n/a					MS,IP	86
	Human	.023-.064	no	AGG	N6AMT1 ^c	exons, TT	activate		20
	Glioblastoma		H3K9me3			Low at X,Y			29
	2n/1n cells	0.004		AG,GA	AlkBH1	Allele-spec.		SMRT	87

811
812 ^a Related to N6AMT2; ^b Related to MT type IC; ^c Related to N6AMT1.
813 Abbreviations: MS, mass-spectrometry; IP, MeDIP-seq; SMRT, SMRT-seq; TS and TT, transcription start
814 and transcription termination; TE, transposable elements; emb, embryos; Early-DF, early-diverging fungi.

Table S2. Putative amino-MTase and demethylase orthologs in the phylum Rotifera.

Species	N4CMT (N6_N4_Mtase) PF01555 SPPY	METTL4 (MT-A70) PF05063 DPPW	N6AMT1 PF05175 NPPY	N6AMT2 (N6aMlase) PF10237 DPPF/Y	MT type IC PF02384 NPPF/Y	AlkBH1	AlkBH4	TET	WGS assembly (source)
<i>Adineta vaga</i> (Av-ref)	+	+	+	+	-	+	+	-	16
<i>Adineta vaga</i> (AvL1)	+	+	+	+	-	+	+	-	33
<i>Adineta ricciae</i>	+	+	+	+	-	+	+	-	18
<i>Adineta steineri</i>	+	+	+	+	-	+	+	-	47
<i>Rotaria magnacalcarata</i>	+	+	+	+	-	+	+	-	18
<i>Rotaria macrura</i>	+	+	+	+	-	+	+	-	18
<i>Rotaria sordida</i>	+	+	+	+	-	+	+	-	47
<i>Rotaria socialis</i>	+	+	+	+	-	+	+	-	47
<i>Rotaria</i> sp. 'Silwood-1'	+	+	+	+	-	+	+	-	47
<i>Rotaria</i> sp. 'Silwood-2'	+	+	+	+	-	+	+	-	47
<i>Didymodactylos carnosus</i>	-	+	+	+	-	+	+	-	47
<i>Brachionus plicatilis</i>	-	-	+	+	-	+	+	-	88
<i>Brachionus calyciflorus</i>	-	-	+	+	-	+	+	-	89
<i>Brachionus koreanus</i>	-	-	+	+	-	+	+	-	90
<i>Brachionus rotundiformis</i>	-	-	+	+	-	+	+	-	91
<i>Brachionus asplanchnoidis</i>	-	-	+	+	-	+	+	-	67
<i>Brachionus</i> sp. 'Tiscar'	-	-	+	+	-	+	+	-	67

Table S3. Properties of *E. coli* strains.

Strain name	Genotype	Methylation marks	Strain source
ER2925 (<i>dam</i> ⁻/ <i>dcm</i> ⁻)	<i>ara-14 leuB6 fhuA31 lacY1 tsx78 glnV44 galk2 galT22 mcrA dcm-6 hisG4 R(zgb210::Tn10)TetS endA1 rspL136 (StrR) dam13::Tn9 (CamR) rfbD1 xylA-5 mtl-1 thi-1 mcrB1 hsdR2</i>	None (except rare EcoKI methylation)	NEB
ER2738 (methyl-free)	<i>F' proA⁺ B⁺ lacI^f Δ(lacZ)M15 zzf::Tn10(Tet^R)/ fhuA2 glnV Δ(lac-proAB) thi-1 Δ(hsdS-mcrB)5</i>	None	NEB
M28	<i>F⁻ galk2(Oc) IN(rrnD-rrnE)1 rpsL200(strR) rph-1</i>	6mA, 5mC	M. Meselson
DH5α TM	<i>F⁻ Φ80lacZΔM15 Δ(lacZYA-argF) U169 recA1 endA1 hsdR17(rk⁻, mk⁺) phoA supE44 thi-1 gyrA96 relA1 λ⁻</i>	6mA, 5mC	Invitrogen
NEB® 5- alpha	<i>fhuA2 Δ(argF-lacZ)U169 phoA glnV44 Φ80Δ (lacZ)M15 gyrA96 recA1 relA1 endA1 thi-1 hsdR17</i>	6mA, 5mC	NEB
Top10	<i>F⁻ mcrA Δ(mrr-hsdRMS-mcrBC) Φ80lacZΔM15 ΔlacX74 recA1 araD139 Δ(ara,leu)7697 galU galk rpsL (StrR) endA1 nupG</i>	6mA, 5mC	Invitrogen
BL21-AI TM	<i>F⁻ ompT hsdS_B (r_B⁻ m_B⁻) gal dcm araB::T7RNAP- tetA</i>	6mA	Invitrogen
Rosetta TM 2(DE3)	<i>F⁻ ompT hsdSB(r_B⁻ m_B⁻) gal dcm (DE3) pRARE2 (CamR)</i>	6mA	Novagen

Table S4. Genometric correlations between DIP-seq methylation marks and gene and TE annotations in Av-ref and AvL1 assemblies ⁹².

Genometric Correlation	Assembly	Av-ref		AvL1	
IP-seq to TE annotations	Test	IP 4mC-TEs	IP 6mA-TEs	IP 4mC-TEs	IP 6mA-TEs
Relative Ks p-value ^a	Kolmogorov-Smirnov test	1.289e-13	0.0358	0.0083	0.0048
Relative ecdf deviation area ^b	Permutation test	0.0459	0.0070	0.0125	0.0137
Relative ecdf area correlation ^b		0.1842	0.0285	0.0502	0.0540
Relative ecdf deviation area p-value ^b		<0.01	0.01	<0.01	<0.01
Jaccard Measure p-value ^c	Jaccard test	<0.01	<0.01	<0.01	<0.01
Jaccard Measure lower tail ^c		FALSE	FALSE	FALSE	FALSE
Projection test p-value ^d	Projection test	0	0	0	0
Projection test lower tail ^d		FALSE	FALSE	FALSE	FALSE

^aRelative Ks p-value: relative distance test measures whether two sets of positions are closer together or further apart than expected. P-value close to zero: non-uniform distribution (query locations are non-independent of the references).

^bRelative ecdf deviation area p-value: compares the two cumulative distribution functions using the area of the region in which they differ as the test statistic. P-value close to zero: query features are closer than expected to the reference features.

^cJaccard Measure lower tail: measures overlaps between two interval sets by measuring the extent of intersection between two interval sets, divided by the length of their union. FALSE: Overlap is more frequent than expected (p-val <0.01).

^dProjection test lower tail: query intervals are represented as midpoints, but the reference should be a set of intervals. Features are closer (TRUE) or away (FALSE) to reference.

Table S5. Genome assembly and gene annotation metrics.

Species	<i>A. vaga</i> Av-ref ^a		
Accession (assembly name)	GCA_000513175.1 ('2013')	AvL1 Initial assembly ^b	AvL1 hybrid assembly ^c
Coverage ^d (mean)	Not calculated	40.74 (Illumina), 29.82 (PacBio)	33.88 (Illumina), 33.33 (PacBio)
Span (Mb)	217.9	197.1	217.1
No. contigs	36,335	19,202	9,859
Contig N50 (kb)	96.7	22.1	87.4
No. scaffolds	36,167	19,202	9,859
Scaffold N50 (kb)	260.3	19.2	87.4
Scaffold N90 (kb)	7.8	5.3	15.8
Scaffold longest (kb)	1,087.3	167368	747.8
Gaps (Ns) span (kb) (% genome)	4,136.3 (1.9%)	None	None
% GC	30.8	29.9	30.4
BUSCO EUK (n = 429)	C: 89%; F: 3%; D: 8%	C: 84%; F: 8%; D: 8%	C: 88%; F: 6%; D: 6%
BUSCO MET (n = 843)	C: 84%; F: 3%; D: 13%	C: 79%; F: 6%; D: 15%	C: 81%; F: 5%; D: 14%
Annotation			
Method	Augustus	Augustus/GeneMark.ES	Braker/Augustus
No. genes	49,300	61,531	65,934

^a*A. vaga* assembly CGA_00513175.1 ¹⁶.

^bAvL1 initial assembly (no contaminants) GCA_013411005.1 ³³.

^cAvL1 curated hybrid assembly (Illumina + PacBio) from this study.

^dAverage read coverage based on trimmed and filtered data.

Abbreviations: GC, guanine-cytosine; BUSCO, Benchmarking Universal Single-Copy Orthologs; EUK, Eukaryota; MET, Metazoa; C, Complete; F, Fragmented; D, Duplicated.

Table S6. SMRT-seq base modification detection.

PacBio base modification: total nucleotides		4mC		6mA	
bases	total N's Assembly	4mC-10x ^a	4mC-20x ^b	6mA-10x ^c	6mA-20x ^d
	32676087 C's + 75836765 A's	21016	10369	17886	6926
PacBio base modification: total bins with modification					
bins ^e	total bins	4mC-10x	4mC-20x	6mA-10x	6mA-20x
1kb	221401	15965	6536	15078	5364
5kb	49278	11611	4473	11371	3918
10Kb	28432	9001	3503	8937	3131
Bins-TEs ^f	total bins	4mC-10x	4mC-20x	6mA-10x	6mA-20x
1kb	13552	679	235	601	204
5kb	9973	1577	505	1471	420
10Kb	9542	2326	694	2262	604
Bins-Genes ^g	total bins	4mC-10x	4mC-20x	6mA-10x	6mA-20x
1kb	181469	15118	6270	14144	5036
5kb	88412	26892	10518	26346	9161
10Kb	76640	35659	14197	35209	12537

^aSMRT-seq 4mC modified bases with minimum 10x PacBio read coverage^bSMRT-seq 4mC modified bases with minimum 20x PacBio read coverage^cSMRT-seq 6mA modified bases with minimum 10x PacBio read coverage^dSMRT-seq 6mA modified bases with minimum 20x PacBio read coverage^eGenome binning into different bin sizes (1, 5 and 10 kb)^fBins containing TE annotations^gBins containing gene annotations

857 **Table S7.** Primers and oligonucleotides.
858

Name	Sequence (5'→3')	Purpose
N4CMT-F	tttGGATCCgtcatttactaacaatatgtcggt	N4CMT ORF amplification
N4CMT-R	tttCTCGAGccacaaatgtacttttgactcgat	
N4CMT-Cbx-R	tttCTCGAGttgtctKMgacgtaatcgataacca	
N4CMT_Seq1	atcgcggttcgacagtcaat	Sequencing
IAY21-SP-F	aattcgaacgccagcacatg	Site-directed mutagenesis to obtain catalytic mutants
IAY21-SP-R	gatctcagtgggtggtggtg	
IAY21-OP-F	ccagccaataaaacttggcttcgtgaagggt	
IAY21-OP-R	tggagctgtaacaacacattgaacgga	
IAY22-OP-F	ccagccaataaaacttggccttcgtga	
IAY21-Y65A-F	ttgttacagctccaccagc	
		Substrates for <i>in vitro</i> assays:
MTase_subst-1a	ttgaatagttccgCCGgaattttCagtcaa	4mC 30-bp from <i>A. vago</i> AvL1 c1882
MTase_subst-1b	ttgactGaaaattcCGGcggaactattcaa	
Av11_tel_GGGTGTGT	gggtgtgtgggtgtgtggg	<i>A. vago</i> AvL1 19-bp telomeric-like repeat
Av11_tel_CCCACACA	cccacacacccacacaccc	
Av_tel_TGTGGG	tgtgggtgtgggtgtggg	<i>A. vago</i> Av-ref 18-bp telomeric repeat
Av_tel_ACACCC	acacccacacccacaccc	
Av11_c2350-F1	aggagacatccttattgaagca	4mC ~460-bp repeat unit from <i>A. vago</i> AvL1 c1882
Av11_c2350-R1	tcaagtctgttgacctacataagaa	
Av11_4mC/6mA_F1	tgtaccctgcagatgtttgtg	4mC/6mA 580-bp from AvL1 c785 (<i>Athena</i> retroelement)
Av11_4mC/6mA_R1	tctcagacgggctacatgat	
Av11_6mA_F1	tccgcccattccataactgt	6mA, 405-bp from <i>A. vago</i> AvL1 c699 (hnRNP-A)
Av11_6mA_R1	catcatgtgtcaaaggaaatcc	
A11motif-HindIII-F	tttAAGGTTacctaatagcacatatagcc	Contains putative N4CMT recognition motif from <i>A. vago</i> AvL1 c1882 repeat
A11motif-BamHI-R	tttGGATCCtttgatatgcttcaataaggatg	
A11motif-3'209-F1	tcaCcctaccctcatggatt	4mC 209-bp part of repeat unit from <i>A. vago</i> L1 c2350. Used with Av11_c1882-R1 primer.
A11motif-5'200-R1	aatcgatgcttggttgagc	4mC 200-bp part of repeat unit from <i>A. vago</i> L1 c1882. Used with Av11_c2350-F1 primer.
AvL1c2350-364-R	ggatctatgttgagtgtgtgg	4mC 371-bp part of repeat unit from <i>A. vago</i> L1 c1882. Used with Av11_c2350-F1 primer.

859

Table S8. N4CMT recombinant proteins.

Protein_ID	Protein_variant	Length, aa	MW, kDa	pI
N4CMT A'	<i>N6_N4_MTase</i> +Cbx (s23)	426	49.73	8.98
N4CMT A	<i>N6_N4_MTase</i> +Cbx (s23) p.L416R	426	49.78	9.05
N4CMT B	<i>N6_N4_MTase</i> +Cbx (s179)	426	49.75	9.06
N4CMT B'	<i>N6_N4_MTase</i> +Cbx (s179) p.R416L	426	49.71	8.99
N4CMT A-ΔCbx	<i>N6_N4_MTase</i> (s23)	287	33.20	8.60
N4CMT B-ΔCbx	<i>N6_N4_MTase</i> (s179)	287	33.13	8.60
N4CMT A-APPA	<i>N6_N4_MTase</i> +Cbx (s23) p.S62A;Y65A;L416R	426	49.63	8.99
N4CMT B-APPA	<i>N6_N4_MTase</i> +Cbx (s179) p.S62A;Y65A	426	49.65	9.07

Table S9. Summary of N4CMT action on *E. coli* genomic DNA *in vitro*.

<i>E. coli</i> strain	Acquisition of 4mC mark after N4CMT treatment	<i>E. coli</i> genetic background	
		6mA	5mC
Rosetta 2(DE3) (n=2)	++	Dam+ / EcoK1+	Dcm-
BL21-AI	++	Dam+ / EcoK1+	Dcm-
M28	+	Dam+ / EcoK1+	Dcm+
ER2925 (n=2)	+	Dam- / EcoK1+	Dcm-
ER2738	-	Dam- / EcoK1-	Dcm-

Table S10. Summary of ChIP-seq peaks identified by MACS2 (diagonal values) and overlap of peaks within Av-ref and AvL1 assemblies.

Av-ref	H3K4me3	H3k9me3	H3K27me3
H3K4me3	5163	13	27
H3K9me3	-	1902	1811
H3K27me3	-	-	4630
AvL1	H3K4me3	H3k9me3	H3K27me3
H3K4me3	5789	43	48
H3K9me3	-	1205	681
H3K27me3	-	-	2378

Table S11. Gene ontology analysis of methylated and unmethylated genes (.XLSX file).

Table S12. Methylation analysis in under-annotated regions.

contig	start	stop	4mC-10x	6mA-10x	annotations
Contig1073 ^a	0	25272	10	17	Chapaev
Contig1204 ^e	66205	74905	4	10	
Contig1251 ^a	29664	39261	11	10	Polinton-9
Contig126 ^b	34974	38796	2	10	Hebe
Contig1397 ^b	24097	31635	9	10	Juno/AthJN
Contig1606 ^e	50662	52438	10	5	
Contig1615 ^e	4175	10552	9	12	
Contig1743 ^e	0	5188	10	10	
Contig18953 ^e	26550	30514	7	14	
Contig2220 ^a	29384	35415	10	6	Ginger
Contig2425 ^b	43254	59525	1	11	CACTA1
Contig2467 ^b	0	19427	13	2	Helitron
Contig286 ^c	0	7155	1	12	ITS
Contig2879 ^e	2234	4406	19	17	
Contig2948 ^a	0	8375	7	10	Helitron
Contig3423 ^e	1154	6396	13	16	
Contig3571 ^d	0	10942	24	14	TR
Contig3784 ^e	0	9651	8	20	
Contig3893 ^b	4644	35659	14	4	Athena-P
Contig4128 ^c	0	5620	24	16	ITS
Contig4313 ^a	8885	19081	10	4	DNA-N26B
Contig4665 ^e	840	5755	34	10	
Contig480 ^b	0	13847	11	5	Athena-I
Contig523 ^e	36990	42099	12	5	
Contig5325 ^b	8714	25663	15	22	Athena-M
Contig6065 ^b	3940	10974	22	18	TelKA1a
Contig61067 ^b	0	10836	12	18	Vesta1
Contig67 ^e	0	10315	11	10	
Contig805 ^e	3363	6455	3	13	
Contig8145 ^e	15264	20597	4	11	
Contig839 ^e	38934	43045	11	5	
Contig865 ^e	0	11998	11	13	
Contig882 ^b	6424	20046	11	4	MuDR/Mariner
Contig89 ^b	14890	25375	7	11	Vesta1
Contig925 ^e	105036	112889	13	9	
Contig991 ^b	2840	24886	10	2	TR/Sola3/Pen

^aRepbases-TE (5); ^bAdineta-TE (12); ^cITS (2); ^dtandem repeat (1); ^eunknown (16).

Supplementary Note

Gene transcription and DNA modifications. According to SMRT-seq, repetitive regions such as TEs or TRs attract the highest base modification density (Fig. 2g). Still, nearly one-half of tag counts originate from genic loci: a total of 10,928 4mC and 9,596 6mA methylation marks, representing *ca.* 52% and 54% of total 4mC and 6mA, respectively, lie within gene annotations. To examine the links between gene transcription and DNA methylation at base-level resolution, we compared AvL1 transcriptomic data for genes carrying one or more methylation marks in the SMRT-seq dataset (Extended Data Fig. 5a-c), distinguishing those with 1, 2, 3 or more marks (with 4mC and 6mA separately and combined). We used RPKM values to divide genes into subsets with higher ($\text{RPKM} \geq 1$) and lower ($\text{RPKM} < 1$) transcription levels, and to examine correlations with the number of methylation marks. In general, numbers of genes with methylated sites (4mC, 6mA or both combined) and higher RPKM (≥ 1) were significantly higher than those with equal methylation levels but lower RPKM (< 1). Further, although genes with > 3 6mA sites did not show significant differences ($p\text{-val} = 0.71$, χ^2 test for 40 and 33 genes for high and low RPKM respectively), the combined numbers for 4mC and 6mA were significant ($p\text{-val} = 3.02\text{E-}5$) (Extended Data Fig. 5c). Even though some of the methyl marks may be false positives, the observed difference between two gene categories suggests association between methyl marks and genes with higher transcription levels.

To discern the connections between gene methylation and transcription, we explored the occupancy of 4mC and 6mA near the TS in genes. Regardless of the number of modified bases in gene body or in a 2-kb window upstream of TS, methylated genes are consistently expressed at higher levels than unmethylated genes (Extended Data Fig. 5d). Notably, the 6mA occupancy shows a characteristic profile, i.e. a double peak, upstream of TS sites, while 4mC does not (Extended Data Fig. 5e, right panel). Inspection of these patterns with cluster analysis (deepTools option `--kmeans` with `--outFileSortedRegions`) shows that they mainly originate from a total of 1212 gene models carrying the double 6mA mark (*ca.* 260 bp and 750 bp upstream of TS) and no significant accumulation of 4mC sites (Extended Data Fig. 5f). The increased 6mA deposition was corroborated by DIP-seq data, with a significant peak observed upstream of TS for these 1212 genes (Extended Data Fig. 5g); a smaller 4mC DIP-seq peak was also visible further upstream. Comparison of expression levels for these 1212 genes shows that their transcription levels are higher than average (Extended Data Fig. 5h). We then checked their homologs in Av-ref for similarity of methylation and expression profiles. After a blastp search with 1212 AvL1 genes as queries, we obtained 909 *A. vaga* homolog gene models, which not

only showed a similar DIP-seq peak profile (Extended Data Fig. 5,i), but also had expression levels significantly higher than average (Extended Data Fig. 5j). Overall, SMRT-seq modification data agree well with DIP-seq profiles of homologous gene sets in two strains. These observations rule out the possibility of residual RNA-derived 6mA signal and support the view that genic 6mA modifications, particularly those near the TS, are positively correlated with high expression levels of the corresponding genes.

We also performed gene ontology (GO) analysis to find out whether specific gene categories are subject to modification. In AvL1, 66% of genes with 6mA SMRT-seq signature at TS had annotated functions, which is comparable with 63% for all gene models (41,340 genes with GO annotations out of 65,934) (Table S11). Several GO categories displayed significant differences between methylated and unmethylated genes (Fisher's exact test), and the shared groups between Av-ref and AvL1 6mA-methylated genes were rather broad (metabolism, development, catalytic activity, biogenesis). These findings are consistent with designation of 6mA as a developmentally dynamic mark^{79,83,93}.

Haplotype-specific 6mA patterns were suggested to affect allele-specific transcription⁸⁷. Since the AvL1 genome displays the same degenerate tetraploid structure as Av-ref, with 40% of the genome organized in quartets^{16,33}, we searched for allele-specific DNA methylation patterns affecting homologs and/or ohnologs (homeologs). We defined collinear block regions in AvL1 (see Methods) and searched for inter-block differences in base modifications (SMRT-seq) and in transcription levels ($\log_2$ RPKM). Initial inspection suggested that any inter-block transcription differences (Extended Data Fig. 6b) originated from blocks with broken collinearity (i.e., when blocks could not be aligned without rearrangements). Upon comparing the number of base modifications between homologous blocks, only a few cases showed disparity in modified bases (Extended Data Fig. 6c). Out of 28705 pairs in AvL1, with 25619 defined as collinear and 3086 with broken collinearity, 615 and 59 showed difference in SMRT-seq methylation of two or more marks (square root of base modification difference) for collinear and broken pairs, respectively. To establish if collinearity (collinear or broken) and difference in methylation between blocks are independent, χ^2 test was performed using the categories of pairs without base modification difference (value 0 for 16000 collinear pairs vs. 1743 broken pairs) and pairs with any difference (value ≥ 1 in 17611 collinear pairs vs. 1343 broken pairs). The test showed that proportions between both categories are not fully independent, with some association between collinearity and methylation level difference between pairs (χ^2 test, p-val = 3.57E-21).

Finally, we analyzed the remaining AvL1 genomic regions lacking *ab initio* annotations but still displaying significant methylation density. Genomic regions without gene models or

946 annotated TEs/TRs were extracted, and each region was examined for the presence of 4mC
947 and/or 6mA indicated by SMRT analysis. Contigs with methyl marks were further inspected for
948 coverage with Illumina, PacBio, and RNA-seq reads, and showed a lack of transcriptomic
949 coverage, indicative of transcriptionally silent regions (Extended Data Fig. 3f; Table S12).
950 Regions covered by methyl marks were extracted and used in BLAST searches to detect
951 homology to known genes or TEs. From a final set of 36 AvL1 loci with significant numbers of
952 4mC and/or 6mA sites (>10 tags), the analysis revealed one under-annotated tandem repeat,
953 two ITS regions and 17 regions showing homology to TEs (12 from combined *Adineta* TE
954 libraries and 5 from Repbase), indicating that one-half of the extensively methylated,
955 transcriptionally silenced regions represents under-annotated TEs (Table S12).

Supplementary Discussion

Base modification, primarily in the form of methylation, constitutes an important facet of epigenetics due to the covalent nature of its linkage to DNA. In eukaryotes, the archetypal 5mC modification dominates the epigenetic landscape, and its distribution patterns are established by concerted action of writers, readers and erasers of epigenetic marks. The maintenance and *de novo* MTases *Dnmt1* and *Dnmt3* act together with demethylases to set the levels of CpG methylation^{5,94}, and may have been doing so since the divergence of plants and animals⁹⁵ (but see⁹⁶). In bacteria, the most widespread DNA modifications added by amino-MTases of R-M systems modify the exocyclic amino groups of adenines and cytosines, with 5mC constituting a distant third³⁰ because of high incidence of 5mC→T transitions prone to deamination. The 6mA modification is widespread in eukaryotes, although its role is still debated, especially when its levels are particularly low^{9,11,36,58}. Our assessment of *A. vago* methylome broadly agrees with the view of 6mA as a dynamic context-dependent mark, associated with higher expression in a subset of genes but also found over repressed TEs.

The overall level of 4mC, as revealed by SMRT-seq, amounted to 0.065% of cytosines, while the somewhat lower 6mA content (0.024% of adenines) is still higher than 6mA levels in *X. laevis* or mouse, and is comparable to values reported for *C. elegans*, *Drosophila*, plants, and humans (Table S1). Lower 6mA levels are not surprising, as this dynamic mark can be modulated in a tissue- and stage-specific fashion by balancing activities of N6A-MTases and AlkB-like demethylases, which are present in bdelloids. Indeed, a higher fraction of 4mC may be due to lack of enzymes responsible for active cytosine demethylation, such as TET or potential homologs of bacterial R-M enzymes recognizing 4mC. Notably, both 4mC and 6mA sites show an asymmetric pattern, in contrast to symmetrical MTases, such as *Dnmt1* in mammals or N6A-MTases of ciliates and early-diverging fungi, which act on hemi-methylated DNA^{4,72,77}. The lack of maintenance MTases acting on symmetric motifs implies that methyl marks should be added *de novo* after DNA replication to be maintained at specific sites.

Interestingly, the preferred symmetrical dinucleotide for 4mC addition (CpG) coincides with that of canonical C5-MTases, although the asymmetric CpA is also frequently utilized. This agrees with higher similarity of N4CMT to bacterial MTases with a CG doublet in their recognition motif, suggesting that recognition by TRD may contribute to target choices. For 6mA addition, the asymmetric ApG or GpA are the preferred sites, as in other metazoans with similar 6mA content (Table S1). Symmetrical 6mA addition to ApT dinucleotides occurs in green algae,

ciliates, and early-diverging fungi, where it is associated with actively transcribed genes and with linker DNA between nucleosomes^{71-73,77}. The sequence context of methylation sites may further contribute to recognition of methyl marks by various readers, often resulting in opposite transcriptional effects, as shown for 5mC or 6mA²².

It is hardly a coincidence that N4CMT is most closely related to MTases of cyanophages rather than bacteria. Indeed, phage-borne orphan MTases, in addition to being prone to horizontal spread, may be under evolutionary pressure to broaden their sequence specificity to protect the phage from multiple bacterial R-M systems⁹⁷. An MTase with strict target specificity is unlikely to cover a broad range of epigenetic targets, limiting its regulatory potential, and would benefit from reduced sequence specificity while acquiring chromatin-based targeting. The intrinsic N4CMT target preference adds an intriguing twist to this view. While this preference is seen in our *in vitro* assays and is manifested *in vivo* as high-density SMRT modification regions, these regions do not show an increased density of H3K9/27me3 histone marks over modified TRs (Extended Data Fig. 3a). In *C. elegans*, a SETDB1 homolog *met-2* adds H3K9me2 marks to suppress transcription of satellite repeats, which in *met-2* null worms yield DNA-RNA hybrids and trigger DNA damage-induced germline lethality⁹⁸. Although investigations of DNA damage are outside the scope of the present work, future studies may uncover additional pathways involving chromodomain-independent N4CMT activity targeting 4mC to silence satellite repeats. Indeed, the 460-bp repeat in AvL1 is fully silenced, since our qRT-PCR experiments failed to yield PCR products (data not shown). It also remains to be seen whether H3K9me and H3K27me exhibit spatial and functional overlap in bdelloids, as in ciliates^{44,45}, or are separated in space/time, with such studies being impeded by syncytial organization. Although in *Drosophila* and mammals H3K9 denotes constitutive and H3K27 - facultative heterochromatin, and the relevant enzymatic machinery is represented by SUV39H, G9a and SETDB1 homologs for H3K9 and the EZH-containing Polycomb repressive complex 2 for H3K27, the lack of SUV39H-like and G9a-like proteins in bdelloids may indicate their replacement with diversified SETDB1 paralogs for H3K9 methylation, and supports a facultative, bivalent nature of their heterochromatin, as indicated by triple Tudor domains in SETDB1 which recognize a combination of active and repressive histone marks to ensure silencing⁹⁹.

Supplementary Figures

Extended Data Fig. 1. Base and dinucleotide composition features in *A. vago*. **a**, Distribution of the observed/expected ratio of CpG dinucleotide frequency in Av-ref and AvL1 assemblies in a 1-kb sliding window (left panel) and in CDS regions (right panel). Its mean value, 1.103 and 1.032 for Av-ref and AvL1, respectively, indicates the lack of pronounced 5mC deamination signatures in gDNA. CDS ratio was calculated per gene, with 1.096 and 0.990 as mean CpG obs/exp values for Av-ref and AvL1, respectively. **b**, Nucleotide and dinucleotide composition frequencies across AvL1 TE annotations and 5' upstream regions.

Extended Data Fig. 2. DIP-seq peak distribution near genes and transposons in AvL1. a-b, Distribution of 4mC and 6mA DIP-seq peaks calculated by MACS (see Methods) around gene (a) and TE (b) annotations in AvL1, with coverage of peak deposition located within and in the proximity of annotated features. Peak coverage is represented in 25-bp bins within 2.5 kb upstream and downstream. The body size feature, representing genes or TEs, is automated and normalized as meta-profile (0-100% of body length). **c-d,** Profiles (top) and heat maps (bottom) in AvL1 showing relative fold enrichment of DIP-seq reads for 4mC and 6mA. (c), gene regions with TS (transcription start) and TT (transcription termination) sites and their vicinity (± 3 kb). (d), TE annotations within 5' (5) and 3' (3) sites and near insertion points (± 3 kb). The right-hand side in C and D represents profiles and heatmaps of DIP-seq reads over features divided into four clusters (deepTools--kmer 4; Methods).

Extended Data Fig. 3. DNA and histone modifications on selected contigs. Circos plots of Av-ref (Av) and AvL1 (As) profiles show 7 layers with annotations, sequencing coverage, methylation sites, and DIP/ChIP peaks. Inside to outside: TEs, grey bars; genes, orange bars; green line, RNA coverage for Av-ref and AvL1 transcriptomes; purple line, small RNA coverage in Av-ref. Blue histogram, % G+C (light-blue, low GC; dark-blue, high GC). In DNA sequencing coverage plots, red and blue represent coverage by Illumina and PacBio reads, respectively (Methods). In AvL1 PacBio layer, DNA methylation sites for 4mC (blue triangle) and 6mA (red square), with height in the ring showing methylation fraction (from 0 to 1). Histone methylation peaks for H3K4, H3K9 and H3K27: orange, green and black bars, respectively. Blue and red bars, DIP-seq 4mC and 6mA peaks. Contig/scaffold ideograms are plotted as black bars, with labels for Av-ref (Av) and AvL1 (As) showing the numbers from source assemblies GCA_000513175 and AvL1, respectively. Label ticks are distanced at 5 kb. Coverage layers were calculated with 1-kb sliding window.

Extended Data Fig. 5. Correlation between gene transcription and SMRT modification marks. Numbers of methylated genes with 1, 2, 3 or more (>3) marks for 4mC (**a**), 6mA (**b**) and both combined (**c**) are plotted with their RNA-seq transcription profiles (high for RPKM $\geq$ 1, low for RPKM<1). Asterisks show statistically significant differences in the numbers of methylated genes between high and low RPKM levels (χ^2 test, ** $p < 10^{-4}$, * $p < 0.05$). (**d**) Boxplot comparing expression (log2RPKM) of genes with no methylation; methylation in the gene body; around TS (2 kb upstream); and both methylated regions combined (TS plus body). Box represents the first and third quartiles; line, the median. RPKM, reads per kb per million mapped reads. **e**, **f**, 4mC and 6mA SMRT-seq methylation in the final gene set ($n=65,934$) (**e**) and clusterized for 6mA signal in the TS region ($n=1212$) (**f**). **g**, Distribution of 4mC and 6mA DIP-seq peaks around AvL1 genes with 6mA modification ($n=1212$) showing coverage of peak deposition within and near genes. In **h**, **j**, genes marked with 6mA around TS exhibit higher expression than genes without it. **i**, **j**, Distribution of 4mC and 6mA DIP-seq peaks (**i**) and RPKM (**j**) in Av-ref genes homologous to AvL1 genes in (**e-h**). In (**g**, **i**), peak coverage is shown in 25-bp bins within ± 2 kb. The body size feature, representing genes, is automated and normalized as a meta-profile (0-100% of body length). In (**h**, **j**), boxplot compares the median and interquartile range of RPKM expression levels. The p-values were calculated by a two-tailed Student's t-test.

Extended Data Fig. 6. Comparisons of AvL1 gene transcription in ChIP-seq peaks and in collinear blocks. **a**, Box plot showing AvL1 gene expression levels (log₂RPKM) associated with co-localized H3K4me3, H3K9me3, H3K27me3, and H3K9-27me3 marks or without histone marks. ANOVA analysis shows significant differences in expression (one way ANOVA, $Df = 4$, $F = 1331$), with genes associated with the H3K4me3 mark displaying the highest RPKM (reads per kilobase per million mapped reads). **b**, Transcription in collinear blocks. Points represent collinear blocks of genes, plotted based on the transcription values (log₂RPKM) on block1 (X-axis) and transcription on block2. Syntenic blocks are differentiated between two groups: collinear (homologous genes with low Ks values) and blocks in which collinearity has been broken (homeologous genes with high Ks values). **c**, Base modification versus transcription differences. X-axis represents the difference between blocks in the number of detected SMRT base modifications (square root), and Y-axis represents the absolute difference in log₂RPKM between them.

Extended Data Fig. 7. Distribution of 4mC and 6mA peak counts between different transposon types. Shown are the mean peak counts from DIP-seq data for 4mC (**a**) and 6mA (**b**) near each type of annotated TEs, classified as Helitron, LTR, non-LTR (LINE), Athena, Penelope or TIR, extending the 5' and 3' transposon ends by 500 bp window size. ANOVA analysis shows differences in distribution of 4mC (one way ANOVA, Df = 5, F = 39.871, p-val < 2.2 E-16) and 6mA peaks (one way ANOVA, Df = 5, F = 27.753, p-val < 2.2 E-16) near specific TE families.

Extended Data Fig. 9. MBD proteins from the *A. vago* genome. **a**, Diagram of recombinant MBDs used in this study. S, S tag; His, His tag; MW, Molecular weight of protein in kDa. $\alpha 1$, $\beta 1$, $\beta 2$, $\beta 3$, secondary structure elements outlined in Extended Data Fig. 8a. **b-c**, Binding of AvMBD to unmethylated (b) and 4mC-methylated by N4CMT_A (c) DNA. For electrophoretic mobility shift assays, 2 ng of ³²P-sAvL1-451 were used with 50-100 ng of purified AvMBD proteins.

METHODS

Rotifer cultures. A clonal culture of *Adineta vaga*, started in 1995 from a single individual, was maintained continuously in filtered spring water and fed with *E. coli* M28. Rotifers were grown in 150x20 mm untreated Petri dishes and transferred into new ones, until the desired biomass was reached. The *A. vaga* L1 natural isolate³³ was collected in 2012, and the clonal culture was maintained in the laboratory under the same conditions.

Plasmid construction. N4CMT ORFs from scaffold_23 (GSADVT00006927001, allele *N4CMT_A*) and scaffold_179 (GSADVT00035445001, allele *N4CMT_B*) (<http://www.genoscope.cns.fr/adineta/cgi-bin/gbrowse/adineta/>) were amplified from cDNA to eliminate introns. The first exon in the annotation is ambiguous and variable in different bdelloids, thus it was omitted from primer design, so that the N-terminus coincides with that used by bacterial MTases. Briefly, RNA was extracted from adult rotifers starved for 24 hours, using Direct-zol™ RNA Miniprep kit (Zymo Research), and cDNA was synthesized from 2 µg of RNA with SuperScript® IV Reverse Transcriptase (Invitrogen) and random hexamers, following the manufacturers' protocols. *N4CMT* was then amplified by PCR from 5% of cDNA reaction with Q5® Hot Start High-Fidelity DNA Polymerase (NEB). All primers used in this study are listed in Table S7. PCR fragments were cloned into pET29b(+) vector (Novagen) using *Bam*HI and *Xho*I sites and propagated in *E. coli* NEB5-alpha (NEB). Catalytically inactive mutants were obtained using Gen-Edit™ site-directed DNA mutagenesis kit (First Biotech). To obtain substrate plasmids pUC19-m97 and pUC19-m119, the insert sequence was amplified from AvL1 genomic DNA with primers A11motif-Hind3-F and A11motif-BamHI-R (Table S7) and OneTaq® Hot Start DNA Polymerase (NEB). Amplicons were treated with *Hind*III (Anza™ 16) and *Bam*HI (Anza™ 5) in 1x Anza™ Red Buffer (Thermo Fisher Scientific) and purified through 1.5% agarose gel using Zymoclean Gel DNA Recovery kit (Zymo Research). The pUC19 vector was prepared in the same way, ligated with insert using Instant Sticky-end Ligase Master Mix (NEB) and transformed into NEB5α competent cells (NEB). Plasmid purifications were done with Zippy Plasmid Miniprep (Zymo Research). Inserts were verified by Sanger sequencing on the ABI3730XL at the W. M. Keck Ecological and Evolutionary Genetics Facility at the Marine Biological Laboratory. Expression plasmids carrying AvMBD sequences in pET29b(+) vector were synthesized by GenScript. All DNA sequences were optimized with GenSmart™ service to produce soluble recombinant proteins in *E. coli*.

Protein expression and purification. Recombinant proteins were expressed in *E. coli* Rosetta 2(DE3) (Novagen) in LB medium, Miller formulation (Amresco) supplied with 50 µg/ml kanamycin (Fisher Scientific), 34 µg/ml chloramphenicol (Acros Organic). First, cells were grown at 37°C, 200 rpm until OD=0.4. After that, cultures were heat-shocked as follows: 10 min at 42°C, 20 min at 37°C, 30 min on ice, and 20 min at 37°C. After final OD check, expression of recombinant proteins was induced by supplying the growth medium with IPTG (Gold Bio) to 500 µM, and the culture was grown for additional 4 h at 32°C, 350 rpm for N4CMT versions or for additional 3 hours at 34°C, 300 rpm for AvMBDs. Bacterial cells were pelleted by centrifugation at 4°C, 4000 g for 30 min and stored at -80°C. Induction of recombinant proteins was confirmed by SDS-PAGE followed by Western blot hybridization as described in¹⁰⁰. For protein purification, cellular lysates were prepared using xTractor™ Buffer (Clontech), supplemented with lysozyme (Sigma), DNase I (Promega) or Benzonase® Nuclease (Sigma), and Roche cComplete™ EDTA-free Protease Inhibitor Cocktail (Sigma), according to the manufacturers' instructions. Soluble proteins were separated from insoluble debris by centrifugation at 4°C, 4000 g for 30 min. Recombinant N4CMT were purified using TALON® Single Step Columns (Clontech), following the manufacturer's protocol. Proteins were concentrated using Pierce™ 9K MWCO Protein Concentrators (Thermo Scientific), and the buffer was exchanged to 50 mM phosphate buffer, 300 mM NaCl, pH 7.0 supplemented with Roche cComplete™ EDTA-free Protease Inhibitor Cocktail (Sigma). Protein concentrations were equalized based on concentration of the full-length His-tagged protein as detected by Western blotting with His-tag-specific antibodies (Aviva Systems Biology) using Image Studio™ Lite 5.2.5 Software (LI-COR). Purified proteins were stored at 4°C for up to 2 weeks. Recombinant AvMBD's were purified on ÄKTA Pure M2 with HiTrap TALON 1 ml columns (Cytiva), concentrated with Pierce™ 3K MWCO Protein Concentrators PES (Thermo Scientific), supplied with EDTA, glycerol and protease inhibitors to the final buffer composition of 40 mM sodium phosphate, pH 7.4; 240 mM NaCl; 102 mM imidazole; 20% glycerol; 4 mM EDTA; 1x cComplete protease inhibitor cocktail; 1x Halt protease inhibitor cocktail; pH 7.4. Proteins were stored in single-use aliquots at -80°C. Proteins were quantified using Micro BCA™ Protein Assay Kit (Thermo Scientific), and their purities verified by SDS-PAGE in 15% resolving gel followed by staining with

InstantBlue Protein Stain (Expedeon) and Western blotting with S-tag (Novagen) and His-tag (Aviva Systems Biology) specific antibodies, both at 1:5000 dilutions, as described in ¹⁰⁰.

DNA substrate preparation for methylation assays. The *A. vaga* cultures were maintained as above but fed with *dam-/dcm- E. coli* (C2925, NEB) strain instead of M28 for a month. Genomic DNA was extracted from adult rotifers starved for 48 hours, following the standard phenol-chloroform extraction protocol ¹⁰¹. To obtain control DNA from different *E. coli* strains (Table S3), bacteria were grown overnight in LB medium Miller formulation (Amresco) at 37°C and 200 rpm, and total DNA was extracted using UltraClean® Microbial DNA Isolation Kit (MoBio Labs).

For N4CMT *in vivo* activity assays, plasmids carrying N4CMT sequences were introduced into Rosetta 2(DE3) strain. Bacteria were grown as above, pelleted and stored at -80°C until expression of recombinant proteins was confirmed by Western hybridization with His-tag-specific antibodies. After that, bacterial pellets were incubated in lysis buffer (10 mM Tris, pH 8.0, 100 mM NaCl, 5 mM EDTA, 120 µg/ml Proteinase K (ThermoFischer), 0.6% SDS) at 53°C overnight. Total DNA was purified following the standard phenol-chloroform extraction protocol ¹⁰¹, including treatment with RNaseONE (Promega). DNA quantity and quality were inspected by agarose gel electrophoresis and NanoDrop 2.0 measurements. Cleavage of genomic DNA by McrBC (NEB) was performed overnight at 37°C as recommended by the manufacturer, followed by DNA separation in 0.8% TAE-agarose gel electrophoresis. Plasmids (pUC19, pBlueScript SK+ etc.) for methyltransferase assays were transformed into methylation-free C2925 competent cells (NEB) and purified using Zippy Plasmid Miniprep (Zymo Research). To obtain a 4mC-positive control for immunoassays, pUC19 was methylated with *M. BamHI* methyltransferase (NEB). To obtain a positive control for 6mA, pUC19 was purified from NEB5α (*dam+*) *E. coli* strain. Oligonucleotides were ordered from Eurofins Genomics and annealed in 1x annealing buffer (10 mM Tris, pH 7.5, 50 mM NaCl, 1 mM EDTA) as follows: the mix was incubated at 95°C for 3 min and allowed to cool down to RT for 1 h. Other dsDNA substrates were obtained by PCR and purified using Monarch PCR clean-up kit (NEB) or Zymoclean Gel DNA Recovery kit (Zymo Research).

In vitro methyltransferase activity assays. Reactions were carried in 1x *M. BamHI* Methyltransferase Reaction Buffer (NEB) supplemented with 80 µM S-adenosyl-L-methionine (SAM) provided with the buffer. Optimal results were obtained with 500 µg/ml as a final concentration of N4CMT recombinant proteins. Reactions were initially incubated at 25°C for 4 h, and incubation was continued for another 16 h after supplementing with additional 80 µM SAM.

DNA dot blot immunoassays. Samples were spotted on BioTrace™ NT Nitrocellulose Transfer Membrane (Pall Corporation), air-dried and UV-cross-linked with 120,000 µJ/cm² exposure using Spectrolinker™ XL-1500 UV crosslinker (Spectronics Corporation). The cross-linked membrane was blocked in 3% non-fat milk in TBST (containing 0.05% v/v Tween) and incubated with 1:40,000 anti-N4-methyl-C antibody or with 1:60,000 anti-N6-methyl-A antibody at 25°C for 1 h. Rabbit primary antibodies raised against 4mC- or 6mA-modified DNA ¹⁰² were a kind gift from Dr. Iain Murray (NEB), and were checked for the absence of cross-reactivity, as well as for lack of reactivity with 5mC on human DNA. The membrane was washed three times with TBST, incubated with 1:10,000 goat anti-rabbit HRP antibody (Sigma) at room temperature for 1 h, washed three times with 1x TBST, and developed using SuperSignal™ West Dura Extended Duration Substrate (Thermo Fisher Scientific). Chemiluminescence was detected using the Amersham Imager 600 chemiluminescence imager (GE Healthcare).

Electrophoretic mobility shift assays. sAvL1-451 DNA were 5'-end-labeled with [γ-³²P]dATP (PerkinElmer) using T4 polynucleotide kinase (NEB) and purified from excess of radioactive nucleotides using Oligo Clean & Concentrator kit (Zymo Research) following the manufacturers' protocols. Binding reactions were set up in 10 µl total volume in a buffer with final concentrations 100 mM KCl, 10 mM Tris, pH7.4, 0.1 mM EDTA, 0.1 mM DTT, supplied with 500 ng LightShift™ Poly (dI-dC) (Thermo Scientific). Addition of 2.5 µl of AvMBD proteins provided 5% glycerol per reaction. Proteins were first pre-incubated with non-radioactive DNA for 15 min at RT. Then, ³²P-labeled DNA was added to a final concentration of 0.05 nM, and reactions were incubated for additional 30 min at RT. After supplying with 6X EMSA gel-loading solution (Thermo Scientific), samples were loaded onto 6% DNA Retardation gels. Samples were run at 90 V in 0.5x TBE buffer (44.5 mM Tris-HCl, pH 8.3, 44.5 mM boric acid and 1 mM EDTA) at 4°C for

90 min. Gels were dried using Model 583 Gel Dryer (BioRad), exposed with phosphorimaging plate (Fujifilm), scanned on Typhoon FLA 7000, and analyzed using Image Quant TL v8.1 software.

DNA extraction for DIP-seq. For genomic DNA extraction, animals were starved for 48 h and treated with ampicillin and tetracycline antibiotics (final concentration of 10 mg/ml and 0.5 mg/ml, respectively) for 24 hours, then harvested as described in ¹⁶. Total DNA was extracted with DNeasy Tissue kit (Qiagen), final eluates were checked by agarose gel electrophoresis and final concentrations were measured by Nanodrop. The isolated genomic DNA was diluted to ~250 ng/μl using TE buffer and sonicated on the 130 μl scale (Covaris microtubes) to 200-400 bp using Covaris S220 focused ultrasonicator (10% duty cycle, 175W peak, 200 cycles, 180 sec, 6°C). After measuring concentration and size distribution with Bioanalyzer High Sensitivity DNA chip (Agilent), 100 ng of fragmented DNA was used for library construction with NuGen Ovation Ultra Low System v2.

DIP-seq (MeDIP-seq). After adaptor ligation and purification steps (NuGen Ovation Ultra Low System v2 protocol), DNA fragments were combined with 0.5 μg of anti-4mC or anti-6mA antibodies (see above) in 500 μl of 1x IP buffer and incubated at 4°C for 6 h. In parallel, 40 μl of Protein A magnetic beads were prepared as in ⁷³. Protein A beads were added to DNA-antibody mixture and incubated at 4°C overnight with rotation. Beads were washed four times with 1x IP buffer on a magnetic rack. 20 μl of proteinase K (20 mg/ml) were used to release the methylated DNA with 3 h of incubation at 50°C. The final eluate was purified using 2x phenol-chloroform-isoamyl alcohol (25:24:1) extraction and ethanol precipitation. DNA was resuspended in 35 μl H₂O, followed by library amplification and bead purification (NuGen RNAClean XP magnetic beads). Quality control and concentration measurement were performed using Bioanalyzer DNA 1000 chip (Agilent) and Qubit sDNA HS Assay kit (Thermo). Libraries were sequenced using the Illumina HiSeq 2500 platform (50-bp SR) at the Brown University Sequencing Core Facility. Base calling was performed with the standard Illumina pipeline (Casava 1.8.2). Illumina adaptors were trimmed with cutadapt ¹⁰³, as well as any sequence with low quality score (<Q20) and/or <16 nucleotides in length (FASTX Toolkit). Reads were aligned to Av-ref ¹⁶ and AvL1 assembly (see below) using Bowtie ¹⁰⁴, with parameters permitting less than one mismatch in the first 30 bases. MACS software ¹⁰⁵ was used to locate enriched regions for 4mC and 6mA in both genomes (using nomodel nolambda parameters). Downstream processing of mapped reads and peaks was performed using a variety of available software and custom Linux scripts.

Genome assembly. The initial *A. vaga* L1 isolate draft assembly was generated with high quality paired-end Illumina MiSeq reads using SPAdes assembler to yield N50 of 18.125 kb ³³. However, the published AvL1 assembly filtered any sequences without blastn matches to Av-ref, which may include recent horizontal transfers and TEs. To improve the initial assembly, DNA was extracted from rotifer eggs as in ³⁹, and a 20-kb library was constructed using BluePippin selection to sequence 15 SMRT cells on a PacBio RS II sequencer (Pacific Biosciences) at the Johns Hopkins University Deep Sequencing and Microarray Core facility with P6-C4 chemistry (accession number PRJNA558051). We used PBJelly from PBSuite 15.8.24 ¹⁰⁶ with PacBio filtered subreads to improve the initial AvL1 assembly. A total of 890,504 PacBio subreads with N50 read length of 16,294 bp was used after SFilter (Pacific Biosciences) and spike-in control removal. The improved hybrid assembly was filtered from contaminants using bacterial single-copy genes, GC-content, k-mer frequencies (k=4), and DNA coverage values (both Illumina and PacBio) following ¹⁰⁷. Assembled contaminant contigs, mostly of bacterial origin, were filtered out to yield a final assembly totaling 217.1 Mb in 9,856 contigs (Table S5), which is very close to the 218-Mb Av-ref assembly ¹⁶ and improves by 20 Mb the Illumina-only assembly, increasing N50 from 22.1 kb to 87.4 kb. We also identified 12 chimeric contigs, listed in Supplementary Data File S4, which were mostly eukaryotic with an attached small stretch of bacterial DNA showing high methylation density. The AvL1 assembly used in this work was deposited in Genbank (accession No. JAGENE000000000) and can be downloaded at <https://jbpc.mbl.edu/media/frodriguez/public/AvL1/>.

PacBio modification analysis. We examined genome-wide distribution of modified bases in SMRT-seq data ¹⁰⁸ with SMRT Analysis Software 2.3.0. Raw data from 15 AvL1 SMRT cells were filtered by SFilter (Pacific Biosciences) to remove reads containing adapters, short reads and low-quality reads with cutoffs for read quality ≤ 0.75, read length ≤ 50 nt, and subread length ≤ 50 nt. Filtered reads were aligned to the AvL1 assembly using RS_Modification_Detection.1 protocol (Pacific Biosciences). Briefly, the cleaned

reads were aligned to AvL1 curated genome assembly using blasr¹⁰⁹. The polymerase kinetics information was processed and reported as IPD ratio, with its fraction (the methylated portion of reads mapped) at each site. The 4mC and 6mA base modifications were identified, and the final report was extracted as csv and gff files for posterior processing. Filtering was performed by selecting only 4mC and 6mA marks with 20x coverage and mQv $\geq$ 22 (Table S6); any sites with coverage <10x were removed. Although SMRT analysis may sometimes erroneously identify 5mC as 4mC, as occurred for the *fig* genome¹¹⁰, which has a full complement of plant 5mC-MTases but no N4C-MTases, we are confident that multiple orthologous methods applied to *A. vaga*, which lacks 5mC-MTases but has the N4C-MTase, validate our SMRT-seq cytosine modification calls as 4mC. Additional analyses were done with custom scripts for plotting results with R. We separated 4mC and 6mA according to their methylation levels: low-fraction sites (0.1-0.5), moderately methylated (0.5-0.8) and highly methylated (0.8-1). The upstream and downstream 10-bp sequences from 4mC and 6mA modification sites were extracted for motif identification in each group by MEME-ChIP¹¹¹.

Dot-blot immunoassays for histone marks. We first assayed, by dot-blot analysis, the reactivity of *A. vaga* histone methylation marks with Premium ChIP-seq grade affinity-purified rabbit polyclonal antibodies H3K4me3, H3K9me3 and H3K27me3, raised against synthetic peptides with the corresponding trimethylated lysines (Diagenode C15410003, C15410056 and C15410195, respectively). These antibodies display a wide range of species reactivity including vertebrates, *Drosophila*, *C. elegans* and plants, and have been tested by ChIP-seq, IF, Western blotting, and ELISA. The H3 N-terminal residues 1-31 display 100% identity between *A. vaga* and humans; although formally cross-reactivity of K9/27 cannot be excluded for *A. vaga*, however none was observed in human peptide arrays spanning identical aa sequence (Diagenode). Protein extracts from Av-ref and AvL1, resuspended in 0.5 v of extraction buffer (10 mM Hepes, 5 mM MgCl₂, 2 mM DTT, 10% glycerol and cOmplete protease inhibitor tablets (Roche)), were spotted on BioTrace™ NT Nitrocellulose Transfer Membrane (Pall Corporation), air dried and blocked in 5% BSA in TBST (containing 0.05% v/v Tween) for 1 h at RT and incubated with 1:10,000 anti-H3K4me3, H3K9me3 or H3K27me3 antibodies at RT for 1 h. The membrane was washed three times with TBST, incubated with 1:10,000 goat anti-rabbit HRP antibody (Sigma) at room temperature for 1 h, washed three times with 1x TBST, then once with TBS and developed using SuperSignal™ West Dura Extended Duration Substrate (Thermo Fisher Scientific). Chemiluminescence was detected using the Amersham Imager 600 chemiluminescence imager (GE Healthcare).

ChIP-seq. Chromatin immunoprecipitation (ChIP) was performed based on the *C. elegans* protocol¹¹² with minor modifications. Briefly, rotifers were starved for 48 h before collection, and live animal pellets were washed with PBS, followed by another round with protease inhibitor (cOmplete Roche tablet). The 1-ml pipette tip was used to drip mix into a porcelain mortar containing liquid nitrogen, and the frozen rotifer “popcorn” was ground to fine powder with a pestle. Nuclear proteins were cross-linked to DNA by adding 1.1% formaldehyde (Thermo) in PBS + 1x protease/phosphatase inhibitors (Halt™ Protease & Phosphatase Inhibitor Cocktail, Thermo) for 10 min at room temperature on a rocking platform. Cross-linking was stopped by adding glycine to a final concentration of 0.125 M and incubating for 5 min at room temperature. The medium was removed, and the cells were washed twice with ice-cold PBS containing 1 mM PMSF. The cells were then collected in FA lysis buffer (FA buffer + 0.1% sarkosyl + protease/phosphatase inhibitors); FA buffer: 50 mM HEPES/KOH pH 7.5, 1 mM EDTA, 1% Triton™ X-100, 0.1% sodium deoxycholate; 150 mM NaCl. Subsequently, the chromatin was isolated, sonicated (Covaris S220: 2% Duty Cycle, 105W Peak, 200 Cycles, 360 sec, 6°C), and immunoprecipitated with anti-H3K4me3 antibody, anti-H3K27me3 antibody, anti-H3K9me3 antibody (all from Diagenode) or no antibody (input control). After reverse-cross link (overnight at 65°C), DNA was purified by using 2x phenol-chloroform-isoamyl alcohol (25:24:1) extraction and ethanol precipitation. DNA was resuspended in 35 μ l 10 mM Tris-Cl, pH 8.5. The ChIP DNA and input DNA were used to construct ChIP-seq libraries using NEBNext Ultra II DNA Library Prep Kit (NEB) following the manufacturer’s procedure. The libraries were sequenced on Illumina NextSeq 500 platform for 75 bp single-end HT at the W.M. Keck Sequencing Facility at the MBL. After demultiplexing and adapter trimming (bcl2fastq software, Illumina), the raw reads were cleaned up to obtain high-quality reads (see parameters in IP-seq). Clean reads were mapped to Av-ref and AvL1 assemblies using bowtie2¹¹³ with default parameters. Genomic regions associated with histone modification were identified using Model-based Analysis of ChIP-Seq (MACS2)¹⁰⁵ using default parameters.

RNA-seq. For *A. vaga* Av-ref transcriptome, total RNA was extracted from animals at all life-stages with TRIzol® (Invitrogen) following manufacturer's protocol with a glass Dounce homogenizer. After DNase I (NEB) treatment on RNA Clean & Concentrator columns C-5 (Zymo Research), *A. vaga* total RNA was eluted and subjected to poly-A selection with Ambion MicroPoly(A) Purist Kit (Thermo Fischer). Libraries were prepared with Encore Complete Library RNA-Seq Library Systems (NuGen). A total of 3 biological replicas were sequenced on a dedicated Illumina NextSeq Mid lane (1x150bp) and, after QC (http://hannonlab.cshl.edu/fastx_toolkit/) and adapter trimming (Cutadapt v1.9.2)¹⁰³, mapped to Av-ref¹⁶ with Tophat 2.1.1¹¹⁴, using default parameters and --max-intron-length 100. Aligned sequence reads were counted by genomic feature with HTSeq-count¹¹⁵, using default parameters. For AvL1 transcriptome, RNA extraction was performed following¹⁶ for the fully hydrated *A. vaga* L1 cultures containing animals at all life-stages. Rotifers were collected by centrifugation at 10,000 rpm. After removal of the supernatant (spring water), total RNA was extracted with Trizol (Invitrogen) followed by ethanol precipitation. After DNaseI treatment (DNA-free, Ambion), 1 µg of total RNA was shipped for QC, library preparation (eukaryotic mRNA protocol) and Illumina sequencing (HiSeq x PE150bp) to Novogene Co., Ltd. Raw reads (~3.3 Gb) from two lanes as technical replicates were processed (see parameters in IP-seq), and properly paired reads were aligned to the AvL1 assembly using TopHat v2.1.1¹¹⁶, using default parameters and --max-intron-length 100. Mapped reads were counted within each feature with HTSeq-count¹¹⁵ using default parameters, which was used to calculate RPKMs of annotated genes.

Prediction of protein-coding genes. BRAKER¹¹⁷, a combination of GeneMark-ET¹¹⁸ and AUGUSTUS¹¹⁹, was used to predict protein-coding genes in the AvL1 genome using aligned RNA-seq data. TopHat alignments were used to generate UTR training examples for AUGUSTUS to train UTR parameters and predict genes. This procedure was done with --softmasking enabled, after masking the genome with RepeatMasker (see *Repeat annotation*). Total predictions comprised 74,569 gene models originating from 74,233 loci. Initial predictions were filtered from TE genes using AvL1 TE annotations (RepeatMasker) and BLAST homology search to known TE proteins. BLAST searches were performed with 74,569 gene predictions using blastp (blast+) and blastx (diamond blast) onto nr and uniref90 databases, respectively. BLAST descriptions with TE- related terms ("transposon", "transposable", "integrase", "reverse transcriptase", "pol", "gag") were considered as TE- associated proteins. A total of 977 genes were classified as AvL1 TE-related. A further quality check of gene annotations filtered incomplete genes. Annotations at the contig boundaries were removed (n = 5205), along with CDS that carried a premature stop codon (n = 282) or without appropriate termination codon at the CDS end (n = 2748, which mostly fall on contig boundaries). A final filter was applied to remove annotations with no BLAST homology (neither nr nor uniprot) and for which CDS sequence was under 300 bp. A final gene set of 65,934 annotations was used for downstream analysis.

Repeat annotation. We used the REPET package with default settings for initial AvL1 *de novo* TE identification and annotation¹²⁰. The automated library of TE families was subjected to extensive manual curation, as was previously done for Av-ref¹⁶, and used as database for searching and annotating TE copies in the AvL1 assembly with RepeatMasker¹²¹. We used RMBlast (National Center for Biotechnology Information Blast modified for use with RepeatMasker) as search engine. Initial RepeatMasker output was filtered for copies covering less than 5% of reference TE length. The output was converted into gff3 format for subsequent analysis. TE annotation was intersected with gene prediction models to eliminate any duplication events spanning both databases and to obtain a list of TE-encoded genes for further analysis. For tandem repeat (TR) identification, AvL1 assembly was uploaded to Tandem Repeats Database¹²². We generated an initial set of TRs by analyzing the sequence of each contig using Tandem Repeats Finder¹²³ with default parameters (match=2, mismatch=7, indels=7, minimal alignment score=50). Further searches with modifications in the alignment score (size of the repeat unit) were performed, and manual correction was carried out when necessary.

Small RNA analysis. *A. vaga* sRNA-seq data (SRA accession no. SRP070765) for two wildtype small RNA replicas were mapped to Av-ref genome as described in⁴⁸. Heatmaps of sRNA-Seq data for genes, TEs, and DIP-seq and ChIP-seq peaks were generated with deepTools¹²⁴ for each annotation. Reads normalized to 1x sequencing depth (RPGC or reads per genomic content) were used for normalization in heatmaps.

Methylation data processing and visualization. For generation of heat maps and profile plots, the DeepTools ¹²⁴ computeMatrix, plotHeatmap and plotProfile scripts were used with specific parameters: RPGC normalization, bin size 10, effective genome size (Av 213837663 and AvL1 217117546), extendReads (IP-seq 50, ChIP-seq 75, sRNA-seq 50), interpolationMethod nearest. The annotatePeaks function from HOMER Tools ¹²⁵ was used to obtain methylation profiles of selected regions of interest, using different window and bin sizes (parameters given in figure legends). Overlapping values of different annotated features (DIP/ChIP-seq peak, base modification) were estimated with bedtools v2.27.1 ¹²⁶, whether they are intersecting (bedtools intersect) or after providing a specific size window (bedtools window). Genome-wide 4mC/6mA visual representations were generated using Circos ¹²⁷; Av-ref reads were plotted from two genomic Illumina libraries (SRP020364) with different insert size (450 and 862 bp); AvL1 reads were plotted from Illumina (SRR8134454) and PacBio (SRX6639068).

Collinearity analysis. Syntenic regions within and between genomes were identified using MCScanX ¹²⁸ after blastp all-versus-all (e-val = 1e-10, maximum number of target sequences = 5) of the protein annotations from both genomes (Av-ref and AvL1). We searched for collinear block regions with at least 3 homologous genes and 20 maximum gaps allowed. The Ks and Ka (synonymous and nonsynonymous substitution, respectively) values between pairs of collinear genes were calculated with the script add_kaks_to_MCScanX.pl (<https://zenodo.org/badge/latestdoi/92963110>). We also searched for collinearity breaks between adjacent homologous blocks, defined as regions where homologous blocks could not be aligned along scaffolds without some rearrangements.

Phylogenetic analyses. MTase homologs in bdelloids were identified by tblastn searches of GenBank WGS databases at NCBI, checked for the presence of metazoan genes in the vicinity, translated with validation of exon-intron structure, and used in blastp searches of REBASE ¹ to obtain MTases with known recognition sequences. Multiple sequence alignments were performed by MUSCLE v.3.8.31 ¹²⁹. Amino acid sequences were clustered by neighbor-joining, as MTases are not amenable to conventional phylogenetic analysis due to hypervariability of the target recognition domain, and the tree was visualized in MEGA ¹³⁰. MBD-containing bdelloid proteins were identified by profile HMM search ¹³¹ with the MBD query (PF01429). Av-ref SETDB1 homologs from Genoscope annotation were manually re-annotated to improve quality, and full-length proteins were used as queries in blastp searches of refseq_protein database at NCBI to obtain additional orthologs from 10 bdelloid species and representative protostome taxa. Maximum likelihood phylogenetic analysis was done with IQTREE v1.6.11 ¹³² using best-fitting model selection and 1,000 ultrafast bootstrap replicates. Ago/Piwi counts in AvL1 were done as in ⁴⁷.

Data and Code Availability: Sequences obtained in this study were deposited under BioProject PRJNA558051 (SRA accession Nos. SRR9886612, SRR9900832-45 for individual SMRT cells). The Avaga_MBL_L1 genome assembly was deposited under accession No. JAGENE000000000. The ChIP-seq, MeDIP-seq and RNA-seq datasets were deposited in GEO under accession Nos. GSE140049-52. All materials and (non-essential) custom scripts are freely available without restrictions.

Reporting Summary: Further information on research design is available in the Nature Research Reporting Summary linked to this article.

References

1. Roberts, R.J., Vincze, T., Posfai, J. & Macelis, D. REBASE--a database for DNA restriction and modification: enzymes, genes and genomes. *Nucleic Acids Res* **43**, D298-9 (2015).
2. Beaulaurier, J., Schadt, E.E. & Fang, G. Deciphering bacterial epigenomes using modern sequencing technologies. *Nat Rev Genet* **20**, 157-172 (2019).
3. Klose, R.J. & Bird, A.P. Genomic DNA methylation: the mark and its mediators. *Trends Biochem Sci* **31**, 89-97 (2006).
4. Edwards, J.R., Yarychkivska, O., Boulard, M. & Bestor, T.H. DNA methylation and DNA methyltransferases. *Epigenetics Chromatin* **10**, 23 (2017).
5. Greenberg, M.V.C. & Bourc'his, D. The diverse roles of DNA methylation in mammalian development and disease. *Nat Rev Mol Cell Biol* **20**, 590-607 (2019).
6. Iyer, L.M., Abhiman, S. & Aravind, L. Natural history of eukaryotic DNA methylation systems. *Prog Mol Biol Transl Sci* **101**, 25-104 (2011).
7. Goll, M.G. et al. Methylation of tRNA^{Asp} by the DNA methyltransferase homolog Dnmt2. *Science* **311**, 395-8 (2006).
8. Raddatz, G. et al. Dnmt2-dependent methylomes lack defined DNA methylation patterns. *Proc Natl Acad Sci U S A* **110**, 8627-31 (2013).
9. Heyn, H. & Esteller, M. An Adenine Code for DNA: A Second Life for N6-Methyladenine. *Cell* **161**, 710-3 (2015).
10. Luo, G.Z., Blanco, M.A., Greer, E.L., He, C. & Shi, Y. DNA N(6)-methyladenine: a new epigenetic mark in eukaryotes? *Nat Rev Mol Cell Biol* **16**, 705-10 (2015).
11. Luo, G.Z. & He, C. DNA N(6)-methyladenine in metazoans: functional epigenetic mark or bystander? *Nat Struct Mol Biol* **24**, 503-506 (2017).
12. Iyer, L.M., Zhang, D. & Aravind, L. Adenine methylation in eukaryotes: Apprehending the complex evolutionary history and functional potential of an epigenetic modification. *Bioessays* **38**, 27-40 (2016).
13. O'Brown, Z.K. et al. Sources of artifact in measurements of 6mA and 4mC abundance in eukaryotic genomic DNA. *BMC Genomics* **20**, 445 (2019).
14. Mark Welch, D.B., Mark Welch, J.L. & Meselson, M. Evidence for degenerate tetraploidy in bdelloid rotifers. *Proc Natl Acad Sci U S A* **105**, 5145-9 (2008).
15. Gladyshev, E.A., Meselson, M. & Arkhipova, I.R. Massive horizontal gene transfer in bdelloid rotifers. *Science* **320**, 1210-3 (2008).
16. Flot, J.F. et al. Genomic evidence for ameiotic evolution in the bdelloid rotifer *Adineta vaga*. *Nature* **500**, 453-7 (2013).
17. Boschetti, C. et al. Biochemical diversification through foreign gene expression in bdelloid rotifers. *PLoS Genet* **8**, e1003035 (2012).
18. Nowell, R.W. et al. Comparative genomics of bdelloid rotifers: Insights from desiccating and nondesiccating species. *PLOS Biology* **16**, e2004830 (2018).
19. Luo, G.Z. et al. N(6)-methyldeoxyadenosine directs nucleosome positioning in *Tetrahymena* DNA. *Genome Biol* **19**, 200 (2018).
20. Xiao, C.-L. et al. N6-methyladenine DNA modification in the human genome. *Mol Cell* **71**, 306-18 (2018).
21. Greer, E.L. et al. DNA methylation on N6-adenine in *C. elegans*. *Cell* **161**, 868-78 (2015).
22. Kweon, S.-M. et al. An adversarial DNA N6-methyladenine-sensor network preserves Polycomb silencing. *Mol Cell* **74**, 1138-47 (2019).
23. Malone, T., Blumenthal, R.M. & Cheng, X. Structure-guided analysis reveals nine sequence motifs conserved among DNA amino-methyltransferases, and suggests a catalytic mechanism for these enzymes. *J Mol Biol* **253**, 618-32 (1995).
24. Chen, H. et al. *Phytophthora* methylomes are modulated by 6mA methyltransferases and associated with adaptive genome regions. *Genome Biol* **19**, 181 (2018).
25. Li, W., Shi, Y., Zhang, T., Ye, J. & Ding, J. Structural insight into human N6amt1-Trm112 complex functioning as a protein methyltransferase. *Cell Discov* **5**, 51 (2019).
26. Woodcock, C.B., Yu, D., Zhang, X. & Cheng, X. Human HemK2/KMT9/N6AMT1 is an active protein methyltransferase, but does not act on DNA in vitro, in the presence of Trm112. *Cell Discov* **5**, 50 (2019).

27. Hamey, J.J. et al. Novel N-terminal and Lysine Methyltransferases That Target Translation Elongation Factor 1A in Yeast and Human. *Mol Cell Proteomics* **15**, 164-76 (2016).
28. Kusevic, D., Kudithipudi, S. & Jeltsch, A. Substrate Specificity of the HEMK2 Protein Glutamine Methyltransferase and Identification of Novel Substrates. *J Biol Chem* **291**, 6124-33 (2016).
29. Xie, Q. et al. N6-methyladenine DNA modification in glioblastoma. *Cell* **175**, 1228-1243.e20 (2018).
30. Blow, M.J. et al. The epigenomic landscape of prokaryotes. *PLOS Genetics* **12**, e1005854 (2016).
31. Timinskas, A., Butkus, V. & Janulaitis, A. Sequence motifs characteristic for DNA [cytosine-N4] and DNA [adenine-N6] methyltransferases. Classification of all DNA methyltransferases. *Gene* **157**, 3-11 (1995).
32. Yap, K.L. & Zhou, M.M. Structure and mechanisms of lysine methylation recognition by the chromodomain in gene transcription. *Biochemistry* **50**, 1966-80 (2011).
33. Vakhrusheva, O.A. et al. Genomic signatures of recombination in a natural population of the bdelloid rotifer *Adineta vaga*. *Nature Communications* **11**, 6421 (2020).
34. Sukackaite, R., Grazulis, S., Tamulaitis, G. & Siksnys, V. The recognition domain of the methyl-specific endonuclease McrBC flips out 5-methylcytosine. *Nucl Acids Res* **40**, 7552-7562 (2012).
35. Sutherland, E., Coe, L. & Raleigh, E.A. McrBC: a multisubunit GTP-dependent restriction endonuclease. *J Mol Biol* **225**, 327-48 (1992).
36. Lentini, A. et al. A reassessment of DNA-immunoprecipitation-based genomic profiling. *Nat Methods* **15**, 499-504 (2018).
37. Clark, T.A. et al. Characterization of DNA methyltransferase specificities using single-molecule, real-time DNA sequencing. *Nucl Acids Res* **40**, e29 (2012).
38. Xu, X. et al. Evolutionary transition between invertebrates and vertebrates via methylation reprogramming in embryogenesis. *National Science Review* **6**, 993-1003 (2019).
39. Arkhipova, I.R., Yushenova, I.A. & Rodriguez, F. Giant reverse transcriptase-encoding transposable elements at telomeres. *Mol Biol Evol* **34**, 2245-57 (2017).
40. Hodges, E. et al. High definition profiling of mammalian DNA methylation by array capture and single molecule bisulfite sequencing. *Genome Res* **19**, 1593-605 (2009).
41. Bujnicki, J.M. & Radlinska, M. Molecular evolution of DNA-(cytosine-N4) methyltransferases: evidence for their polyphyletic origin. *Nucleic Acids Res* **27**, 4501-9 (1999).
42. Mozzetta, C., Boyarchuk, E., Pontis, J. & Ait-Si-Ali, S. Sound of silence: the properties and functions of repressive Lys methyltransferases. *Nat Rev Mol Cell Biol* **16**, 499-513 (2015).
43. Ahringer, J. & Gasser, S.M. Repressive chromatin in *Caenorhabditis elegans*: Establishment, composition, and function. *Genetics* **208**, 491-511 (2018).
44. Frapporti, A. et al. The Polycomb protein Ezh1 mediates H3K9 and H3K27 methylation to repress transposable elements in *Paramecium*. *Nat Commun* **10**, 2710 (2019).
45. Zhao, X. et al. RNAi-dependent Polycomb repression controls transposable elements in *Tetrahymena*. *Genes Dev* **33**, 348-364 (2019).
46. Howe, F.S., Fischl, H., Murray, S.C. & Mellor, J. Is H3K4me3 instructive for transcription activation? *Bioessays* **39**, 1-12 (2017).
47. Nowell, R.W. et al. Evolutionary dynamics of transposable elements in bdelloid rotifers. *eLife* **10**, e63194 (2021).
48. Rodriguez, F. & Arkhipova, I.R. Multitasking of the piRNA silencing machinery: Targeting transposable elements and foreign genes in the bdelloid rotifer *Adineta vaga*. *Genetics* **203**, 255-68 (2016).
49. Arita, K., Ariyoshi, M., Tochio, H., Nakamura, Y. & Shirakawa, M. Recognition of hemi-methylated DNA by the SRA protein UHRF1 by a base-flipping mechanism. *Nature* **455**, 818-21 (2008).
50. Anosova, I. et al. A novel RNA binding surface of the TAM domain of TIP5/BAZ2A mediates epigenetic regulation of rRNA genes. *Nucleic Acids Res* **43**, 5208-20 (2015).
51. Diagouraga, B. et al. PRDM9 Methyltransferase Activity Is Essential for Meiotic DNA Double-Strand Break Formation at Its Binding Sites. *Mol Cell* **69**, 853-865.e6 (2018).
52. Engert, C.G., Droste, R., van Oudenaarden, A. & Horvitz, H.R. A *Caenorhabditis elegans* protein with a PRDM9-like SET domain localizes to chromatin-associated foci and promotes spermatocyte gene expression, sperm production and fertility. *PLoS Genet* **14**, e1007295 (2018).
53. Allis, C.D. & Jenuwein, T. The molecular hallmarks of epigenetic control. *Nat Rev Genet* **17**, 487-500 (2016).

- 1588 54. Epsztejn-Litman, S. et al. De novo DNA methylation promoted by G9a prevents reprogramming of
1589 embryonically silenced genes. *Nat Struct Mol Biol* **15**, 1176-1183 (2008).
- 1590 55. Leung, D.C. et al. Lysine methyltransferase G9a is required for de novo DNA methylation and the
1591 establishment, but not the maintenance, of proviral silencing. *Proc Natl Acad Sci U S A* **108**,
1592 5718-23 (2011).
- 1593 56. Ohki, I. et al. Solution structure of the methyl-CpG binding domain of human MBD1 in complex
1594 with methylated DNA. *Cell* **105**, 487-97 (2001).
- 1595 57. Chandler, S.P., Guschin, D., Landsberger, N. & Wolffe, A.P. The methyl-CpG binding
1596 transcriptional repressor MeCP2 stably associates with nucleosomal DNA. *Biochemistry* **38**,
1597 7008-18 (1999).
- 1598 58. Schiffers, S. et al. Quantitative LC-MS provides no evidence for m6dA or m4dC in the genome of
1599 mouse embryonic stem cells and tissues. *Angew Chem Int Ed Engl* **56**, 11268-11271 (2017).
- 1600 59. Zhu, S. et al. Mapping and characterizing N6-methyladenine in eukaryotic genomes using single-
1601 molecule real-time sequencing. *Genome Res* **28**, 1067-1078 (2018).
- 1602 60. Schubeler, D. Function and information content of DNA methylation. *Nature* **517**, 321-6 (2015).
- 1603 61. Du, J., Johnson, L.M., Jacobsen, S.E. & Patel, D.J. DNA methylation pathways and their
1604 crosstalk with histone methylation. *Nat Rev Mol Cell Biol* **16**, 519-32 (2015).
- 1605 62. Le Thomas, A. et al. Transgenerationally inherited piRNAs trigger piRNA biogenesis by changing
1606 the chromatin of piRNA clusters and inducing precursor processing. *Genes Dev* **28**, 1667-80
1607 (2014).
- 1608 63. Jih, G. et al. Unique roles for histone H3K9me states in RNAi and heritable silencing of
1609 transcription. *Nature* **547**, 463-467 (2017).
- 1610 64. Batki, J. et al. The nascent RNA binding complex SFINX licenses piRNA-guided heterochromatin
1611 formation. *Nat Struct Mol Biol* **26**, 720-731 (2019).
- 1612 65. Zoch, A. et al. SPOCD1 is an essential executor of piRNA-directed de novo DNA methylation.
1613 *Nature* **584**, 635-639 (2020).
- 1614 66. Yu, R., Wang, X. & Moazed, D. Epigenetic inheritance mediated by coupling of RNAi and histone
1615 H3K9 methylation. *Nature* **558**, 615-619 (2018).
- 1616 67. Blommaert, J., Riss, S., Hecox-Lea, B., Mark Welch, D.B. & Stelzer, C.P. Small, but surprisingly
1617 repetitive genomes: transposon expansion and not polyploidy has driven a doubling in genome
1618 size in a metazoan species complex. *BMC Genomics* **20**, 466 (2019).
- 1619 68. Park, M., Patel, N., Keung, A.J. & Khalil, A.S. Engineering Epigenetic Regulation Using Synthetic
1620 Read-Write Modules. *Cell* **176**, 227-238.e20 (2019).
- 1621 69. Jain, R., Rivera, M.C. & Lake, J.A. Horizontal gene transfer among genomes: the complexity
1622 hypothesis. *Proc Natl Acad Sci U S A* **96**, 3801-6 (1999).
- 1623 70. Ku, C. et al. Endosymbiotic origin and differential loss of eukaryotic genes. *Nature* **524**, 427-32
1624 (2015).
- 1625 71. Wang, Y., Chen, X., Sheng, Y., Liu, Y. & Gao, S. N6-adenine DNA methylation is associated with
1626 the linker DNA of H2A.Z-containing well-positioned nucleosomes in Pol II-transcribed genes in
1627 Tetrahymena. *Nucleic Acids Res* **45**, 11594-606 (2017).
- 1628 72. Beh, L.Y. et al. Identification of a DNA N6-adenine methyltransferase complex and its impact on
1629 chromatin organization. *Cell* **177**, 1781-96 (2019).
- 1630 73. Fu, Y. et al. N6-methyldeoxyadenosine marks active transcription start sites in Chlamydomonas.
1631 *Cell* **161**, 879-92 (2015).
- 1632 74. Liang, Z. et al. DNA N(6)-adenine methylation in Arabidopsis thaliana. *Dev Cell* **45**, 406-416.e3
1633 (2018).
- 1634 75. Zhou, C. et al. Identification and analysis of adenine N(6)-methylation sites in the rice genome.
1635 *Nat Plants* **4**, 554-563 (2018).
- 1636 76. Zhang, Q. et al. N(6)-methyladenine DNA methylation in japonica and indica rice genomes and its
1637 association with gene expression, plant development, and stress responses. *Mol Plant* **11**, 1492-
1638 1508 (2018).
- 1639 77. Mondo, S.J. et al. Widespread adenine N6-methylation of active genes in fungi. *Nat Genet* **49**,
1640 964-8 (2017).
- 1641 78. Dabe, E.C., Sanford, R.S., Kohn, A.B., Bobkova, Y. & Moroz, L.L. DNA methylation in basal
1642 metazoans: Insights from ctenophores. *Integr Comp Biol* **55**, 1096-110 (2015).
- 1643 79. Zhang, G. et al. N6-methyladenine DNA modification in Drosophila. *Cell* **161**, 893-906 (2015).

1644 80. Falckenhayn, C. et al. Comprehensive DNA methylation analysis of the *Aedes aegypti* genome.
1645 *Sci Rep* **6**, 36444 (2016).

1646 81. Wang, X. et al. DNA methylation on N6-adenine in lepidopteran *Bombyx mori*. *Biochim Biophys*
1647 *Acta Gene Regul Mech* **1861**, 815-25 (2018).

1648 82. Liu, J. et al. Abundant DNA 6mA methylation during early embryogenesis of zebrafish and pig.
1649 *Nat Commun* **7**, 13052 (2016).

1650 83. Koziol, M.J. et al. Identification of methylated deoxyadenosines in vertebrates reveals diversity in
1651 DNA modifications. *Nat Struct Mol Biol* **23**, 24-30 (2016).

1652 84. Wu, T.P. et al. DNA methylation on N(6)-adenine in mammalian embryonic stem cells. *Nature*
1653 **532**, 329-33 (2016).

1654 85. Yao, B. et al. DNA N6-methyladenine is dynamically regulated in the mouse brain following
1655 environmental stress. *Nat Commun* **8**, 1122 (2017).

1656 86. Kigar, S.L. et al. N(6)-methyladenine is an epigenetic marker of mammalian early life stress. *Sci*
1657 *Rep* **7**, 18078 (2017).

1658 87. Pacini, C.E., Bradshaw, C.R., Garrett, N.J. & Koziol, M.J. Characteristics and homogeneity of N6-
1659 methylation in human genomes. *Sci Rep* **9**, 5185 (2019).

1660 88. Franch-Gras, L. et al. Genomic signatures of local adaptation to the degree of environmental
1661 predictability in rotifers. *Sci Rep* **8**, 16051 (2018).

1662 89. Kim, H.S. et al. The genome of the freshwater monogonont rotifer *Brachionus calyciflorus*. *Mol*
1663 *Ecol Resour* **18**, 646-55 (2018).

1664 90. Park, J.C. et al. The genome of the marine rotifer *Brachionus koreanus* sheds light on the
1665 antioxidative defense system in response to 2-ethyl-phenanthrene and piperonyl butoxide. *Aquat*
1666 *Toxicol* **221**, 105443 (2020).

1667 91. Kang, H.M. et al. The genome of the marine monogonont rotifer *Brachionus rotundiformis* and
1668 insight into species-specific detoxification components in *Brachionus* spp. *Comp Biochem Physiol*
1669 *Part D Genomics Proteomics* **36**, 100714 (2020).

1670 92. Favorov, A. et al. Exploring massive, genome scale datasets with the GenometriCorr package.
1671 *PLoS Comput Biol* **8**, e1002529 (2012).

1672 93. Li, Z. et al. N6-methyladenine in DNA antagonizes SATB1 in early development. *Nature* (2020).

1673 94. Williams, B.P. & Gehring, M. Stable transgenerational epigenetic inheritance requires a DNA
1674 methylation-sensing circuit. *Nat Commun* **8**, 2124 (2017).

1675 95. Jurkowski, T.P. & Jeltsch, A. On the evolutionary origin of eukaryotic DNA methyltransferases
1676 and Dnmt2. *PLoS One* **6**, e28104 (2011).

1677 96. Bewick, A.J. et al. Diversity of cytosine methylation across the fungal tree of life. *Nat Ecol Evol* **3**,
1678 479-490 (2019).

1679 97. Murphy, J., Mahony, J., Ainsworth, S., Nauta, A. & van Sinderen, D. Bacteriophage orphan DNA
1680 methyltransferases: insights from their bacterial origin, function, and occurrence. *Appl Environ*
1681 *Microbiol* **79**, 7547-55 (2013).

1682 98. Padeken, J. et al. Synergistic lethality between BRCA1 and H3K9me2 loss reflects satellite
1683 derepression. *Genes Dev* **33**, 436-451 (2019).

1684 99. Jurkowska, R.Z. et al. H3K14ac is linked to methylation of H3K9 by the triple Tudor domain of
1685 SETDB1. *Nat Commun* **8**, 2057 (2017).

1686 100. Yushenova, I.A. & Arkhipova, I.R. Biochemical properties of bacterial reverse transcriptase-
1687 related (rvt) gene products: multimerization, protein priming, and nucleotide preference. *Curr*
1688 *Genet* **64**, 1287-1301 (2018).

1689 101. Sambrook, J. & Russell, D.W. *Molecular cloning: A laboratory manual*, (Cold Spring Harbor
1690 Laboratory Press, Cold Spring Harbor, NY, 2001).

1691 102. Kong, H. et al. Functional analysis of putative restriction-modification system genes in the
1692 *Helicobacter pylori* J99 genome. *Nucl Acids Res* **28**, 3216-23 (2000).

1693 103. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *2011*
1694 **17**, 3 (2011).

1695 104. Langmead, B., Trapnell, C., Pop, M. & Salzberg, S.L. Ultrafast and memory-efficient alignment of
1696 short DNA sequences to the human genome. *Genome Biol.* **10**(2009).

1697 105. Zhang, Y. et al. Model-based analysis of ChIP-Seq (MACS). *Genome Biol* **9**, R137 (2008).

1698 106. English, A.C. et al. Mind the gap: upgrading genomes with Pacific Biosciences RS long-read
1699 sequencing technology. *PLoS One* **7**, e47768 (2012).

107. Delmont, T.O. & Eren, A.M. Identifying contamination with advanced visualization and analysis practices: metagenomic approaches for eukaryotic genome assemblies. *PeerJ* **4**, e1839 (2016).
108. Flusberg, B.A. et al. Direct detection of DNA methylation during single-molecule, real-time sequencing. *Nat Methods* **7**, 461-5 (2010).
109. Chaisson, M.J. & Tesler, G. Mapping single molecule sequencing reads using basic local alignment with successive refinement (BLASR): application and theory. *BMC Bioinformatics* **13**, 238 (2012).
110. Usai, G. et al. DNA Modification Patterns within the Transposable Elements of the Fig (*Ficus carica* L.) Genome. *Plants (Basel)* **10**(2021).
111. Ma, W., Noble, W.S. & Bailey, T.L. Motif-based analysis of large nucleotide data sets using MEME-ChIP. *Nat Protoc* **9**, 1428-50 (2014).
112. Askjaer, P., Ercan, S. & Meister, P. Modern techniques for the analysis of chromatin and nuclear organization in *C. elegans*. *WormBook*, 1-35 (2014).
113. Langmead, B. & Salzberg, S.L. Fast gapped-read alignment with Bowtie 2. *Nat Meth* **9**, 357-359 (2012).
114. Kim, D. et al. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol* **14**, R36 (2013).
115. Anders, S., Pyl, P.T. & Huber, W. HTSeq—a Python framework to work with high-throughput sequencing data. *Bioinformatics* **31**, 166-169 (2015).
116. Trapnell, C. et al. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat Protoc* **7**, 562-78 (2012).
117. Hoff, K.J., Lomsadze, A., Borodovsky, M. & Stanke, M. Whole-genome annotation with BRAKER. *Methods Mol Biol* **1962**, 65-95 (2019).
118. Lomsadze, A., Burns, P.D. & Borodovsky, M. Integration of mapped RNA-Seq reads into automatic training of eukaryotic gene finding algorithm. *Nucleic Acids Res* **42**, e119 (2014).
119. Stanke, M., Schoffmann, O., Morgenstern, B. & Waack, S. Gene prediction in eukaryotes with a generalized hidden Markov model that uses hints from external sources. *BMC Bioinformatics* **7**, 62 (2006).
120. Flutre, T., Duprat, E., Feuillet, C. & Quesneville, H. Considering transposable element diversification in de novo annotation approaches. *PLoS One* **6**, e16526 (2011).
121. Smit, A.F.A., Hubley, R. & Green, P. RepeatMasker Open-4.0. 2013-2015
<http://www.repeatmasker.org>. (2015).
122. Gelfand, Y., Rodriguez, A. & Benson, G. TRDB--the Tandem Repeats Database. *Nucleic Acids Res* **35**, D80-7 (2007).
123. Benson, G. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res* **27**, 573-80 (1999).
124. Ramirez, F. et al. deepTools2: a next generation web server for deep-sequencing data analysis. *Nucleic Acids Res* **44**, W160-5 (2016).
125. Heinz, S. et al. Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Mol Cell* **38**, 576-89 (2010).
126. Quinlan, A.R. & Hall, I.M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841-2 (2010).
127. Krzywinski, M. et al. Circos: an information aesthetic for comparative genomics. *Genome Res* **19**, 1639-45 (2009).
128. Wang, Y. et al. MCScanX: a toolkit for detection and evolutionary analysis of gene synteny and collinearity. *Nucleic Acids Res* **40**, e49 (2012).
129. Edgar, R.C. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics* **5**, 113 (2004).
130. Kumar, S., Stecher, G. & Tamura, K. MEGA7: Molecular Evolutionary Genetics Analysis version 7.0 for bigger datasets. *Mol Biol Evol* **33**, 1870-1874 (2016).
131. Eddy, S.R. Accelerated Profile HMM Searches. *PLoS Comput Biol* **7**, e1002195 (2011).
132. Trifinopoulos, J., Nguyen, L.-T., von Haeseler, A. & Minh, B.Q. W-IQ-TREE: a fast online phylogenetic tool for maximum likelihood analysis. *Nucleic Acids Research* **44**, W232-W235 (2016).

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [nreditorialpolicychecklist.pdf](#)
- [nrreportingsummary.pdf](#)
- [accesstosequencingdata.docx](#)
- [SupplementaryTableS11.xlsx](#)
- [SupplementaryDataFilesS1S4.xlsx](#)