Science
AAAS

Supplementary Materials for

Critical assessment of DNA adenine methylation in eukaryotes using quantitative deconvolution

Yimeng Kong, Lei Cao, Gintaras Deikus, Yu Fan, Edward A. Mead, Weiyi Lai, Yizhou Zhang, Raymund Yong, Robert Sebra, Hailin Wang, Xue-Song Zhang and Gang Fang*

Correspondence to: gang.fang@mssm.edu

This PDF file includes:

Other Supplementary Material for this manuscript includes the following:

Materials and Methods

E. coli K12 MG1655 and ER3413

E. coli K12 MG1655 was obtained from American Type Culture Collection (ATCC, cat#47076, Manassas, VA, USA). The methyltransferase-deficient *E.coli* strain ER3413 was obtained from the Yale Coli Genetic Stock Center (Yale CGSC, cat#14167, New Haven, CT, USA) and both *E. coli* strains were grown on Miller Luria Broth (LB) and LB agar (Thermo Fisher Scientific, cat#BP1426-500, and cat#BP1425-500, respectively, Waltham, MA, USA). The gDNA of both *E. coli* strains was extracted with DNeasy Blood & Tissue Kits (Qiagen, Hilden, Germany) by the manufacturer's protocol.

C. reinhardtii

C. reinhardtii wild type strain CC-125 mt+ was obtained from Dr. Peter Hegemann's lab at the Humboldt University of Berlin (www.chlamy.de) and grown in 50 ml of standard Tris-Acetate-Phosphate (TAP) medium (Thermo Fisher Scientific, cat#A1379801) in 250 ml flasks at 25 ºC with 220 rpm shaking and constant light. The genomic DNA of *C. reinhardtii* was extracted with NucleoSpin Plant II kit (MACHEREY-NAGEL, Düren, Germany) according to the manufacturer's protocol.

T. thermophila

T. thermophila wild-type strain CU428 gDNA was obtained from the national Tetrahymena Stock Center at Cornell University (Cornell TSC, Ithaca, NY, USA).

D. melanogaster

Fly stocks were obtained from Dr. Bing Yao lab at Emory University and maintained under standard culture conditions. The w^{1118} strain was used for collect embryos and adults. For optimal egg collection, healthy young male and female flies were crossed and maintained in tubes with new food. Prior to egg collection, files were fed with yeast paste on fresh grape juice agar plate for at least 2 hours. The 45 minutes embryos were then collected on grape juice agar plate and rinsed with water on a fine nylon mesh filter. Genomic DNA for both samples were extracted as described previously (*42*). Briefly, to isolate the gDNA, lysis buffer containing 100mM Tris-HCl (pH 8.0), 5mM EDTA (pH 8.0), 0.2% SDS, 200mM NaCl, and 20-50 µL of protease K (20mg/ml) was added to the fly samples, homogenized and incubated at 55ºC overnight. After the overnight digestion, the lysates were brought to room temperature, and incubated with 5-10 µL of RNase A solution (20mg/ml) for at least 1 hour at 37 ºC. DNA was extracted by adding equal volume of phenol: chloroform: isoamyl alcohol at a ratio of 25:24:1 and centrifuging at 12,000 rpm for 5 min. Supernatant was transferred to clean tubes and mixed with 10% of 3M NaOAC and two times of 100% Ethanol at -80 ºC for at least 30 min. The gDNA pellets were collected by 12,000 rpm centrifugation for 15 min. After washing the gDNA pellet with 70% ethanol, the pellets were dissolved in 30-50 µL of nuclease-free water.

A. thaliana

A. thaliana Col-0 was grown as previously described (*43*). Briefly, *A. thaliana* seeds were surface sterilized with 75% ethanol, plated on half strength Murashige and Skoog (MS) agar medium at 4°C in the dark for 3 days until uniform germination, and then were germinated on at 22°C in constant light for 10 days. The seedlings with about similar size were transferred to potting mix for growth in 22°C growth chambers with 16-h light/8-h dark cycles. The plant materials of whole seedlings and roots were collected at day 21 after initiation of germination. Genomic DNA was extracted using a DNeasy Plant Mini kit (Qiagen) following manufacturers' instructions.

HEK293

Human Embryonic Kidney (HEK) 293 cells were obtained from ATCC (cat#CRL-1573). All media and tissue culture supplies were obtained through ThermoFisher Scientific. The cell growth protocol was based upon Bulanenkova *et al.* (*44*). Cells were grown under sterile conditions in a 2:1:1 mixture of RPMI-1640 and DMEM:F12 and (supplemented at 10% with Fetal Bovine Serum (FBS)), splitting as needed based on confluence (~75%) and on media color. Cells were grown as an adherent culture in Nalgene Nunc tissue culture-treated flasks in a 37 °C/5% CO_2 incubator. Cells screened as negative for mycoplasma during culture, by testing with a MycoAlert Mycoplasma Detection Kit (Lonza, Inc, Morristown, NJ, USA). Cells were harvested at <25 passages with 0.25% trypsin, and gDNA was isolated using a DNeasy Blood & Tissue Kits (Qiagen) by the manufacturer's protocol.

HEK293 Whole Genome Amplification DNA (HEK-WGA)

HEK293 WGA was performed with the REPLI-g mini kit (Qiagen, cat#150023), following the manufacturer's instructions. Briefly, 5ng of genomic DNA was used in the reaction and incubated at 30 °C for 16 hours, then heat-inactivated at 65 °C for 3min. All WGA samples were purified using 0.8X Ampure XP beads (Beckman Coulter, cat#A63881, Indianapolis, IN, USA). The DNA concentration of the product was determined with the Qubit dsDNA High Sensitivity Assay kit (Life Technologies, cat#Q32854, Carlsbad, CA, USA).

Dam-treated HEK293 DNA (HEK-dam)

One microgram of HEK293 WGA DNA was incubated at 37 °C for 4 hours in a 50ul-reaction containing 16U of Dam methyltransferase (Dam; New England Biolabs (NEB), cat#M0222S, Ipswich, MA, USA) and 1mM S-adenosylmethionine (SAM), followed by inactivation of the enzyme at 65 °C for 15 min.

M.SssI treated HEK-WGA (HEK-WGA-MsssI)

HEK293 WGA was performed as described above. To add the cytosine modification on CpG sites, 2ug of HEK293 WGA was subjected to M.SssI methyltransferase (Zymo Research, cat# E2010, Irvine, USA) treatment based upon the standard Zymo methylation protocol, but with increased SAM (16mM concentration), increased enzyme concentration (8U) and longer incubation time (18h).

TaqI/EcoRI/Dam-treated HEK293, HEK-WGA and HEK-WGA-MsssI

To get methylated TCGA, GAAGTTC, and GATC motifs on the HEK293 genome (HEK293-3M) or HEK-WGA (HEK-WGA-3M), 2ug of HEK293 native DNA or HEK-WGA DNA was first incubated for 4h at 65°C with 20U TaqI methyltransferase (methylating TCGA, NEB cat#M0219S). The purified DNA was further incubated with 80U EcoRI methyltransferase (methylating GAATTC, NEB cat#M0211S) at 37°C for 4h and finally with 16U Dam (methylating GATC) at 37 °C for 4h. All methylation reactions were carried out in the presence of 1mM SAM. DNA samples were purified using 1.0X Ampure XP beads (Beckman Coulter cat# A63881). To get HEK-WGA-MsssI-3M, HEK-WGA-3M was further treated with M.SssI methyltransferase following the protocol described above.

H. pylori

The *H. pylori* strain J99 was grown on Brucella agar plates (Difco Laboratories, Detroit, MI, cat# 211086) supplemented with 10% newborn calf serum (Serologicals Corp., Norcross, GA) at 37°C in a 5% CO_2 atmosphere. Genomic DNA was extracted from fresh *H. pylori* cells harvested from 24h cultures using the Wizard genomic DNA purification kit (Promega Corp, Madison, WI cat#A1620) (*45*).

S. cerevisiae

S. cerevisiae strain JEL1 was a kind gift from Dr. Hiten D. Madhani. A single colony was picked for starter culture at 30°C in 100ml YPGL medium, then used to inoculate into 2L YPGL medium to a starting OD600 of 0.03. The cultured cells were grown at 30°C to an OD600 of 1.0, then pelleted. Cells were lysed in a ball mill (6x 3 min at 15 Hz), followed by genomic DNA extraction with a DNeasy Blood & Tissue Kit (Qiagen cat# 69506) by the manufacturer's protocol.

H. vesiculosa

H. vesiculosa NRRL 3301 was obtained from Agricultural Research Service Culture Collection (NRRL, Peoria, IL, USA) and grown in Potato Dextrose Agar medium (NRRL Medium No. 3) The gDNA was isolated with CTAB protocol as previously described (*46*).

E. coli/ *H. pylori*/ *S. cerevisiae* dilution

1ug *E. coli* K12 MG1655 gDNA and 1ug *H. pylori* J99 gDNA were mixed and then the mixture was diluted with *S. cerevisiae* gDNA with the ratios of gDNA amount at 1:2.16 (D1), 1:8.99 (D2), 1:30.55 (D3), 1:98.71 (D4), 1:314.09 (D5), 1:994.69 (D6), 1:3145.37 (D7), 1:9941.52 (D8), and 1:31417.38 (D9), which will create samples of expected 6mA levels from $10^{-2.0}$ to $10^{-6.0}$, with $10^{0.5}$ as an interval. D9 result was not shown because of the extremely low number of *E.coli* / *H. pylori* CCS reads.

Targeted mitochondrial genome amplification

Mitochondrial DNA (mtDNA) was selectively amplified from a HEK 293 DNA sample using the REPLI-g Mitochondrial DNA Kit (Qiagen, cat#151023) according to the manufacturer's protocol. qRT-PCR analysis (not shown) indicated that mtDNA enrichment was 1,011-fold over native DNA.

Glioblastomas

Two glioblastoma tissues were obtained from surgical resection of lesions from patients with pathological diagnosis from the Mount Sinai Hospital. Informed consent from both patients was obtained in accordance with an IRB-approved protocol (HS# 10-00135). Glioblastoma-1 was derived from the anterior hippocampal lesion of an 85-year old female patient and glioblastoma-2 was derived from the nodular inferolateral component of a 58-year old male patient. Genomic DNA from both tissues was extracted using a DNeasy Kit (Qiagen) following manufacturers' instructions.

PBMC

Cryopreserved PBMC ($15x10^6$ cells) were obtained from ZenBio (lot# PBMC041019ABCD; Research Triangle Park, NC, USA). This sample was a pool of PBMC cells isolated from 4 individuals, which helps to reduce lot-to-lot variation. Genomic DNA was extracted using a DNeasy Blood and Tissue Kit (Qiagen) following manufacturers' instructions.

HEK 293 with transfection of an empty pCI plasmid (HEK-pCI)

HEK 293 cells were maintained by the procedures of Bulanenkova *et al*. (*44*). T25 flasks with approximately 10^6 HEK 293 cells were transfected using Lipofectamine 3000 (Invitrogen, cat#L3000015, Carlsbad, CA, USA) according to manufacturer's recommendations. Briefly, immediately before transfection, media was changed on the flask. Lipofectamine:DNA complexes for the pCI null vector (Promega, cat#E1731, Madison, WI, USA) were added to the media with 6.5ug DNA administered. Lipofectamine:DNA complexes were prepared in OptiMEM medium (Gibco, cat#31985062, Gaithersburg, MD, USA), then added to HEK growth media (RPMI 1640:DMEM:F-12 (2:1:1) medium supplemented with 10% fetal bovine serum [Gibco: RPMI-1640: cat#11875093; DMEM/F12: cat#11320033; FBS: cat#A3160402]) currently in the flask. Following transfection, flasks were returned to incubate at 37 ºC /5% CO_2. At 48 hrs post infection, the transfection medium was replaced with fresh HEK growth medium in each flask. At 72 hrs post infection, flasks were harvested by scraping cells in PBS, and DNA was isolated by a Wizard Genomic DNA Purification Kit (Promega, cat#A1120).

Library preparation and sequencing

The short-insert DNA SMRTbell library was performed according to the manufacturer's instructions by first shearing the sample to ~200bp using a Covaris LE220 Focused-ultrasonicator, followed by damage and end repair and ligation of single-stranded barcoded adaptors. Sequencing primer was then annealed to the SMRTbells and incubated with the

template for 1 hour at 20 ºC using a ratio of 20:1. The Sequel 2.1 polymerase was then bound to the annealed template complex using a ratio of 30:1 polymerase to SMRTbell at 0.5 nM for 4 h at 30 °C. The binding complex was cleaned up, diluted to the final loading concentration, and held at 4 °C until diffusion-based loading on the Sequel II instrument. Diluted binding complex at concentration of 125-175pM was loaded on the 8M sequencing chip and 30 hour movies were recorded according to the manufacturer's guidelines.

Raw reads pre-processing

The raw SMRT reads pre-processing was performed with SMRTlink v8.0 (Pacific Biosciences). Raw multiplexed files were first de-multiplexed with lima. For each DNA molecule, Circular Consensus reads (CCS reads) were generated using CCS software (v4.0.0) with *--min-rq 0.99* (*i.e. –minPredictedAccuracy 0.99* in older CCS version). Only CCS reads with passes ≥ 20 were kept for the following analysis. Subreads were hereafter mapped to the CCS read with pbalign (v0.4.1) with default parameters. IPD ratio and modification QV were calculated with ipdSummary (v2.4.1) from the mapping results. Ten bases at the head and at the tail of the CCS reads were trimmed off to ensure a correct *in silico* IPD prediction. Consistent results were observed between SMRTlink v7.0 and SMRTlink v8.0.

Deconvolution of CCS reads to different species of origin

With the reference-free method and single molecule analysis, 6mA was examined across all individual DNA molecules in a gDNA sample. Filtered CCS reads were then mapped to the reference genome of species of interest, by minimap2 (v2.17-r941) (*47*) with *–x sr*. Potential inter-species chimeric reads (reads with over 50bp at either end that do not have a match on the reference genome) were excluded for further analysis. After removing the mapped reads, the unmapped reads, if they exist, would hereafter be mapped to the NCBI-NT database using BLASTN (2.9.0+). The best match was then summarized by the contamination source (species or higher taxonomic ranks when CCS reads have more than one best match among very close species) and quantified separately. The 6mA contribution of each subgroup, *i*, was calculated by

$$\frac{L_i \times P_i}{\sum_{k=1}^{t}(L_k \times P_k)}$$

where *t* denotes number of subgroups identified, *L* denotes the 6mA/A level quantified by 6mASCOPE, *P* denotes the proportion of CCS reads mapped to the subgroup among the total.

Machine learning model for 6mA/A quantification by 6mASCOPE

Training data

All training data were constructed by mixing the positive and negative benchmark datasets. The gDNA of an experimental model human cell line HEK293 with extremely low 6mA/A level (0.79 ppm by UHPLC-MS/MS), was used as the negative control. The abundant 5mC events in native DNA may influence IPD-based 6mA detection as previously reported (*16*, *25*), so we designed three negative controls with different 5mC status for the model training: native HEK293 (with 70-80% natural 5mC at CpG sites), HEK-WGA (HEK293 subjected to whole

genome amplification, free of 5mC) and HEK-WGA-MsssI (HEK-WGA subjected M.SssI treatment, ~100% 5mC at CpG sites). These three negative control samples were each methylated *in vitro* using three sequence-specific bacterial 6mA methyltransferases (Dam: G<u>A</u>TC; TaqI: TCG<u>A</u>; and EcoRI: GA<u>A</u>TTC) to create three positive controls: HEK293-3M, HEK-WGA-3M, HEK-WGA-MsssI-3M, respectively. Please refer to the sample preparation section for a detailed description. To avoid the trace amount 6mA contamination from bacterial source enzymes, only human reads were used for both positive and negative controls. To exclude rare non-methylated sites within the 3 motifs in positive controls, only 6mAs with IPD ratio 2 were retained. Because the number of G<u>A</u>TC sites account for the majority of 6mAs within the 3 motifs, over-sampling was applied on the TCG<u>A</u> and GA<u>A</u>TTC motifs to balance the number of events within different motifs. For both negative and positive controls, adenines with extremely high IPD ratio (>7) were filtered out because they do not represent the kinetic signature of 6mA events in SMRT sequencing data (*14*, *25*). Finally, the positive 6mA sites for 3 positive controls includes 692,063 G<u>A</u>TC, 644,466 TCG<u>A</u>, and 695,856 GA<u>A</u>TTC sites in HEK293-3M; 1,295,986 G<u>A</u>TC, 1,247,655 TCG<u>A</u>, and 1,216,144 GA<u>A</u>TTC sites in HEK-WGA-3M, and 1,057,344 G<u>A</u>TC, 997,900 TCG<u>A</u>, and 991,904 GA<u>A</u>TTC sites in HEK-WGA-MsssI-3M. By random mixing the 6mA sites from positive controls and non-methylated sites from negative controls *in silico* at specific ratios, a total of 51 levels of 6mA abundance were constructed between $10^{-1.0}$ to $10^{-6.0}$, each with 100 million adenines in total for each matched negative-positive control. Training data of each 6mA/A level were constructed with 100 replicates for three pairs of negative-positive controls and combined for model training.

<u>QV distribution features and model training</u>

Modification quality value (QV) is a transformed score, -10*log(*p-value*), where *p-value* is the statistical significance that the observed IPD values on the subreads deviate from the *in silico* estimation based on the sequence context. QV captures both the magnitude of IPD ratio and the statistical confidence of 6mA analysis. Based on the observation that QV distribution varies across different 6mA/A levels, 6mASCOPE captures the variation of QV distribution to enable 6mA/A estimation. Coverage between 20 and 240 were divided into 11 bins. In total, 11 slopes, from 1.0 to 2.0 were selected to capture the proportion of linearly increased 6mA deviation from non-methylated adenines. Within coverage bin *i*, the ratio of adenines counts over slope *j* were calculated by adding one pseudo-count to ensure no ratio, $Z_{i,j}$, was zero, so that the log-transformed ratio $Z'_{i,j}$ matrix could be derived:

$$\begin{bmatrix} Z'_{1,1} & \cdots & Z'_{1,j} \\ \vdots & \ddots & \vdots \\ Z'_{i,1} & \cdots & Z'_{i,j} \end{bmatrix}$$

and reshaped to the vector (length V=121) of ratio features for each QV distribution. Random forest regression with 1000 random trees were trained with the feature vectors combining three negative-positive controls and further evaluated using independent-dataset-based validation.

<u>Leave-one-out cross validation</u>

Leave one-level out cross validation (LOOCV) was used to compare different machine learning models, including Linear Regression, Support Vector Classification (linear, poly and rbf kernel),

Decision Tree and Random Forest. Briefly, for each round of cross validation, one of the 51 levels, including all 300 repeated instances, were partitioned as the testing dataset while all other 50 levels were the training dataset. 51 rounds of LOOCV were performed using all 51 levels as the testing dataset iteratively.

Estimation of the 95% confidence intervals

The confidence intervals (CIs, representing a 95% probability that the calculated confidence interval encompasses the true value being estimated) of 6mA/A levels were first estimated for three negative-positive training datasets separately. To begin with, 6mA/A levels estimated by 6mASCOPE depend on the number of CCS reads (*i.e.* sequencing yield). To define the CIs across different numbers of CCS reads, we subsampled and mixed datasets from the methylated 6mA in positive and non-methylated adenine in negative controls at different ratios, creating datasets with 10 levels of CCS read numbers, 8K,10K, 20K, 50K, 80K, 100K, 200K, 500K, 800K, and 1M. In practice, the CI of a prediction by 6mASCOPE is inferred with the nearest CCS reads number. For simplicity, each CCS read contains 100 adenines. For each level of CCS read numbers, 51 levels of 6mA/A abundance were constructed between $10^{-1.0}$ to $10^{-6.0}$. For a specific level of CCS read number and a specific 6mA/A level, three parallel testing datasets were built to capture the possible variation between 6mA events from different motifs (G<u>A</u>TC, TCG<u>A</u>, or GA<u>A</u>TTC), each with 100 replicates.

For each of the 1,530 datasets (10x51x3), which represents a specific combination of CCS read number, 6mA/A abundance and 6mA motif, the probability density function, $Pr_{(x)}$, was captured based on the approximately normal distribution of 100 testing estimates by the random forest model. To obtain the CI for 6mA/A estimation *i* (log-transformed) with molecule number *j* and 6mA motif *k*, we calculated the probability of getting 6mA/A estimation *i* from different 6mA/A levels:

$$Pr_{j,k}[\mathrm{i}-0.05 \leq \mathrm{x} \leq \mathrm{i}+0.05]$$

where x$\pm$0.05 facilitates the estimation of the probability from the distribution.

The lower 95% CI is defined with a maximum of $m_{j,k}$ satisfying,

$$\frac{\sum_{t=1}^{m_{j,k}} Pr_{j,k,t}[\mathrm{i}-0.05 \leq \mathrm{x} \leq \mathrm{i}+0.05]}{\sum_{t=1}^{51} Pr_{j,k,t}[\mathrm{i}-0.05 \leq \mathrm{x} \leq \mathrm{i}+0.05]} \leq 2.5\%$$

and the upper 95% CI is defined with a minimum of $n_{j,k}$ satisfying:

$$\frac{\sum_{t=n_{j,k}}^{51} Pr_{j,k,t}[\mathrm{i}-0.05 \leq \mathrm{x} \leq \mathrm{i}+0.05]}{\sum_{t=1}^{51} Pr_{j,k,t}[\mathrm{i}-0.05 \leq \mathrm{x} \leq \mathrm{i}+0.05]} \leq 2.5\%$$

where *t* is the sorted 6mA/A levels (51 in total), with the same number of CCS reads and methylation motif, from low to high. To adjust for possible variation between different motifs, we integrated the three sets of 95% CIs for each motif and the final 95% CI represents the integration of 95% CI from three pairs of negative-positive training datasets.

Evaluation of CCS read accuracy

CCS reads for *E. coli* strains K12 MG1655 and ER3413 were generated using CCS software (v4.0.0) within *SMRTLink v8.0* with *--min-rq 0.8.* The CCS reads were aligned to their reference genomes (RefSeq accessions CP014225.1 and CP009789.1) using minimap2 (v2.17-r941) (*47*) with *–x sr*, separately. The CCS accuracy (Q, Phred-transformed) with passes *s*, were calculated as:

$$Q_s = -10 \times \log\left(\frac{\sum_1^n M_s}{\sum_1^n L_s}\right)$$

where *n* denotes all CCS reads with passes *s*, *M* is the number of matched bases on each read, L is the length of each read. At 20 passes with *--min-rq* of 0.99, CCS reads achieves Q28 accuracy.

6mASCOPE analysis of *C. reinhardtii* and *T. thermophila*

Genome-wide map of nucleosome positioning

The MNase-seq data of *C. reinhardtii* and *T. thermophila* were obtained from the NCBI Sequence Read Archive (SRA; http://www.ncbi.nlm.nih.gov/sra) with accession codes SRX1008144 (*4*) and SRR5337752 (*5*). The MNase-seq reads were mapped to *C. reinhardtii* (JGI version 9.1) and *T. thermophila* (version T_thermophila_June2014, http://ciliate.org/) reference genomes using *bowtie* (v1.1.2) (*48*). The mapped files were subjected to iNPS (v1.2.2) with default parameters for a genome-wide map of nucleosomes. iNPS was designed with high sensitivity in identifying nucleosome boundaries from MNase-seq (*49*). 73 bp, half of the nucleosome length, was extended on both sides from the center (dyad) of the nucleosome blocks identified by iNPS for complete nucleosomes.

Periodical 6mA pattern around transcriptional start sites (TSS)

With single molecule level detection of 6mA among all individual DNA molecules in gDNA samples of *C. reinhardtii* and *T. thermophila*, CCS reads were mapped to their reference genomes using minimap2 (v2.17-r941) (*47*) with *–x sr*, separately. For the genomic regions with CCS reads mapped, the average IPD ratio of each site was calculated among all CCS reads mapped to this location. Only IPD ratios with coverage ≥ 10 on the CCS reads were kept to ensure an accurate IPD ratio on individual DNA molecules. The IPD ratios of adenines around [−2,000 nt, +2,000 nt] regions surrounding the TSS across different genes were aggregated and normalized to the frequency of adenines.

6mA/A quantification relative to nucleosome and linkers

With the mapped CCS reads on the references, genomic regions of *C. reinhardtii* and *T. thermophila* were divided into ten bins across the genome, including five evenly divided bins (~15bp) between nucleosome dyad and nucleosome boundary, as well as the five evenly divided bins, depending on the length, between nucleosome boundary and the middle of linker. All genomic regions, including both strands, were only binned to the nearest nucleosomes. Using

6mASCOPE, 6mA/A levels within each bins and 3 motif groups, including VATB (V= C,G, or A; B=C,G or T), TATN (N= any nucleotide) or NATA, and others, were quantified, separately.

6mASCOPE analysis of *D. melanogaster* and *A. thaliana*

Contamination identification in *D. melanogaster* embryo ~0.75h and adult gDNA samples

CCS reads were firstly mapped to the *D. melanogaster* reference genome (FlyBase Release 6) using minimap2 (v2.17-r941) (*47*) with *–x sr*. The unmapped CCS reads were hereafter mapped to the NCBI-NT database using BLASTN (2.9.0+) and summarized by species or genus. The main contamination sources included the species *S. cerevisiae*, and genera *Acetobacter* (including the recently proposed genus *Komagataeibacter* (*50*)), and *Lactobacillus*. 6mA/A level for *S. cerevisiae* was re-checked with an additional Sequel II run. The other reads mapped to low abundant species or species without an annotation in NCBI-NT database were summarized together. Although a small proportion of reads from subgroup "Others" may also originate from *D. melanogaster* (possible incomplete reference genome), the 6mA/A level predictions for all subgroups remain essentially the same when we aligned SMRT CCS reads to public long PacBio reads of *D. melanogaster* (SRR9969842 from NCBI-SRA database), which better captures *D. melanogaster* genome than the standard reference.

Contamination identification and phyla classification in *A. thaliana* 21-day-old seedlings and roots

CCS reads were first mapped to the *A. thaliana* reference genome (TAIR 10) using minimap2 (v2.17-r941) (*47*) with *–x sr*. Taxonomic classification of the unmapped reads were performed by Kraken2 (v 2.0.8-beta) with the pre-build database minikraken2_v2_8GB_201904_UPDATE (https://ccb.jhu.edu/software/kraken2/).

Motif enrichment analysis in PBMC and HEK-WGA-MsssI

For both samples, adenines with IPD ratio from 4 to 7 and sequence coverage ≥ 20 were selected and MEME (v 5.1.0) (*51*) was used to detect the motif enriched within the neighborhood (+/- 5bp).

6mASCOPE analysis of HEK-pCI

CCS reads were first deconvolved by 6mASCOPE using the human genome reference (GRCh38). The unmapped CCS reads were hereafter deconvolved by 6mASCOPE using a pCI reference (RefSeq accession U47119.2) and the unmapped CCS reads were subjected to a 3rd round of 6mASCOPE with the *E. coli* reference (RefSeq accession CP014225.1). The other reads ("Others") were summarized together and all subgroups were quantified by 6mASCOPE.

6mA/A quantification by 6mASCOPE on human L1 elements

Human LINE-1 (L1) transposon annotations were collected from the RepeatMasked genome (version 20150807) in Repbase (https://www.girinst.org/repbase/). Those ~6-kb- long L1s were treated as full-length L1s (*52*). CCS reads from human samples, including HEK-dam, native samples (PBMC, glioblastomas and HEK293) and HEK-WGA, were first mapped to the human reference genome (GRCh38) to exclude the non-uniquely mapped reads. The filtered reads were hereafter mapped to all L1 sequences. Both processings were performed using minimap2 (v2.17-r941) (*47*) with *–x sr* and only uniquely mapped CCS reads (MAPQ60) were kept. The evolutionary age for each L1 subfamily was based on the method of Castro-Diaz *et al.* (*53*).

Ultra-high-performance liquid chromatography-quadruple mass spectrometry (UHPLC-MS/MS) analysis

For the DNA digestion, the DNA solutions were prepared in 10 mM Tris-HCl (pH 8.0) plus 1.0 mM $MgCl_2$, and 1.0 U SuperNuclease (Sino Biological, SSNP01, Beijing, China), 0.02 U phosphodiesterase I (Worthington, cat#LS003926, Lakewood, NJ, USA) and 2.0 U Quick CIP (NEB, cat#M0525S). Of note, the DNA solutions was adjusted to 50 μL with 10 mM Tris-HCl (pH 8.0) plus 1.0 mM $MgCl_2$. The DNA solutions was incubated at 37 °C for 3 h, and then filtered by ultra-filtration tubes (MW cutoff: 3 KDa, Pall, cat#OD003C35, Port Washington, NY, USA). Finally, the filtrates were subjected to UHPLC-MS/MS analysis for detection of 6mA and dA.

The method of UHPLC-MS/MS analysis was described as previously but with a few minor modifications (*18*, *54*). Briefly, the UHPLC-MS/MS analysis was performed on an Agilent 1290 Infinity II ultra-high performance LC system coupled with an Agilent 6470 triple quadrupole mass spectrometer. Prior to mass spectrometry analysis, a reversed-phase Agilent Zorbax Eclipse Plus C18 column (2.1 × 50 mm I.D., 1.8 μm particle size) was used for the UHPLC separation. Mass spectrometer was operated in positive ionization mode with 4000 V capillary voltage as well as 150 °C and 11 L/min nitrogen drying gas. A multiple reaction monitoring (MRM) mode was adopted: m/z 252→136 for dA (collision energy, 5 eV), and m/z 266→150 for 6mA (15 eV).

Supplementary Text

The analytical issues faced by previous 6mA studies that used the Single Molecule Real Time (SMRT) Sequencing

Existing methods (*e.g. SMRTanalysis* or *SMRTLink*, Pacific Biosciences, Inc) for 6mA analysis using SMRT sequencing depends on a reference known *a priori* and uses IPD ratio as the indicator for 6mA detection. The core method first aligns the subreads to a reference genome. For a specific genomic location, the IPD values are aggregated from the mapped subreads, either from the same or several different sequencing molecules. IPD value is the measurement of the time for the DNA polymerase to translocate from one nucleotide to the next during SMRT sequencing (*25*). The deviation of an IPD value from the expected level, namely IPD ratio, is highly correlated with the presence of corresponding modifications of the nucleotide (*3, 14, 25, 29, 30*). The expected IPD value can be estimated by either whole genome amplification (WGA) control, or *in silico* from the built-in model based on the genome context around the analyzed base (from -10bp to +4bp, with which the DNA polymerase physically interacts) on the reference (*3, 14, 29*). IPD ratios between ~3 and ~7 are highly correlated with 6mA events (*3, 14, 25, 29, 30*). This reference-based method ignores the reads from potential bacterial contamination beyond the reference genome, yet the bacterial reads can carry abundant 6mA events .

Although 6mA-DIP-seq can also detect bacterial contamination, it does not provide a quantitative estimation of 6mA and only report 6mA peaks detected from the eukaryotic species of interest. In addition, 6mA-DIP-seq is prone to false positives due to 6mA antibody bias(*15, 20*) in the eukaryotic genomes.

A method for reference-free 6mA analysis

To address this fundamental challenge and reliably analyze 6mA in eukaryotic genomes, instead of using a defined genomic reference, we use the circular consensus sequence (CCS) as the molecule-specific reference for 6mA analysis. We designed the short SMRT insert libraries, 200-400bp in size, to obtain more accurate sequence determination and collect repeated measures of IPD for more reliable 6mA detection.

Our data quality control requires at least 20 passes and *--min-rq* (i.e. *--minPredictedAccuracy* in older CCS software version) at least 0.99 (**Methods; Fig. S2**). The Predicted Accuracy from CCS software has a median phred-transformed score of Q 31.46 (0.9993). Averaged base score for CCS read has a median of Q 92.39. The accuracy measured by mapping the CCS reads to the reference genome achieves Q28 and stays saturated at 20 passes for *E. coli* K12 MG1655 and ER3413 (**Fig. S2**), indicating the CCS accuracy is reliable. It is worth noting although CCS improves base calling accuracy, there can be some level of remaining errors, which are effectively accounted for in our machine learning model based on data with a wide range of 6mA/A levels that were processed with the same quality control procedures (**Methods**).

6mA/A quantification by 6mASCOPE

The 6mA level estimated by 6mASCOPE is associated with a 95% confidence interval, which represents a 95% probability that the calculated confidence interval encompasses the true value being estimated **(Methods)**. As shown in **Fig. 5A**, except for the *D. melanogaster* embryo and *A. thaliana* gDNA samples (both are contaminated by bacteria), 6mA level quantifications along with 95% confidence intervals are consistent with the measurements of independent UHPLC-MS/MS among multiple samples.

Confidence interval facilitates dataset-specific estimation for each 6mA/A level quantification by 6mASCOPE. As expected, the confidence interval is narrower for high 6mA/A levels, but it becomes wider as 6mA/A level decreases. Also as expected, at the same 6mA/A level, the confidence interval is narrower with more CCS reads (**Fig. 2F, Fig. S5B)**. In another word, the detection limit of 6mASCOPE, the level when the confidence interval covers 1ppm (the lowest 6mA level in our training dataset), depends on the DNA molecule numbers sequenced by SMRT-sequencing. Deeper sequencing is expected to enable lower detection limit. However, it is worth noting, even with a large number of CCS reads, 6mASCOPE does not precisely differentiate 6mA levels below 10ppm because the lower confidence interval includes 1ppm, which is the lowest 6mA level in our training dataset **(Fig. 2F).**

Supplementary Figures

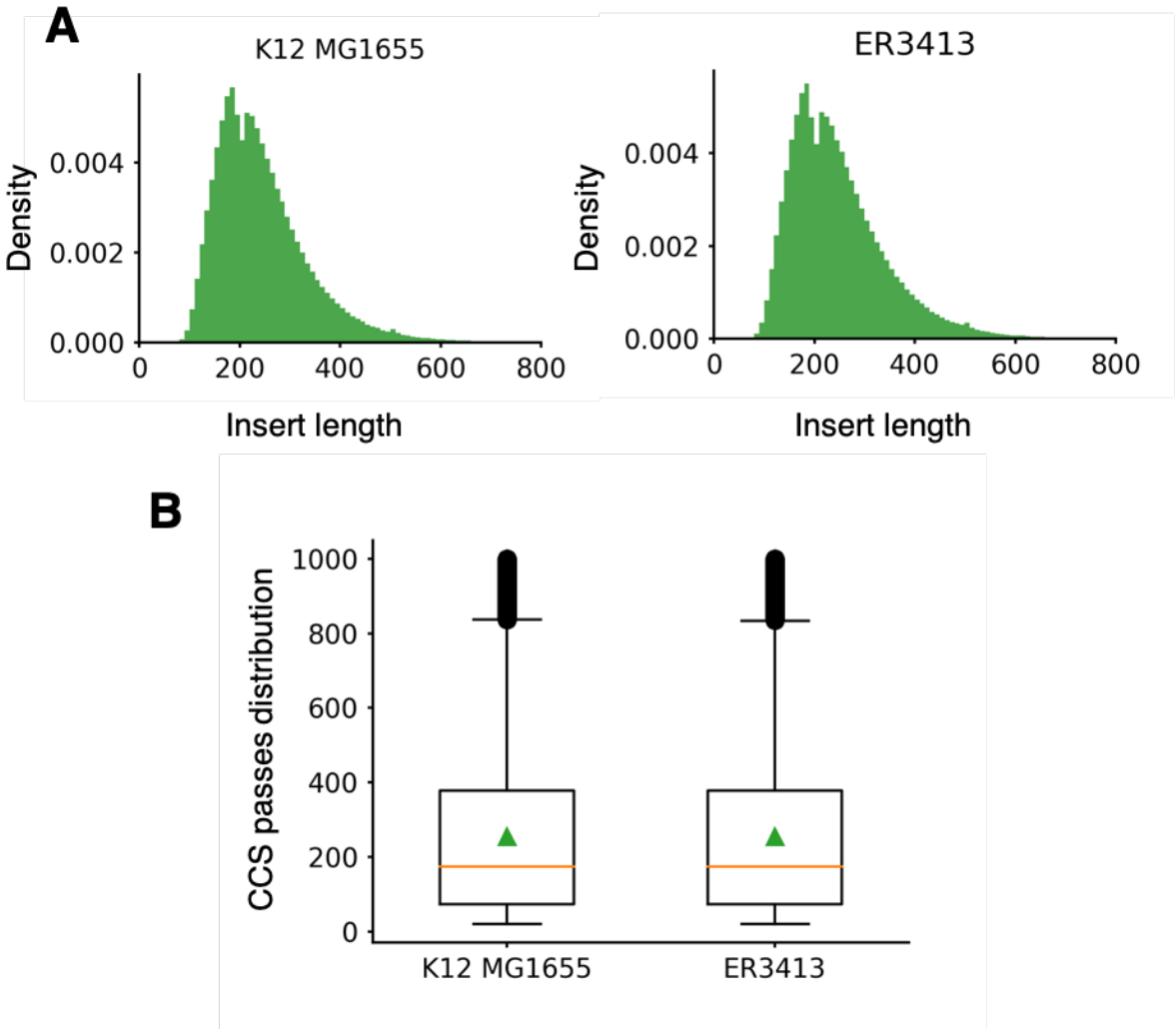

Fig. S1. Short insert libraries by SMRT sequencing.

With the short insert libraries, ~200bp in size (**A**), DNA molecules are sequenced for a large number of passes by the SMRT DNA polymerase (**B**). The bottom and top of the boxplot represent the first and third quartiles. The orange lines and green triangles inside the boxes denote the medium (second quartile) and the mean, respectively. The whiskers extend 1.5 interquartile range (third quartile minus first quartile) upward from the third quartile, and downward from the first quartile. Outliers are marked with circles.

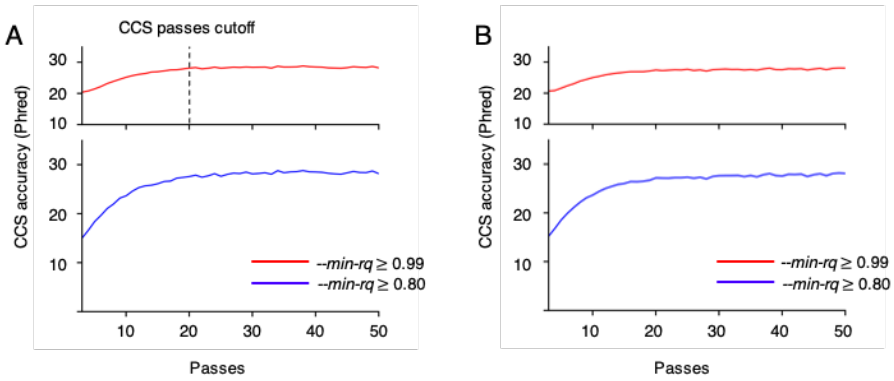

Fig. S2. CCS accuracy over an increasing number of passes.

CCS accuracy is Phred-transformed and a higher value indicates better consensus accuracy. CCS reads from *E. coli* K12 MG1655 (**A**) and ER3413 (**B**) are mapped to their own references respectively. *--min-rq*, a parameter of minimum score for CCS polishing in the CCS generating software. At 20 passes with *--min-rq* of 0.99, CCS reads achieve Q28 accuracy.

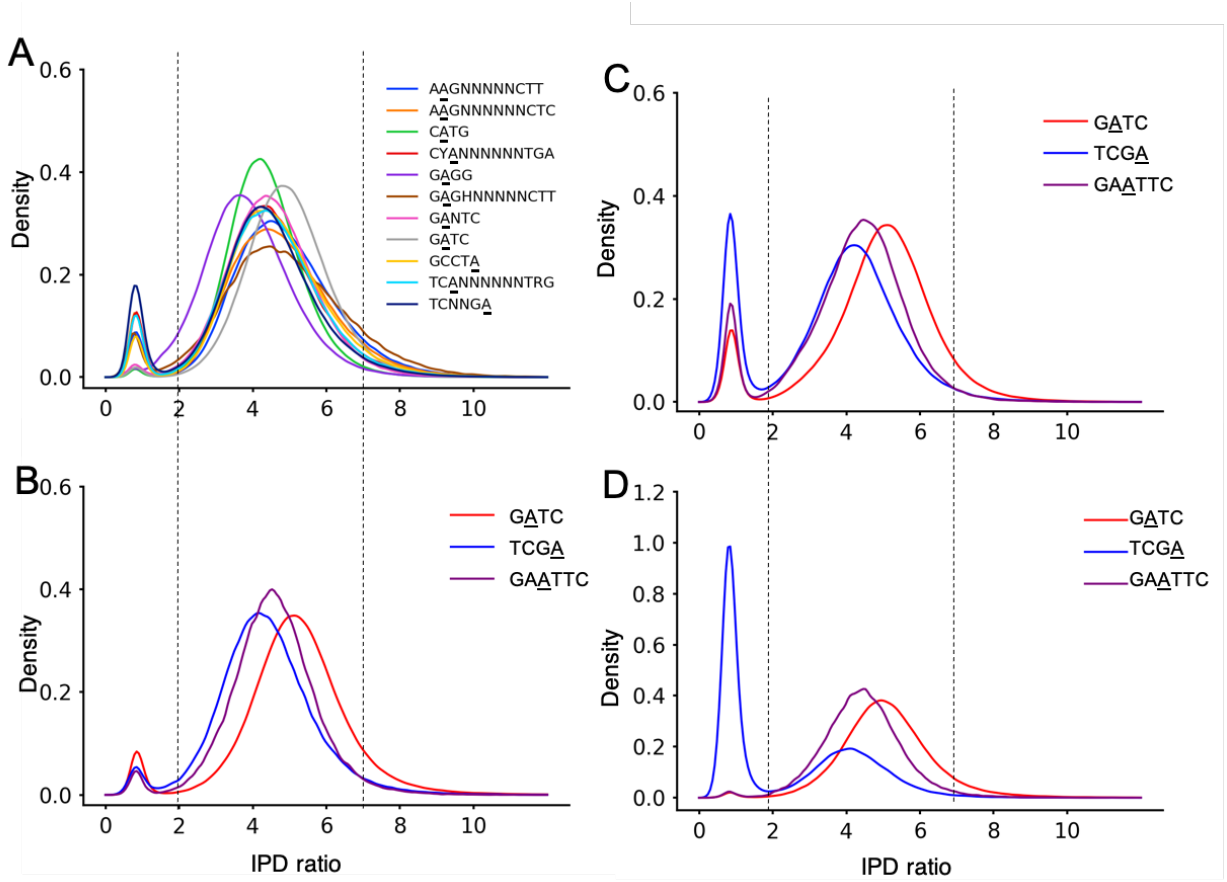

Fig. S3. The three motifs covered by the three commercially available methyltransferases well represent the mean IPD ratios across the highly diverse motifs in *H. pylori*.

H. pylori is shown in (**A**). The IPD ratios of the three motifs (mean between ~4 and ~5.5) are evaluated across three positive controls: HEK-WGA-MsssI-3M (**B**), HEK-WGA-3M (**C**) and HEK293-3M (**D**). IPD ratios from 2 to 7, aligned across the four subfigures, are used to select positive 6mA events as the positive controls for machine learning model training and testing (Methods).

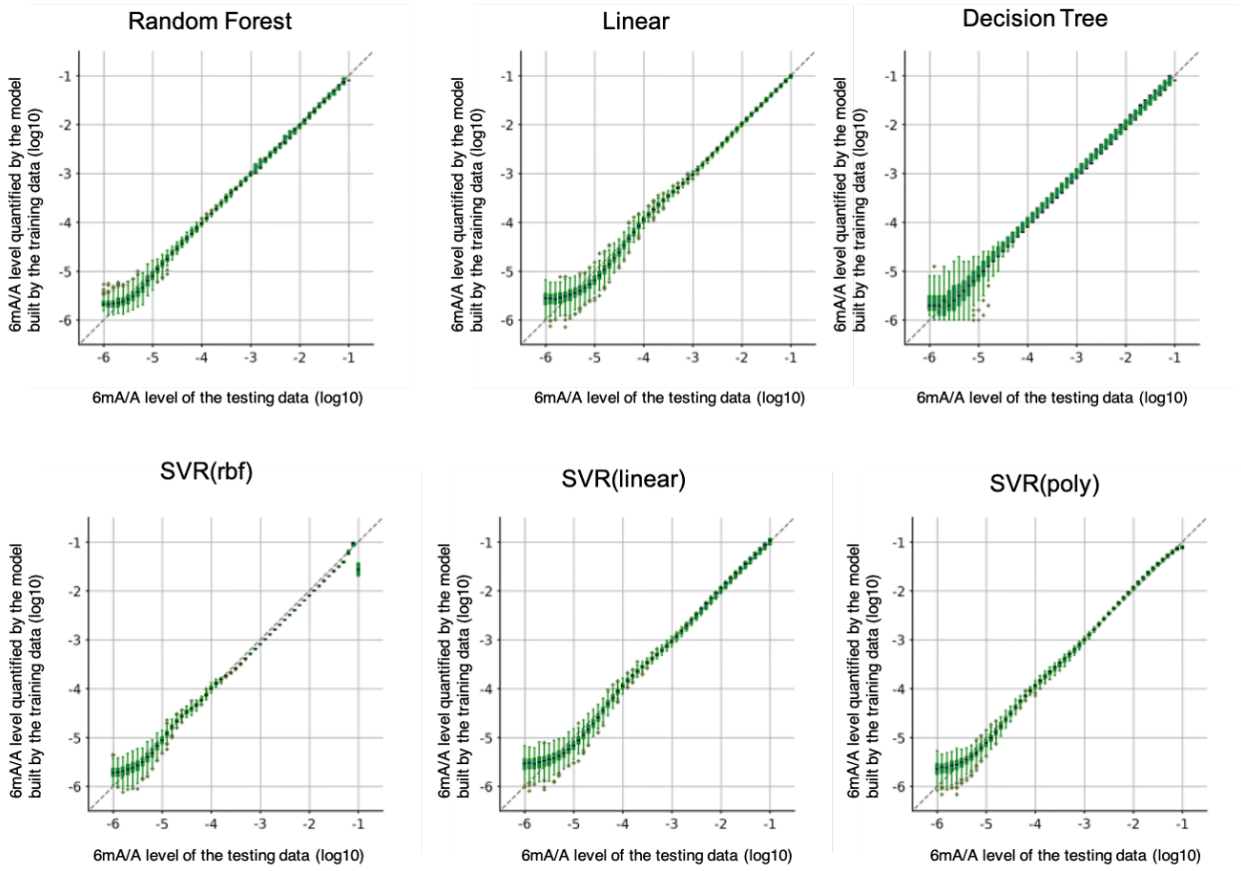

Fig. S4. Leave-One-Out independent cross validation of different machine learning models.

For each model, one of the 51 6mA/A levels (*x-axis*), including all 300 repeated instances, were held out as the testing data for each round of cross validation, while the other 50 levels were used as the training dataset (51 rounds of cross validation). The distribution of the 6mA/A quantification for each round of the leave-one-out cross validation is shown as a boxplot.

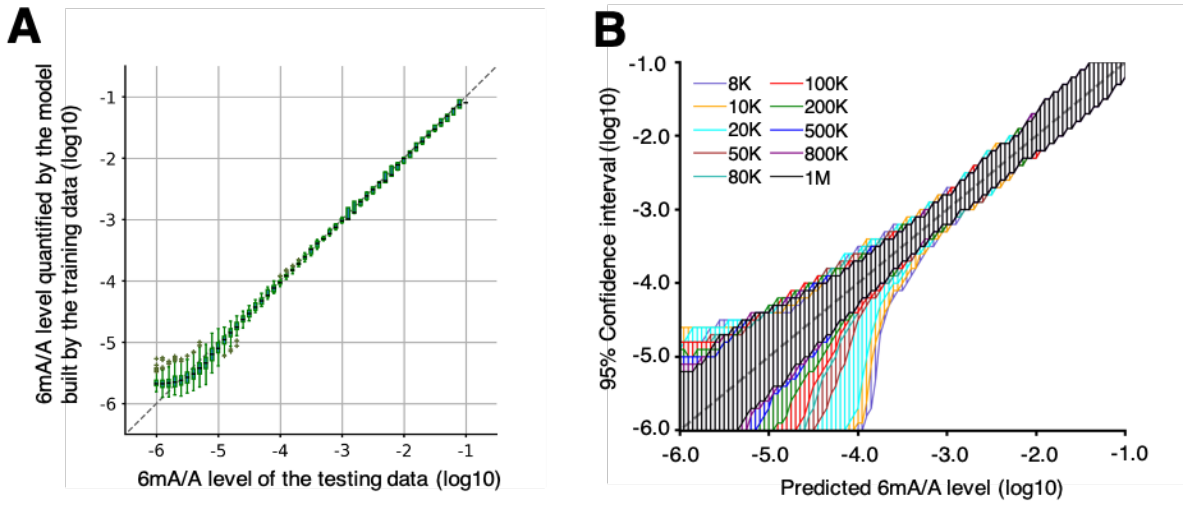

Fig. S5. Cross validation and confidence interval of 6mASCOPE.

(**A**) Leave-One-Out independent cross validation of the machine learning model of 6mASCOPE. Leave-One-Out independent cross validation for each model was performed and illustrated as described in Fig. S4. (**B**) Same description as Fig. 2F, with additional confidence intervals for smaller number of CCS reads.

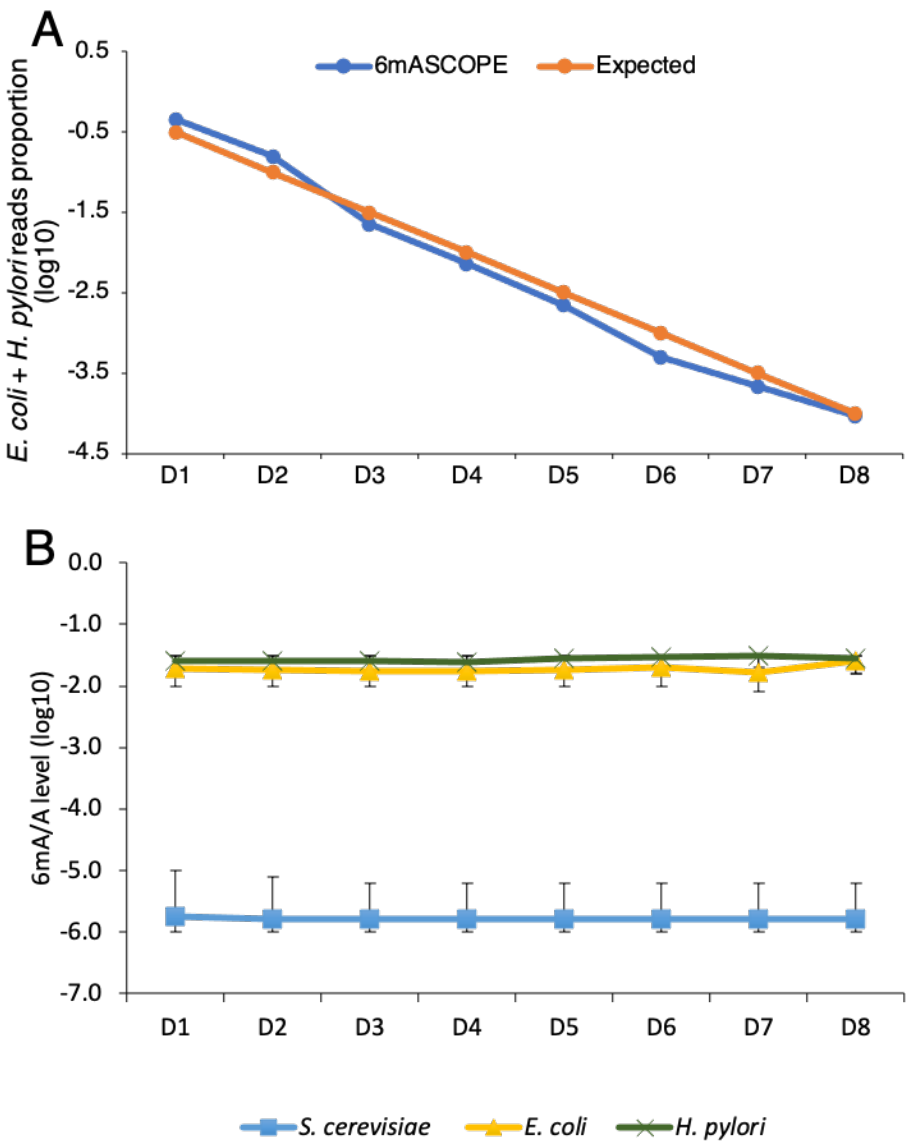

Fig. S6. Serial dilution validation of 6mASCOPE.

6mASCOPE analysis on a serial dilution of *E. coli* (high 6mA/A level) *and H. pylori* (high 6mA/A level) gDNA using *S. cerevisiae* (6mA/A level <1 ppm) at different dilution ratios (D1:D8). Expected ratios of *E. coli* & *H. pylori* (log10-transformed) among the total are correlated with ratios deconvolved by 6mASCOPE (log10-transformed) (**A**). 6mA/A levels of the three resources are quantified by 6mASCOPE (log10-transformed, 95% confidence intervals as error bars) (**B**).

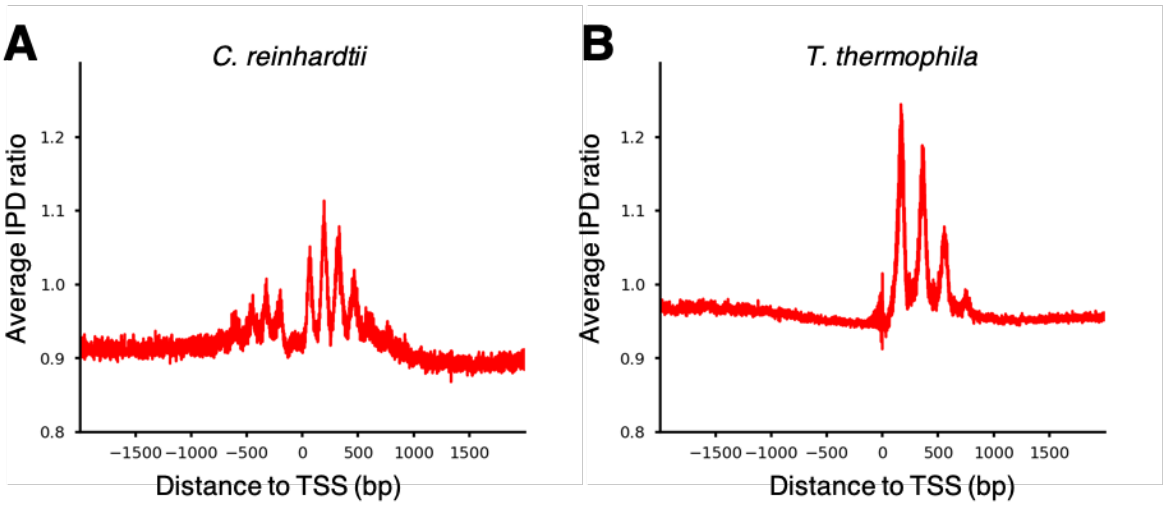

Fig. S7. Periodical pattern of 6mA in *C. reinhardtii* and *T. thermophila.*

6mA has a periodical pattern inversely correlated with nucleosomes near TSSs in *C. reinhardtii* (**A**) and *T. thermophila* (**B**). The average IPD ratios of adenines surrounding the TSS across different genes were aggregated and normalized to the frequency of adenines.

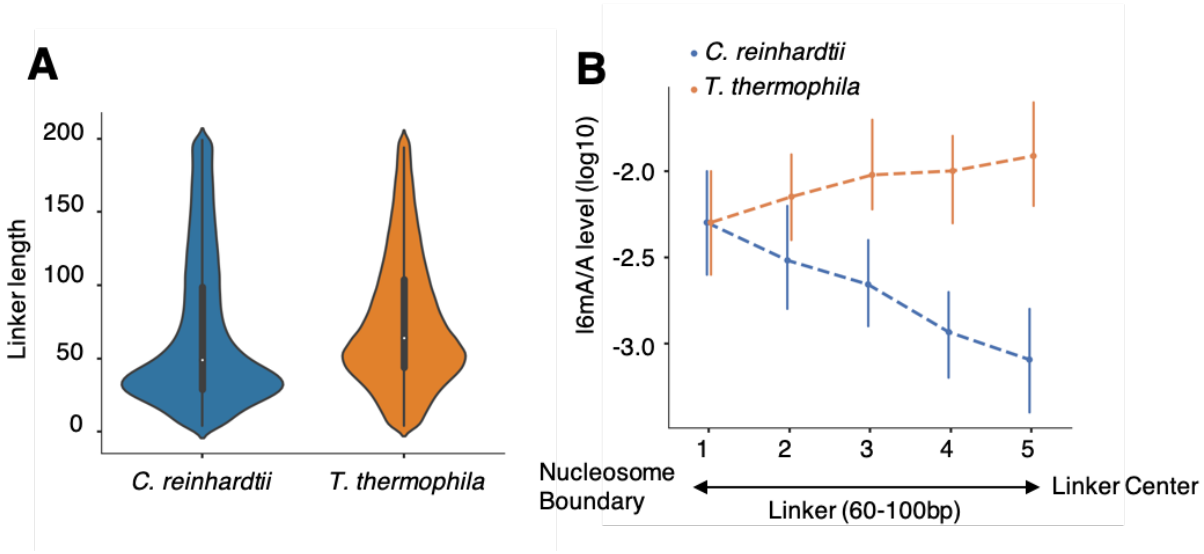

Fig. S8. The different 6mA deposition between *T. thermophila* and *C. reinhardtii* as shown in Fig. 3A was not due to the slightly longer nucleosome linkers in *T. thermophila* than in *C. reinhardtii*.

(**A**) Distribution of linker length (0-200bp) in *C. reinhardtii* and *T. thermophila*. The shape of the violins denotes the probability density of the data at different values. The inside boxplot shows the four quantiles as described in Fig S1. (**B**) Different 6mA depositions in *C. reinhardtii* and *T. thermophila* within linkers of the same length. Linker regions with a length of 60-100bp are divided into five bins (*x-axis*) between the nucleosome boundary and the middle of the linker across the genome. For each bin, 6mA/A level (*y-axis*, log10-transformed) was quantified with 6mASCOPE. Dashed lines join the 6mA quantifications from five bins with error bars showing 95% confidence interval for the prediction. Within the same bin, two species are plotted separately along the *x-axis* to ease visualization.

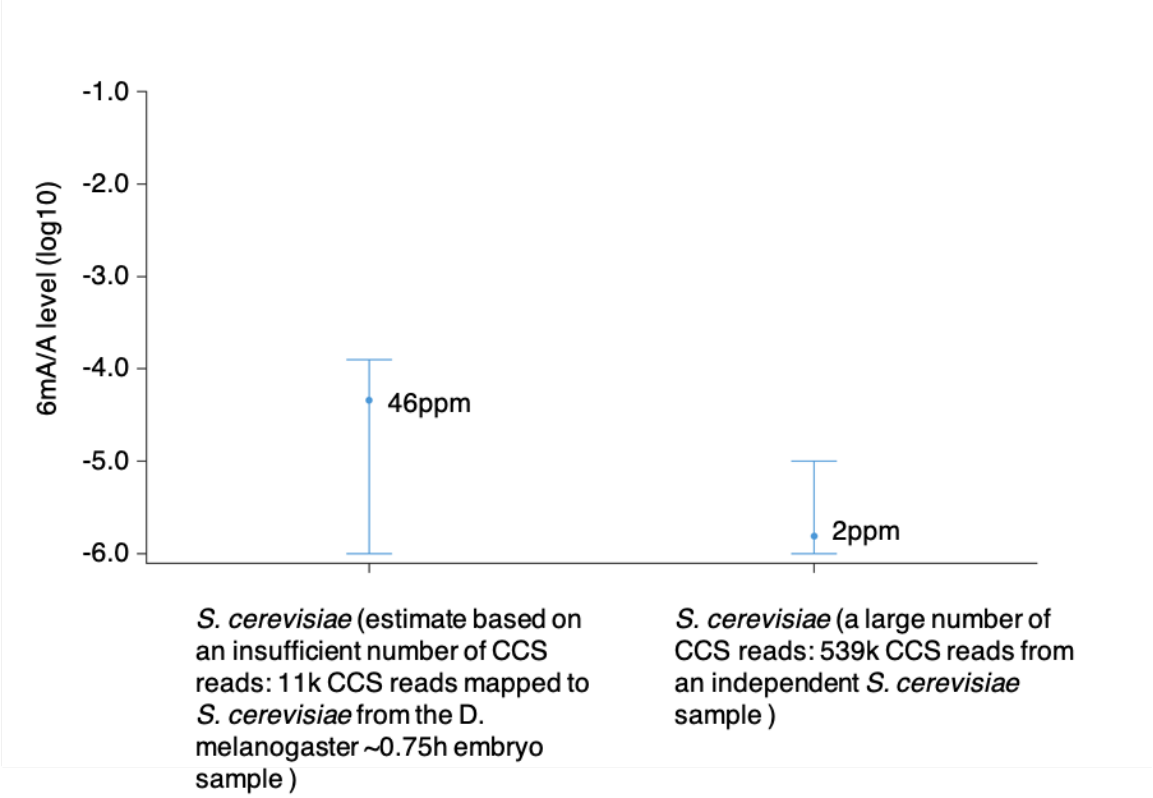

Fig. S9. The extremely low 6mA/A level in *S. cerevisiae* quantified by 6mASCOPE.

(Left) 6mA/A level estimated based on a small number of CCS reads mapped to *S. cerevisiae* from the *D. melanogaster* ~0.75h embryo sample (11k CCS reads only). Importantly, the estimate is not reliable due to the small number of CCS reads. (Right) 6mA/A level estimated based on a large number of CCS reads from an independent *S. cerevisiae* sample (539k CCS reads). The large number of CCS reads provides a more reliable estimation.

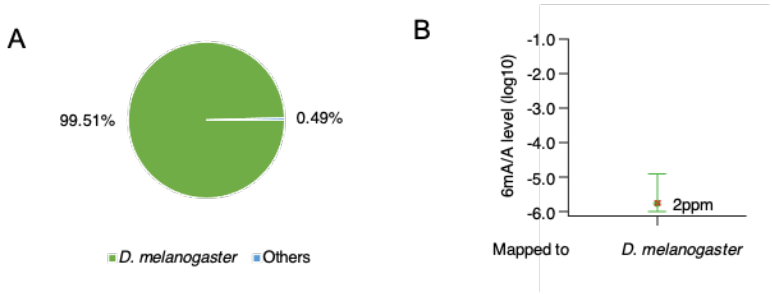

Fig. S10. 6mASCOPE analysis of *D. melanogaster* adult sample shows very low bacteria contamination; consistently, the UHPLC-MS/MS estimation of 6mA/A is very low.

(**A**) Compositions in the *D. melanogaster* adult sample. Percentage shows CCS reads mapped to different subgroups among the total. (**B**) 6mA quantification by 6mASCOPE on CCS reads mapped to *D. melanogaster* genome. Log10-transformed 6mA/A level predictions are shown with a green dot, log10-transformed 95% confidence interval with error bars. The 6mASCOPE prediction is consistent with the measurements of UHPLC-MS/MS (the red cross, on top of the green dot).

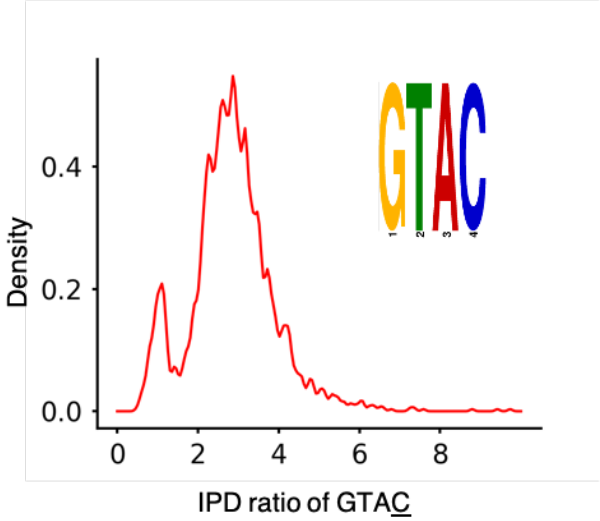

Fig. S11. IPD ratios of cytosines in the 4mC motif GTAC discovered from *Acetobacter* in the fly embryo sample.

5

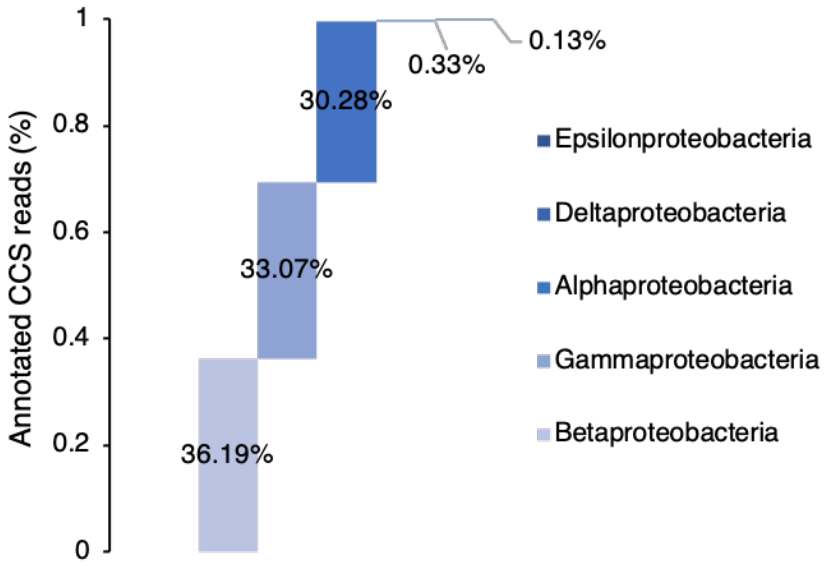

Fig. S12. Main classes of Proteobacteria phylum detected in the CCS reads in *A. thaliana* 21-day-old seedling sample.

Percentage shows CCS reads mapped to the different class among the total phylum.

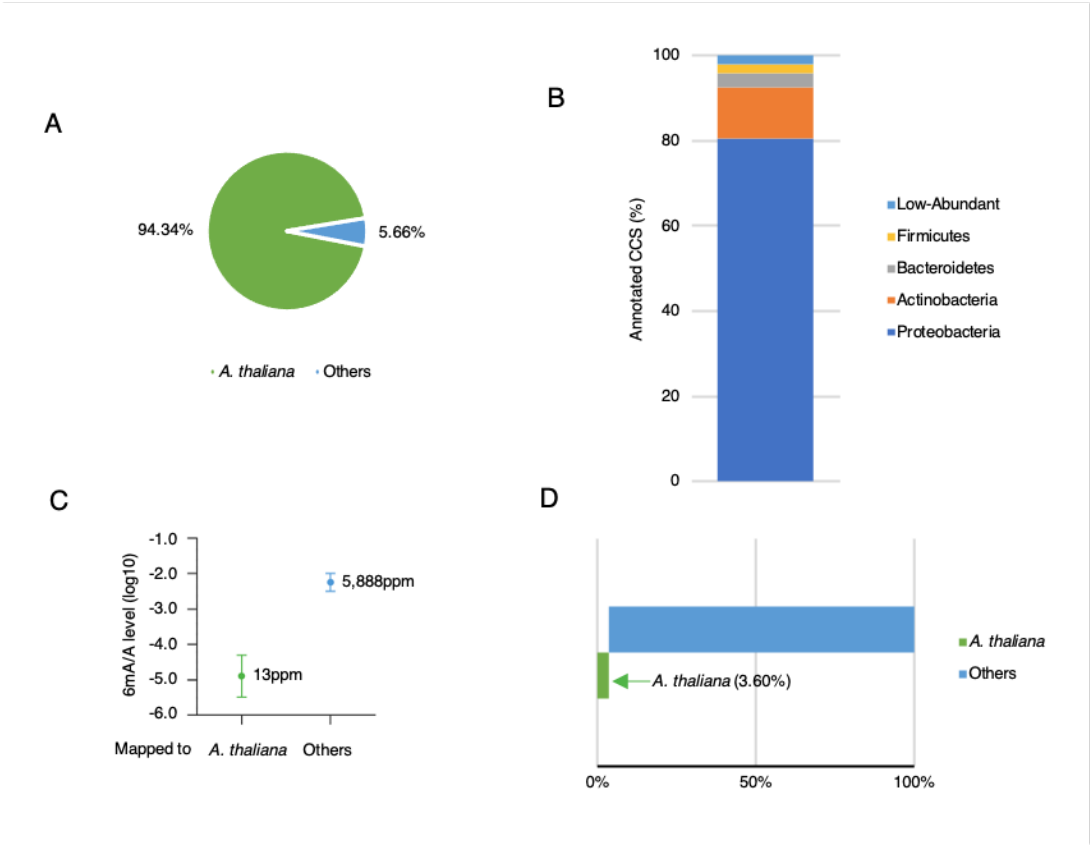

Fig. S13. 6mASCOPE analysis of *A. thaliana* 21-day-old root sample shows more relative microbiome contamination and less relative contribution from *A. thaliana* to the total 6mA abundance.

(**A**) Contamination detected in *A. thaliana* 21-day-old root sample. Percentage shows CCS reads mapped to *A. thaliana* genome and unmapped ("Others") among the total. (**B**) Main phyla detected in contaminated CCS reads. The CCS reads in subgroup "Others" are annotated and taxonomy classified with Kraken2. (**C**) 6mA quantification by 6mASCOPE on the CCS reads mapped to *A. thaliana* genome and Others. For both subgroups, log10-transformed 6mA/A level predictions are shown with dots, log10-transformed 95% confidence interval as error bars. (**D**) Contribution of CCS reads mapped to *A. thaliana* and others in the *A. thaliana* root sample: CCS reads mapped to *A. thaliana* only contribute to 3.60% of the total 6mA events in the total gDNA sample (green arrow).

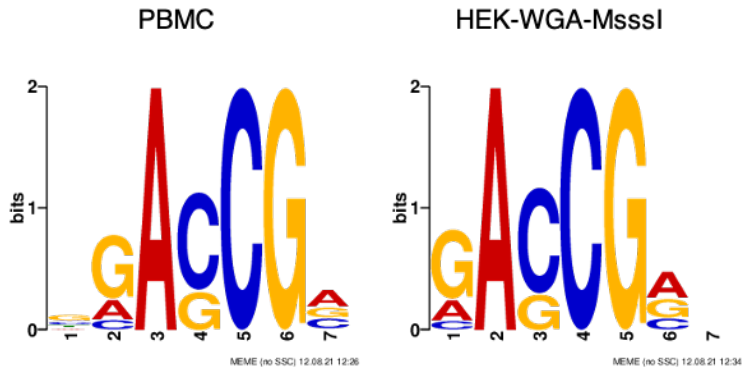

Fig. S14. The predominant motif found in PBMC was also identified in the negative control sample (HEK-WGA-MsssI).

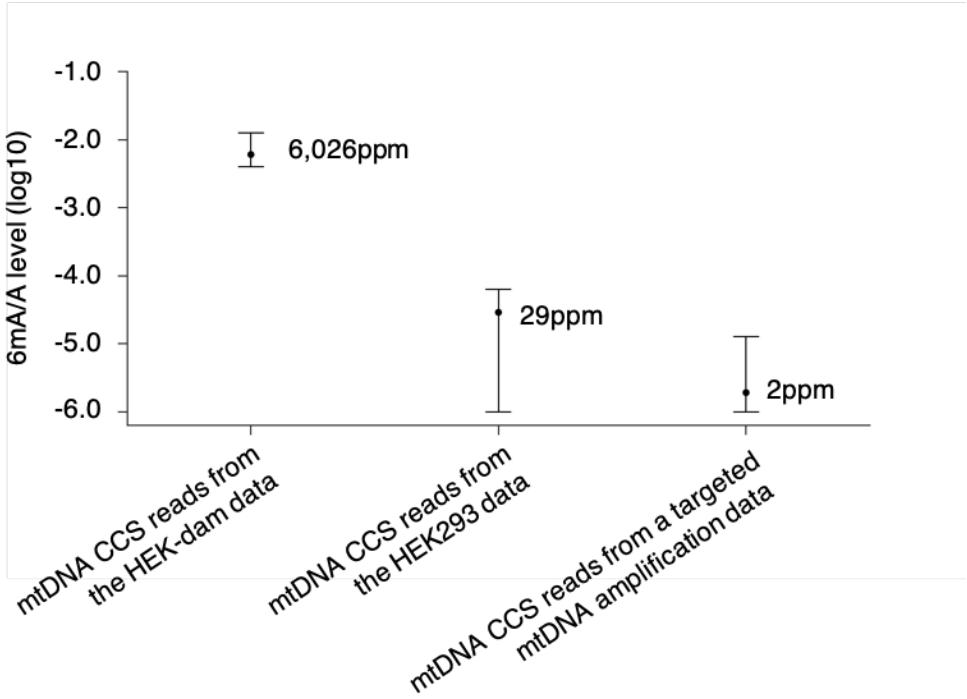

Fig. S15. No evidence supporting a high 6mA/A level on mtDNA in HEK293.

6mA/A levels estimated by 6mASCOPE for reads mapped to mtDNA in HEK-dam (positive control, 5,837 CCS reads), mtDNA in HEK293 (7,205 CCS reads), and an mtDNA sample from targeted mitochondrial genome amplification (negative control, 288,814 CCS reads). It is worth noting that the 6mA/A level of 29ppm with 95% confidence interval of 1-63ppm (based on 8000 CCS reads) does not support a high 6mA/A on mtDNA in the HEK293 sample.

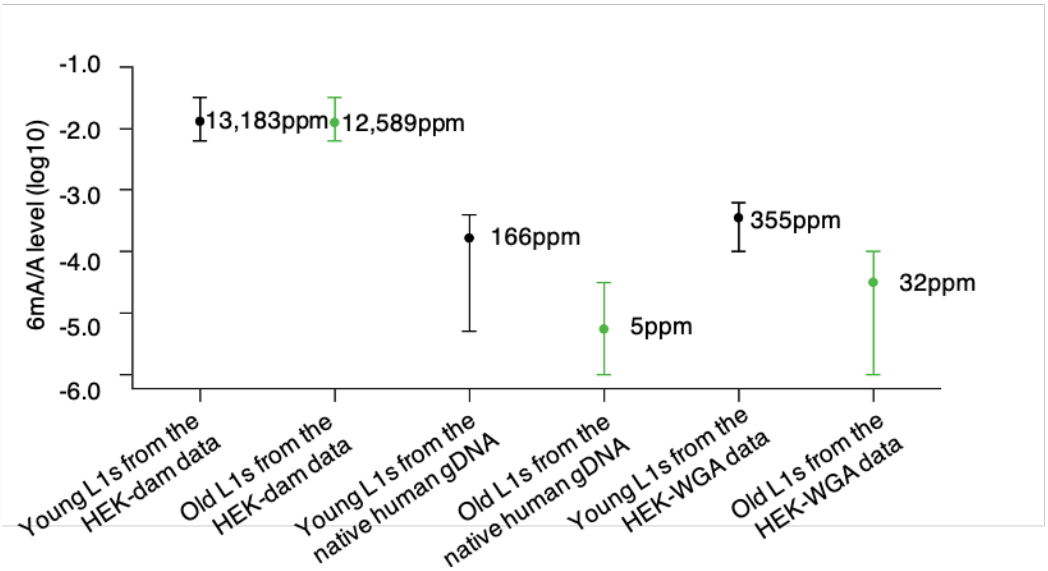

Fig. S16. 6mA/A levels estimated by 6mASCOPE for reads mapped to human L1 elements.

Two types of L1 elements are evaluated: young full length L1 (evolutionary age<10 Myr, including L1HS and L1PA2), and older L1 (evolutionary age>10 Myr), defined by Castro-Diaz *et al.*(*53*) (Methods). HEK-dam, where 6mA events were added to GATC sites *in vitro* to HEK-WGA gDNA, is used as the positive control: both young and old L1s have high 6mA/A levels as expected. Native (young or old) L1 elements are based on CCS reads uniquely mapped to human L1 elements from four datasets: HEK293, PBMC and the two glioblastomas. WGA (young/old) L1 elements are based on CCS reads uniquely mapped to human L1 elements from the HEK-WGA data. The confidence intervals of young L1 are defined based on 8000 CCS reads. The comparison of 6mA/A level in young full length L1 and older L1 in Native vs. WGA samples does not support the enrichment of 6mA in young full length L1 of native gDNA samples.

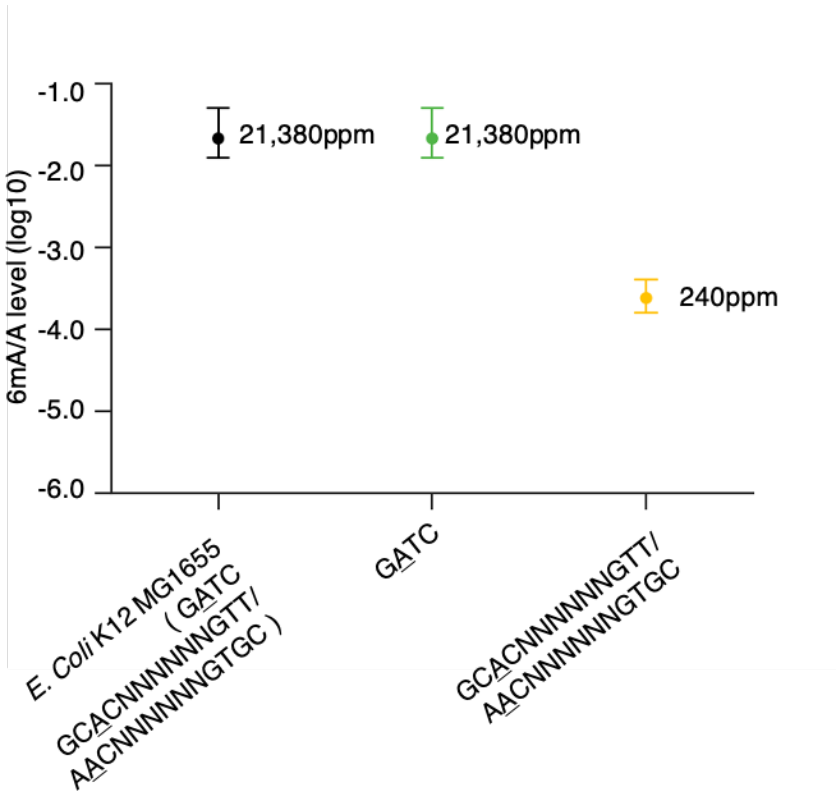

Fig. S17. Dam MTase mutant *E.coli* still has a substantial amount of 6mA events.

Dam MTase mutant *E.coli* has the remaining 6mA MTase hsdM, which methylates 1,194 AACNNNNNNGTGC and GCACNNNNNNGTT motif sites across the *E. coli K12* MG1655 genome (*28*, *55*). This is further confirmed with 6mA/A level quantification by 6mASCOPE on *E. coli K12* MG1655 (containing Dam on GATC and hsdM on GCACNNNNNNGTT/ AACNNNNNNGTGC), *E. coli* with Dam only (GATC), and *E. coli* with HsdM only (GCACNNNNNNGTT / AACNNNNNNGTGC).

Supplementary Table

Table S1. Advantages and limitations of commonly used 6mA detection methods

Methods	Advantages	Limitations
Anti-6mA antibody-based dot blotting	• Rapid detection • Semi-qualitative estimation	• Semi-quantitative estimation • Antibody bias • Confounding by bacterial DNA contamination and eukaryotic RNA contamination
Liquid chromatography coupled with tandem mass spectrometry (LC–MS/MS)	• Quantitative estimation • Highly sensitive	• Confounding by bacterial contamination in samples and/or bacteria-origin enzymes used in MS sample preparation.
Restriction enzyme digest followed by next-generation sequencing (RE-seq)	• Single nucleotide resolution • Highly sensitive • Limited input DNA requirement • Specificity on the known target sequence	• Limited choice of restriction enzymes
Anti-6mA antibody-based DNA immunoprecipitation followed by Illumina sequencing (DIP-seq)	• Widely used	• Antibody bias • Confounding by bacterial DNA contamination and eukaryotic RNA contamination • Large amount of input DNA required
Single-molecule, real-time sequencing (SMRT-seq)	• Single-nucleotide resolution • Single-molecule resolution	• Lack of sensitivity / False positive calls when 6mA/A level is low • Currently available computational methods mainly for prokaryotes and protozoa with haploid genome and high 6mA/A abundance

Table S2. Sequencing information of all data (post-filtering) generated in this paper.

Sample	Platform	No. of CCS reads	Insert length (Mean)	Insert length(Median)	Passes(Mean)	Passes(Median)
***E. coli* K12 MG1655**	Sequel II	966,574	247.07	227.00	272.41	181.00
***E. coli* ER3413**	Sequel II	1,087,810	247.66	229.00	271.33	180.00
C. reinhardtii	Sequel II	862,205	251.61	236.00	241.37	162.00
T. thermophila	Sequel II	975,050	235.86	221.00	288.22	193.00
***D. melanogaster* (embryo)**	Sequel II	674,650	207.16	190.00	331.32	222.00
***D. melanogaster* (adult)**	Sequel II	405,549	246.84	223.00	248.76	160.00
***A. thaliana* (seedling)**	Sequel II	535,030	251.69	230.00	292.10	196.00
***A. thaliana* (root)**	Sequel II	616,493	184.39	164.00	305.44	204.00
S. cerevisiae	Sequel II	548,647	212.51	193.00	271.61	179.00
H. pylori	Sequel II	989,319	449.04	420.00	155.41	114.00
H. vesiculosa	Sequel II	768,431	245.09	216.00	246.29	160.00
HEK293	Sequel II	818,207	315.74	288.00	233.12	145.00
HEK293-3M	Sequel II	1,266,786	227.72	207.00	262.08	170.00
HEK-WGA	Sequel II	910,585	298.87	274.00	237.48	149.00
HEK-WGA-3M	Sequel II	1,213,454	275.27	250.00	260.28	173.00
HEK-WGA-MsssI	Sequel II	1,954,129	252.76	228.00	274.90	194.00
HEK-WGA-MsssI-3M	Sequel II	1,164,880	225.12	205.00	252.65	165.00
HEK-dam	Sequel II	722,190	309.92	281.00	233.18	147.00
mtDNA-WGA	Sequel II	637,761	218.78	201.00	319.88	215.00
PBMC	Sequel II	570,283	251.14	226.00	292.83	191.00
Glioblastoma-1	Sequel II	247,700	259.29	245.00	270.40	171.00

Glioblastoma-2	Sequel II	280,763	278.80	265.00	258.10	164.00
HEK-pCI	Sequel II	741,558	272.46	255.00	258.87	171.00
***E. coli*/*H. pylori*/*S. cerevisiae*-D1**	Sequel II	196,805	207.38	188.00	235.28	143.00
***E. coli*/*H. pylori*/*S. cerevisiae* -D2**	Sequel II	146,337	232.91	211.00	228.11	143.00
***E. coli*/*H. pylori*/*S. cerevisiae* -D3**	Sequel II	212,934	226.85	204.00	235.81	146.00
***E. coli*/*H. pylori*/*S. cerevisiae* -D4**	Sequel II	245,254	240.36	217.00	226.41	141.00
***E. coli*/*H. pylori*/*S. cerevisiae* -D5**	Sequel II	189,598	217.31	194.00	243.00	150.00
***E. coli*/*H. pylori*/*S. cerevisiae* -D6**	Sequel II	270,119	205.18	186.00	249.46	156.00
***E. coli*/*H. pylori*/*S. cerevisiae* -D7**	Sequel II	280,214	212.12	191.00	235.66	148.00
***E. coli*/*H. pylori*/*S. cerevisiae* -D8**	Sequel II	264,110	237.50	211.00	226.15	141.00
HEK293	RSII	18,342	184.83	164.00	121.32	107.00
HEK-WGA	RSII	64,643	188.20	170.00	144.42	121.00
ER3413	RSII	52,144	180.82	161.00	143.16	112.00