RESEARCH ARTICLE

MOLECULAR BIOLOGY

Critical assessment of DNA adenine methylation in eukaryotes using quantitative deconvolution

Yimeng Kong[1], Lei Cao[1]†, Gintaras Deikus[1]†, Yu Fan[1]†, Edward A. Mead[1]†, Weiyi Lai[2], Yizhou Zhang[3], Raymund Yong[3], Robert Sebra[1,4,5], Hailin Wang[2], Xue-Song Zhang[6], Gang Fang[1]*

The discovery of N^6-methyldeoxyadenine (6mA) across eukaryotes led to a search for additional epigenetic mechanisms. However, some studies have highlighted confounding factors that challenge the prevalence of 6mA in eukaryotes. We developed a metagenomic method to quantitatively deconvolve 6mA events from a genomic DNA sample into species of interest, genomic regions, and sources of contamination. Applying this method, we observed high-resolution 6mA deposition in two protozoa. We found that commensal or soil bacteria explained the vast majority of 6mA in insect and plant samples. We found no evidence of high abundance of 6mA in *Drosophila*, *Arabidopsis*, or humans. Plasmids used for genetic manipulation, even those from Dam methyltransferase mutant *Escherichia coli*, could carry abundant 6mA, confounding the evaluation of candidate 6mA methyltransferases and demethylases. On the basis of this work, we advocate for a reassessment of 6mA in eukaryotes.

For decades, N^6-methyldeoxyadenine (6mA) has been known to be widespread in prokaryotes as a regulator of DNA replication, repair, and transcription (*1*–*3*). Recently, 6mA has also been reported to be prevalent in eukaryotes. Unlike the generally high abundance of 6mA in bacteria, 6mA/A levels (6mA events relative to all adenines) in eukaryotic organisms vary over several orders of magnitude (*4*–*13*). A few unicellular organisms have very high 6mA/A levels: 0.4% in *Chlamydomonas reinhardtii* (*4*), 0.66% in *Tetrahymena thermophila* (*5*), and as much as 2.8% in early-diverging fungi (*6*). In contrast, 6mA/A levels reported in multicellular eukaryotes are much lower: ~0.1% to ~0.0001%, or undetectable (*8*, *10*–*12*, *14*, *15*). Nonetheless, important functions have been assigned to 6mA in eukaryotes, suggesting additional epigenetic mechanisms in basic biology and human diseases (*11*). However, other studies have cast doubt on the existence and levels of 6mA in eukaryotic DNA (*15*–*19*). For example, liquid chromatography coupled with tandem mass spectrometry (LC-MS/MS) can reliably quantify 6mA with high sensitivity, but it cannot discriminate eukaryotic 6mA from bacterial 6mA contamination (*16*, *20*). Unique metabolically generated stable isotope labeling can address this limitation of LC-MS/MS (*17*, *18*); however, it can only be used in cultured cells. Anti-6mA antibody–based dot blotting is commonly used to estimate 6mA levels (*4*, *5*, *7*, *9*–*12*), but it cannot rule out bacterial contamination. In addition, anti-6mA antibody–based DNA immunoprecipitation sequencing (DIP-seq) is often used for 6mA mapping (*7*, *8*, *10*, *13*, *21*), but it can be confounded by 6mA-independent factors such as DNA secondary structures (*20*) and RNA contamination (*15*). Restriction enzyme–based 6mA analyses are constrained by their limited recognition motifs (*4*, *22*). Single-molecule real-time (SMRT) sequencing (*23*) and nanopore sequencing (*24*) provide opportunities for directly mapping 6mA events (*3*, *25*, *26*), but the existing methods are mainly for mapping 6mA in prokaryotes and protozoa with high 6mA abundance (*3*, *14*, *26*–*29*). For eukaryotes with low 6mA abundance, these methods are prone to yield many false positive calls due to low sensitivity (*14*–*16*).

The lack of a reliable technology that accurately quantifies 6mA/A levels in eukaryotic genomes motivated us to develop a method, named 6mASCOPE, for quantitative 6mA deconvolution (Fig. 1). The method, based on a short-insert SMRT library design (Fig. 1A), examines all DNA molecules sequenced in a genomic DNA (gDNA) sample, separates the total sequences into different sources, and quantitatively deconvolves the total 6mA events into each of the sources (Fig. 1B). We first validated our method over a wide range of 6mA/A levels, from 10^{-6} to 10^{-1}, and then examined a number of eukaryotes.

[1]Department of Genetics and Genomic Sciences and Icahn Institute for Genomics and Multiscale Biology, Icahn School of Medicine at Mount Sinai, New York, NY 10029, USA. [2]State Key Laboratory of Environmental Chemistry and Ecotoxicology, Research Center for Eco-Environmental Sciences, Chinese Academy of Sciences, Beijing 100085, China. [3]Department of Neurosurgery and Oncological Sciences, Icahn School of Medicine at Mount Sinai, New York, NY 10029, USA. [4]Black Family Stem Cell Institute, Icahn School of Medicine at Mount Sinai, New York, NY 10029, USA. [5]Sema4, a Mount Sinai Venture, Stamford, CT 06902, USA. [6]Center for Advanced Biotechnology and Medicine, Rutgers University, New Brunswick, NJ 08854, USA.
***Corresponding author. Email: gang.fang@mssm.edu**
†These authors contributed equally to this work.

A method for quantitative 6mA deconvolution

Existing SMRT sequencing–based methods for modification detection require a reference genome, as they compare the interpulse duration (IPD) associated with a base of interest in the native DNA to the expected IPD value estimated according to the base and its flanking DNA sequence in the provided reference genome (*25*, *29*, *30*). Within this design, only those sequencing reads that map to the provided reference genome are analyzed for 6mA, ignoring potential bacterial contamination, which is known to have abundant 6mA events.

To help solve this problem, we took a metagenomic approach. First, in contrast to existing methods that depend on a reference genome for IPD analysis, we took a reference-free approach by using the circular consensus sequence (CCS, a feature of SMRT sequencing for error correction) of an individual DNA molecule as its molecule-specific reference for IPD analysis (*23*, *25*, *31*) (Fig. 1A), thus examining all the sequenced genetic contents for 6mA analysis. We designed relatively short SMRT insert libraries of 200 to 400 base pairs (fig. S1A) (*31*) so that each DNA molecule could be sequenced for a large number of passes (mean, 272×; median, 181×; Fig. 1A and fig. S1B), which facilitated a CCS base calling accuracy of >99.84% (Phred score 28; fig. S2) (*31*) and enabled reliable IPD analysis on single molecules (Fig. 2, A and B). We then used a metagenomic approach to map the CCS reads to a comprehensive collection of genomes (*31*) and performed 6mA quantification (described below) separately for each subgroup of genetic contents in a gDNA sample: species of interest, genomic regions of interest, and sources of contamination.

The current standard method to detect 6mA from SMRT sequencing is based on a defined cutoff on a modification quality value (QV; essentially a transformed *P* value) (*3*, *28*, *31*, *32*). Because QV varies markedly over sequencing depth or number of CCS passes on individual molecules (Fig. 2C) (*28*, *30*), a fixed cutoff can create false positive 6mA calls, especially from genomic regions with high sequencing depth (e.g., mitochondrial genomes). We built on a critical observation of linear increase (slope ~1.7 for 6mA events) of QV over CCS passes (better separation from nonmethylated adenines at higher coverages; Fig. 2, C and D) and developed a machine learning model for 6mA quantification from QV values calculated in the reference-free single-molecule IPD analysis. The core idea was to train the machine learning model across a wide range of 6mA/A levels (training datasets described below) and to use the model to predict 6mA/A levels of newly sequenced gDNA samples based on the collective QV distribution instead of an arbitrary QV cutoff (Fig. 2D) (*31*).

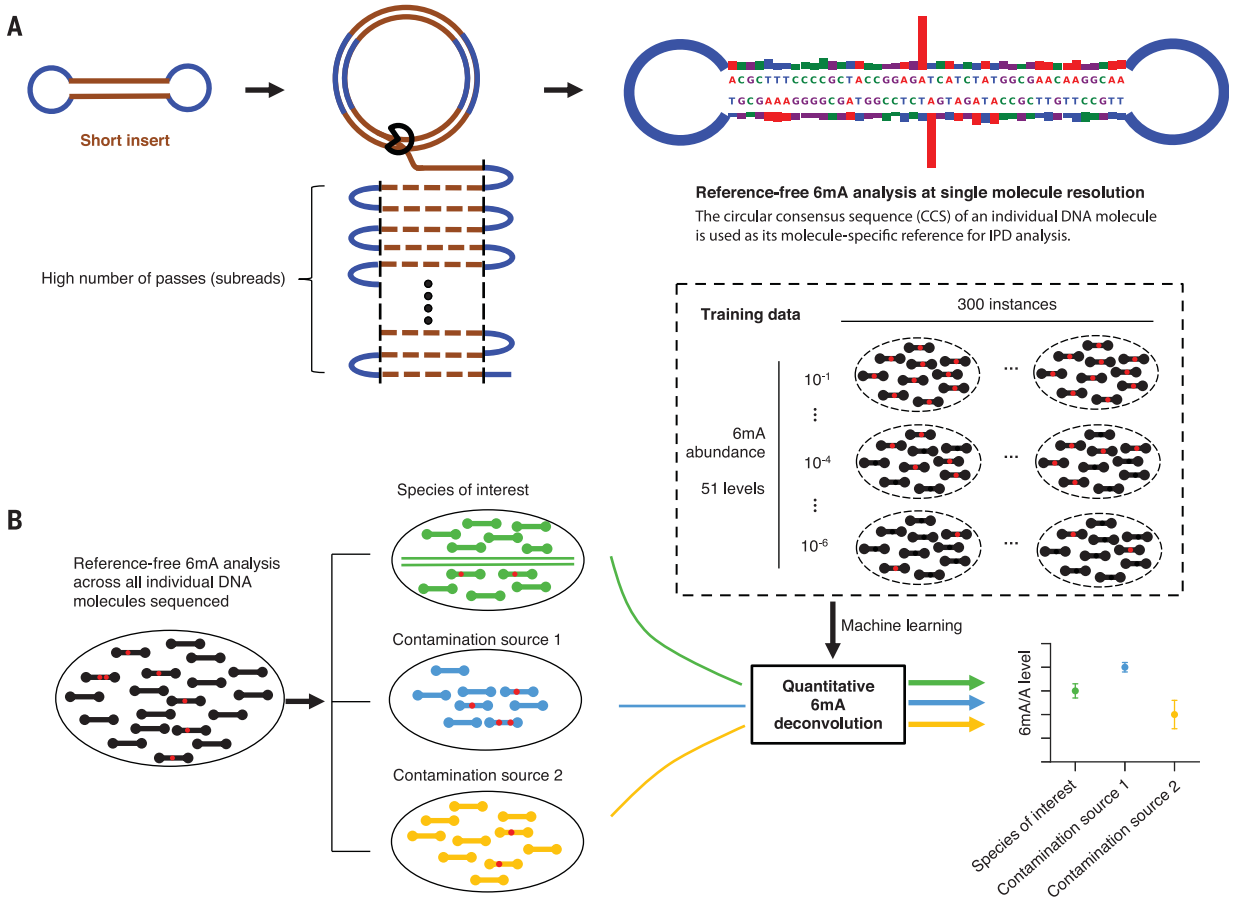

Fig. 1. Overview of 6mASCOPE for quantitative 6mA deconvolution. (**A**) Reference-free 6mA analysis of single molecules. Each molecule (short insert) is sequenced for a large number of passes (subreads). The subreads are combined to a circular consensus sequence (CCS), serving as the molecule-specific reference for in silico IPD estimation, and they provide repeated measures of IPD values for 6mA analysis (*31*). Blue segment denotes SMRT adapter. (**B**) After single-molecule 6mA analysis (a red dot indicates a 6mA event), CCSs (black rods) from a sequenced gDNA sample are separated into the eukaryotic genome (green) and contamination sources (blue and yellow). The 6mA/A levels of each species (or genomic region) are estimated using a machine learning model trained across a wide range of 6mA abundance, with defined confidence intervals.

We constructed high-quality benchmark datasets for the machine learning model training. For 6mA negative controls, we used HEK-WGA [whole-genome amplification of human embryonic kidney (HEK)–293 cell gDNA, 6mA/A level < 10^{-6} by ultrahigh-performance liquid chromatography–tandem mass spectrometry (UHPLC-MS/MS)], HEK293 (native gDNA, 6mA/A level <10^{-6} by UHPLC-MS/MS), and HEK-WGA-MsssI (CpG sites in vitro methylated using a 5mC methyltransferase, MsssI), with the latter two representing the influence of 5mC events on IPD (*16*, *25*, *31*). These samples were each methylated in vitro using three bacterial 6mA methyltransferases (Dam, GATC; TaqI, TCGA; and EcoRI, GAATTC) to create three positive controls: HEK-WGA-3M, HEK293-3M, HEK-WGA-MsssI-3M (fig. S3). By mixing negative and positive controls in silico at different ratios, we created a wide range of 6mA/A levels (10^{-1} to 10^{-6}) for the model training (Fig. 2E) (*31*). Using leave-one-out cross-validation, we compared several models (fig. S4) and selected Random Forest. Our model showed reliable quantification of 6mA/A levels with defined 95% confidence intervals (CIs; Fig. 2F and fig. S5) (*31*). CI depends on both 6mA/A level and number of CCS reads (Fig. 2F and fig. S5B) (*31*), which facilitated dataset-specific CI estimation along with 6mA quantification.

In contrast to existing methods (table S1), 6mASCOPE takes a metagenomic approach and specifically quantifies 6mA events in eukaryotic genomes over contamination, because CCS reads, grouped by species (or specific genomic regions), are separately quantified for 6mA/A levels. For validation, we applied 6mASCOPE on a series of in vitro mixed *E. coli*, *Helicobacter pylori*, and *Saccharomyces cerevisiae* samples with a wide range of 6mA/A levels (10^{-2} to 10^{-6} by UHPLC-MS/MS) and found that 6mASCOPE reliably deconvolved different sources into expected ratios along with stable 6mA quantification (fig. S6).

High-resolution insights of 6mA deposition in two protozoans

Although previous studies reported enrichment of 6mA events in the linkers near transcription start sites (TSSs) in two protozoans, *C. reinhardtii* and *T. thermophila* (*4*, *5*), it remains unclear which specific regions within the linkers are enriched for 6mA events. We sequenced both organisms using the SMRT method and obtained 862,205 and 975,050 CCS reads, respectively, for single-molecule 6mA analysis (table S2) (*31*). We first verified that 6mA has a periodic pattern inversely correlated with nucleosomes near TSSs (fig. S7) (*31*). Next, by dividing genomic regions between the nucleosome dyad and the middle of each nucleosome linker into 10 bins (*31*) and quantifying 6mA/A levels in each bin using 6mASCOPE, we found that

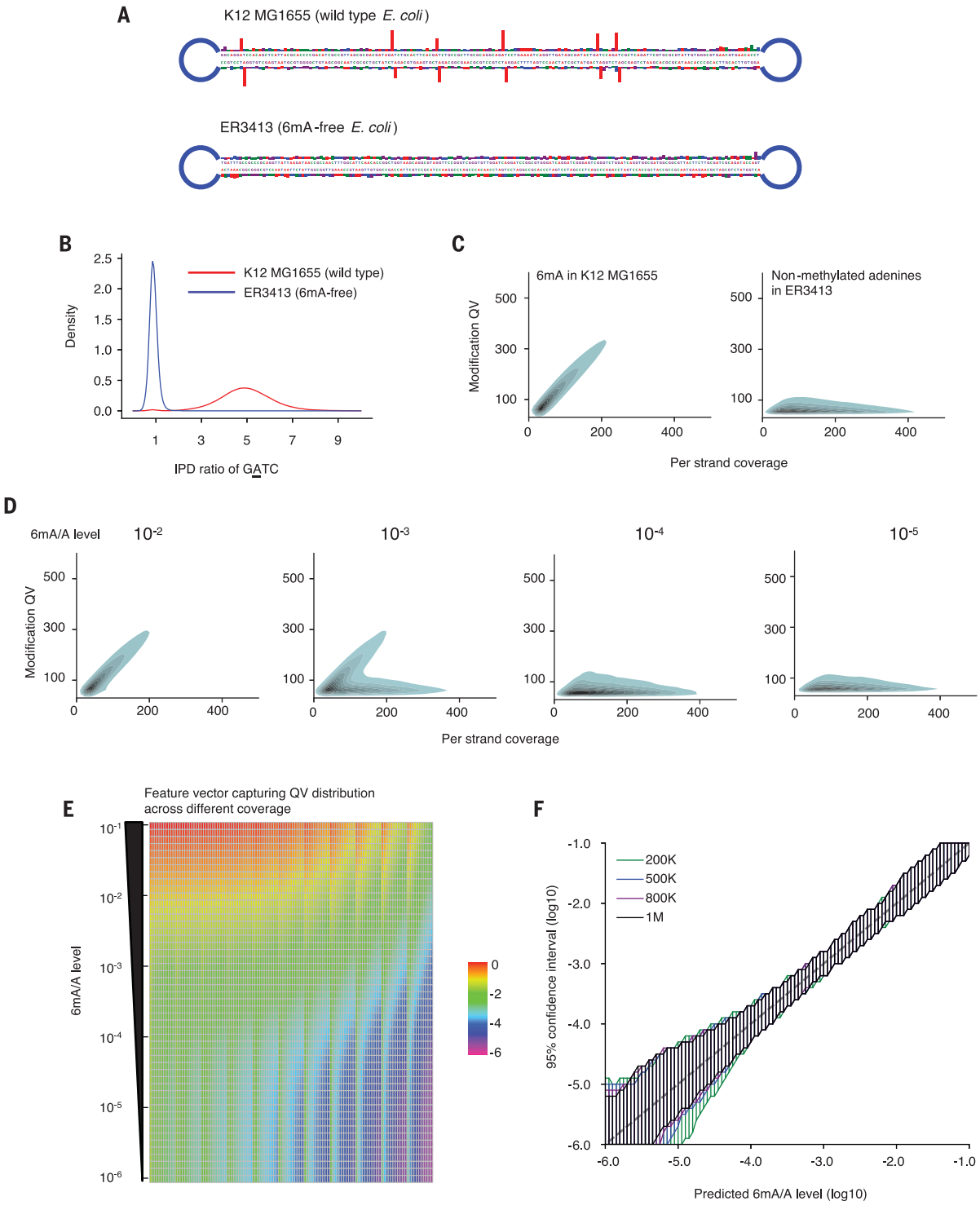

Fig. 2. 6mASCOPE method evaluation. (**A**) IPD ratios (the mean IPD in the native sample divided by the IPD expected from the in silico model) on illustrative molecules from *E. coli* wild-type strain K12 MG1655 and 6mA-free strain ER3413. Blue segment denotes SMRT adapter. (**B**) IPD ratio of adenines on GATC motif in *E. coli* K12 MG1655 and ER3413. 6mA events have IPD ratios of ~5; nonmethylated adenines have IPD ratios of ~1. (**C**) Modification quality values (QVs) of 6mA linearly deviate from the nonmethylated adenines (slope ~1.7), with better separation at high numbers of CCS passes. For illustration, kernel density estimation of adenines with QC > 50 is shown. Left: 6mA in GATC, GCACNNNNNNGTT, and AACNNNNNNTGC from *E. coli* K12 MG1655. Right: Nonmethylated adenines in *E. coli* ER3413. (**D**) QV distribution varies across different 6mA/A levels. (**E**) Feature vectors used for machine learning model training. In each row, one of 51 6mA/A levels (10^{-1} to 10^{-6}) is constructed by mixing negative and positive controls in silico at different ratios. Each column represents the percentage (averaged across 300 replicates, $\log_{10}$-transformed) of adenines over a number of slopes across CCS pass numbers 20 to 240, divided into 11 bins (*31*). (**F**) For each 6mA quantification (*x* axis), 6mASCOPE also provides the 95% confidence interval (*y* axis) (*31*). Colors represent the number of CCS reads used for 6mA quantification.

6mA was enriched at the nucleosome-linker boundaries in *C. reinhardtii* (Fig. 3, A and D) instead of at the middle of the linkers, as previously reported. In contrast, 6mA/A levels of *T. thermophila* increased from the nucleosome boundaries to the middle of linkers (Fig. 3, A and E, and fig. S8). We further used 6mASCOPE to examine the enrichment of 6mA across different motifs. For *C. reinhardtii*, we confirmed that 6mA is enriched in the VATB motif (Fig. 3B; V = A, C, or G; B = C, G, or T) and is essentially absent in non-VATB motifs; for *T. thermophila*, although 6mA was reported to be enriched across the NATN motif (*5*), our 6mASCOPE analysis revealed that VATB sites have a higher 6mA/A level than TATN and NATA sites by a factor of 2 to 3 (Fig. 3C).

6mA from commensal bacteria contribute to most 6mA events in insect and plant samples

A previous study quantified 6mA in *D. melanogaster* using UHPLC-MS/MS and reported that 6mA/A reaches the peak level of ~700 ppm (parts per million) in ~0.75-hour embryos and falls to ~10 ppm at later stages such as adult tissues (*8*). We first collected the fly embryo sample at ~0.75 hours and got

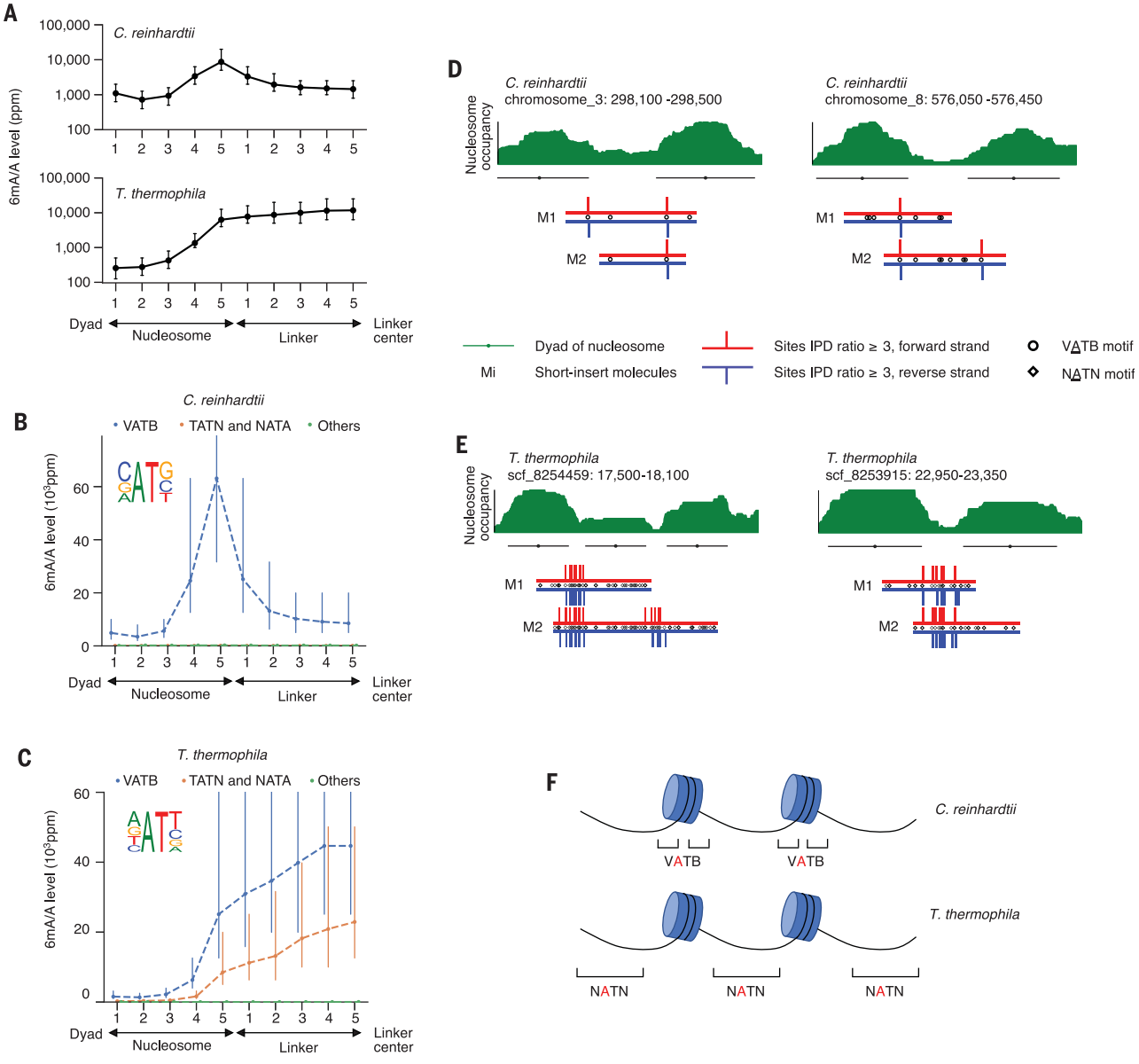

Fig. 3. 6mASCOPE reveals high-resolution 6mA deposition in *C. reinhardtii* and *T. thermophila*. (**A**) 6mA deposition relative to nucleosomes and linkers in *C. reinhardtii* and *T. thermophila*. Genomic regions between the nucleosome dyad and the linker center are divided into 10 bins (*x* axis) across the genome. The 6mA/A level (*y* axis) was quantified with 6mASCOPE. Error bars are 95% CIs. (**B**) 6mA is enriched in the VATB motif at nucleosome-linker boundaries in *C. reinhardtii*. Adenines in each bin are divided into three groups: VATB, TATN/NATA, and others. The dashed line indicates the trend of 6mA/A levels from nucleosome dyad to linker center; *x* and *y* axes are the same as in (A). Error bars are 95% CIs. (**C**) 6mA is enriched across the NATN motif at linkers in *T. thermophila*. (**D** and **E**) Illustrative examples of 6mA enrichment in *C. reinhardtii* (D) and *T. thermophila* (E). Nucleosome occupancy (green stack) is based on MNase-seq data (*31*). Nucleosomes (green lines) and dyads (green dots) are determined by iNPS (v1.2.2). SMRT CCS reads (*Mi*) are shown with red (forward strand) and blue (reverse strand) lines. IPD ratios of 3 or higher are shown. (**F**) Schematic of 6mA enrichment at the nucleosome-linker boundaries in *C. reinhardtii* and the gradual 6mA increase from nucleosome boundaries to linker centers in *T. thermophila*.

674,650 SMRT CCS reads for single-molecule 6mA analysis (table S2). Despite strict measures to avoid contamination (*31*), we found that 96.12% of the CCS reads mapped to the *D. melanogaster* genome reference, whereas 3.88% of the CCS reads mapped to a few microbes (Fig. 4A). Specifically, the contamination reads came from *S. cerevisiae* (1.65%), the major food source of *Drosophila* (*33*), and two genera of bacteria, *Acetobacter* (0.86%) and *Lactobacillus* (0.23%), the main gut commensal bacteria of *D. melanogaster* (*34*). We separately quantified 6mA/A levels in the *D. melanogaster* genome and in each contamination source and found that the level of 6mA/A in total gDNA was 100 ppm (CI, 50 to 200 ppm, consistent with the ~121 ppm UHPLC-MS/MS estimate), 2 ppm in *D. melanogaster* (CI, 1 to 10 ppm), 2 ppm in *Saccharomyces* (CI, 1 to 10 ppm), 5495 ppm in *Acetobacter* (CI, 3162 to 10,000 ppm), 977 ppm in *Lactobacillus* (CI, 501 to 1995 ppm), and 7413 ppm in Others (including additional bacterial genera and unannotated sequences; CI, 3981 to 12,589 ppm) (Fig. 4B and fig. S9) (*31*). Despite their relatively low abundance (3.88%), bacteria contributed to most of the 6mA events in the total gDNA (Fig. 4C). In *Acetobacter*, we observed a high-confidence bacterial 6mA motif (GANTC) (Fig. 4B), consistent with the REBASE database (*35*). The 6mA/A level of 2 ppm (CI, 1 to 10 ppm) estimated for *D. melanogaster*, in contrast to the ~700 ppm previously reported, only explains 1.44% of the total 6mA events in the gDNA sample (considering taxonomy abundances; Fig. 4C).

We next applied 6mASCOPE to examine a *D. melanogaster* adult sample (whole animal), which showed very different microbiome composition with extremely low bacteria contamination, yet still no evidence of a high 6mA/A level in *Drosophila* (fig. S10). We also reanalyzed the 6mA DIP-seq data from a previous *D. melanogaster* study (*8*) and found reads that mapped to multiple bacterial genomes. It is also worth noting that N^4-methylcytosine (4mC), another form of DNA methylation prevalent in bacteria, was also detected in CCS reads from *Acetobacter* enriched at GTAC sites (fig. S11), a motif previously reported in *Acetobacter* (*35*). This observation shows that 4mC analysis for eukaryotic organisms also should be cautiously examined for possible bacterial contamination.

In addition to insects, we hypothesized that soil bacteria can confound 6mA analysis in plants. We applied 6mASCOPE to *A. thaliana* 21-day-old seedlings (*31*), which were reported as having ~2500 ppm 6mA/A by LC-MS/MS (*9*). Among the total 535,030 SMRT CCS reads for single-molecule 6mA analysis, 98.52% could be mapped to the *A. thaliana* genome (Fig. 4D). Among the other 1.48% (subgroup Others), 24.12% were annotated and classified (using Kraken2) into several phyla: Proteobacteria

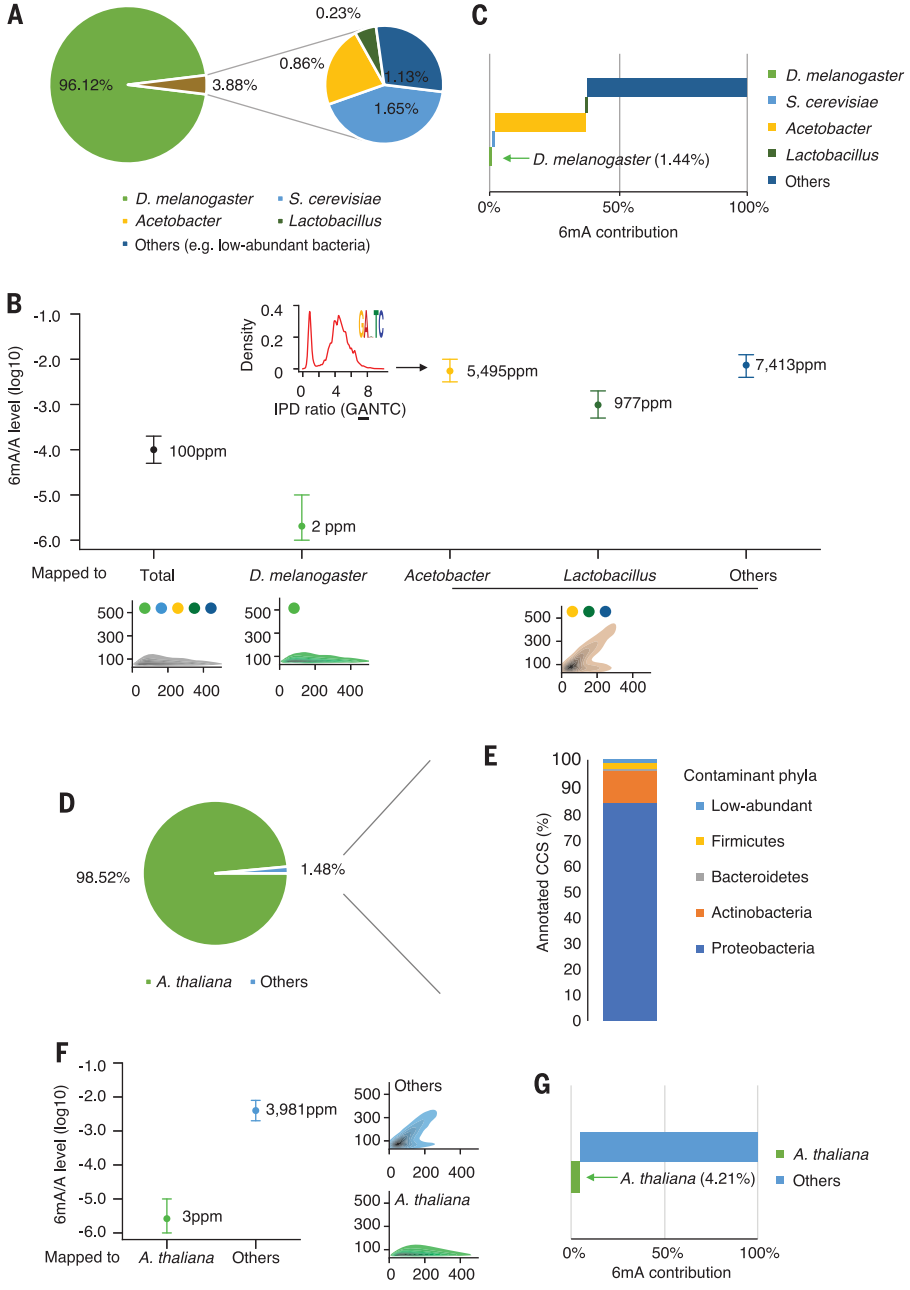

Fig. 4. 6mASCOPE analyses show that commensal bacteria contribute to the vast majority of 6mA events in insect and plant samples. (**A**) Taxonomic compositions (percent) in the *D. melanogaster* embryo ~0.75-hour gDNA sample. CCS reads mapped to *Acetobacter* or *Lactobacillus* are summarized by genus. (**B**) 6mA quantification of the *D. melanogaster* genome and contaminations. For each subgroup, 6mA/A levels are quantified by 6mASCOPE (error bars are 95% CIs). QV distributions are shown at bottom (colored dots refer to species/genus colors in main panel). 6mA/A level of *S. cerevisiae* is further examined with additional sequencing (fig. S9). CCS reads from *Acetobacter*, *Lactobacillus*, and Others (e.g., low-abundant bacteria) are grouped together because CCS read counts within each subgroup are low; CIs are defined on the basis of 8000 CCS reads. Arrow denotes the density of IPD ratios in the GANTC motif in *Acetobacter*. (**C**) 6mA contribution (percent) from each subgroup in the *D. melanogaster* embryo sample. (**D** and **E**) Taxonomic compositions (percent) in the *A. thaliana* 21-day seedling gDNA sample. The CCS reads in subgroup "Others" (D) are classified with Kraken2. Main classes of Proteobacteria are shown in fig. S12. (**F**) 6mA quantification of the *A. thaliana* genome and the contamination (Others). (**G**) 6mA contribution (percent) from each subgroup in the *A. thaliana* seedling sample.

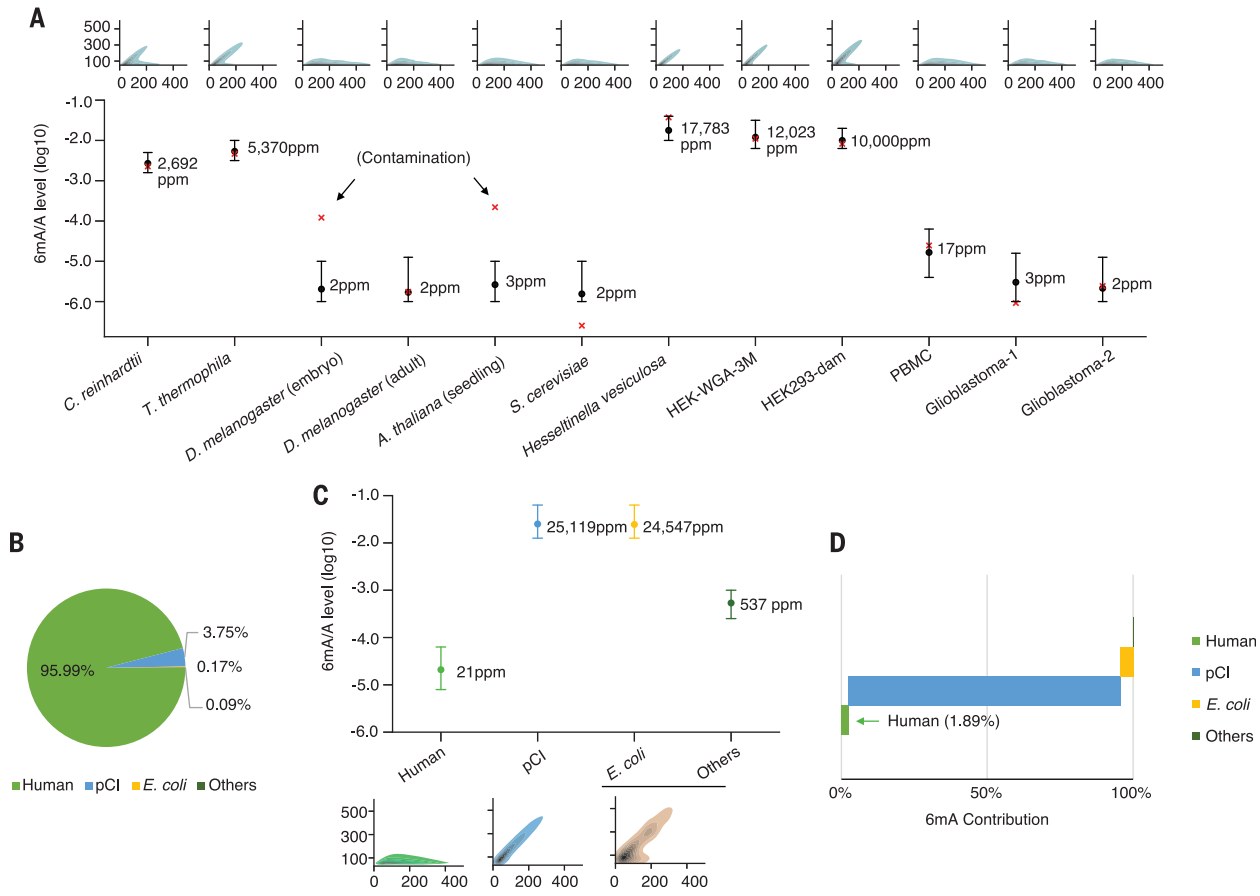

Fig. 5. 6mASCOPE-based quantitative deconvolution across multiple human gDNA samples. (**A**) 6mA/A levels on the genome of interest quantified by 6mASCOPE (error bars are 95% CIs). The 6mA/A level in *S. cerevisiae* is consistent with independent UHPLC-MS/MS measurement (0.3 ppm, lower than the minimum 6mA/A level used in the 6mASCOPE training dataset). Except for *D. melanogaster* embryo and *A. thaliana* gDNA samples (both are contaminated by bacteria), 6mA/A levels by 6mASCOPE are consistent with UHPLC-MS/MS (red cross). For all samples except HEK-WGA-3M and HEK293-dam, the UHPLC-MS/MS is performed independently using the same batch of gDNA samples. For HEK-WGA-3M and HEK293-dam, the UHPLC-MS/MS estimates are mimicked: Nearly all the expected motif(s) are methylated in vitro by the methyltransferase(s). The QV distribution for each gDNA sample is shown at the top. (**B**) Sources (percent) of CCS reads in the HEK-pCI sample (transfection of an empty pCI plasmid into HEK 293 cells). (**C**) 6mA quantification (percent) of different sources in HEK-pCI. CCS reads from *E. coli* and Others are grouped together, and their CIs are determined on the basis of 8000 CCS reads. (**D**) 6mA contribution (percent) from the subgroups in the HEK-pCI sample.

(fig. S12), Actinobacteria, Bacteroidetes, and Firmicutes. These phyla and classes (Fig. 4E and fig. S12) are consistent with *A. thaliana* root microbiome (*36*). Using 6mASCOPE, we separately quantified 6mA/A levels for *A. thaliana* (3 ppm; CI, 1 to 10 ppm) and Others (3981 ppm; CI, 1995 to 7943 ppm) and found that CCS reads mapped to *A. thaliana* contributed to only 4.21% of the total 6mA events in the total gDNA sample (Fig. 4, F and G). Consistently, 6mASCOPE analysis of the *A. thaliana* 21-day-old root sample also demonstrated remarkable microbiome contamination (greater than the seedlings), with a smaller contribution from *A. thaliana* to the total 6mA events (fig. S13).

6mASCOPE finds no evidence of high abundance of 6mA in the human cells examined

We next examined the abundance of 6mA in human cells and tissues. We chose to investigate peripheral blood mononuclear cells (PBMCs), which are composed of 70 to 90% lymphocytes (*37*), because lymphocytes have been shown to have a high 6mA/A level of ~0.051% (510 ppm) (*12*). We also collected and examined two glioblastoma brain tissue samples because glioblastoma stem cells and primary glioblastoma were reported to have a 6mA/A level of ~1000 ppm by dot blotting and mass spectrometry (*11*).

We obtained 570,283, 247,700, and 280,763 SMRT CCS reads from the PBMC sample and the two glioblastoma brain tissues, respectively, for single-molecule 6mA analysis. Of these, 99.53%, 99.88%, and 99.86% of CCS reads were mapped to the human reference genome, indicating highly pure samples. The 6mA/A levels estimated by 6mASCOPE in glioblastoma samples were ~10^{-6}, with 3 ppm for glioblastoma-1 (CI, 1 to 16 ppm) and 2 ppm for glioblastoma-2 (CI, 1 to 13 ppm) (Fig. 5A) (*31*). This level is comparable to the negative controls with extremely low 6mA/A levels: HEK-WGA (1 ppm; CI, 1 to 6 ppm) and native HEK293 (1 ppm; CI, 1 to 6 ppm), when the confidence intervals are taken into consideration. In the PBMC sample, the 6mA/A level estimation of 17 ppm (CI, 4 to 63 ppm) by 6mASCOPE is consistent with the measurements of UHPLC-MS/MS (Fig. 5A). These data suggested either that the abundance of 6mA, if present in glioblastoma and PBMCs, was much lower than the reported levels in the recent studies (glioblastoma, ~1000 ppm; lymphocytes, ~510 ppm) or that 6mA/A levels may be highly heterogeneous or variable between different samples of the same cell type, the same tissue, or a specific disease. Motif enrichment analysis did not support a reliable motif in these samples (fig. S14).

Across all the samples examined in this study, we observed largely consistent 6mA/A

level estimates between 6mASCOPE and UHPLC-MS/MS (Fig. 5A) except the *D. melanogaster* embryo and *A. thaliana* samples, for which the much higher 6mA/A estimates by UHPLC-MS/MS were due to bacterial contamination (Fig. 4), highlighting the capability and reliability of 6mASCOPE. In addition to 6mA quantification of individual species, our method was also able to quantify 6mA/A levels in specific genomic regions of interest. Previous studies have reported enrichment of 6mA in mitochondrial DNA (mtDNA) (*12*, *13*, *21*, *38*) and in young full-length LINE-1 elements (L1s) (*10*, *11*, *21*). For mtDNA, 6mASCOPE did not find 6mA enrichment in the 7205 CCS reads from the HEK293 sample that mapped to mtDNA, in comparison to a negative control (targeted mitochondrial genome amplification, $10^{-5.72}$; CI, $10^{-6.00}$ to $10^{-4.90}$; fig. S15). For L1 elements, although 6mASCOPE appeared to suggest a higher 6mA/A level in the young full-length L1s than in older L1s, a further comparison with a WGA negative control did not support 6mA enrichment in young L1 elements (fig. S16), highlighting the importance of using negative controls to capture possible uncharacterized biases (*14*, *39*). This result was consistent with our previous study of human lymphoblastoid cells, in which increased IPD patterns exist not only in adenines but also in cytosines, guanines, and thymines of young L1 elements, which suggested confounding factors such as secondary structure (*14*).

Plasmids used for genetic manipulation can carry confounding bacterial-origin 6mA

Genetic manipulation is commonly used in epigenetic research to characterize putative methyltransferases and demethylases. *E. coli* is often used as a host for plasmid selection and expansion. As a result, the plasmids can contain 6mA events written by bacterial methyltransferase(s) and can confound 6mA study in eukaryotic cells.

To illustrate this, we transfected an empty pCI plasmid vector from *E. coli* into HEK293 cells, following the standard lipofection-based protocol (*31*). Total gDNA harvested at 72 hours after transfection was sequenced using SMRT technology and analyzed using 6mASCOPE. Among the 741,558 CCS reads, 95.99% were mapped to the human genome and 3.75% came from the pCI vector (Fig. 5B), and the remaining 0.26% of CCS reads (Fig. 5B) included reads that mapped to the *E. coli* genome (*31*), implying possible carryover of gDNA from *E. coli* to the HEK293 cells during transfection. By separately quantifying the 6mA/A level in each subgroup, pCI showed a high 6mA/A level of $10^{-1.60}$ (25,119 ppm), about the same as *E. coli* (Fig. 5C). Considering its abundance, pCI contributed to 93.91% of the total 6mA events in this post-transfection HEK293 total gDNA (Fig. 5, C and D). Hence, genetic manipulation experiments involving plasmids may confound the characterization of putative 6mA methyltransferases and demethylases. Although the use of methylation-free bacteria as the host for plasmid preparation can avoid this type of contamination, it is worth noting that the Dam methyltransferase mutant *E.coli*, previously used in a few studies (*7*, *38*), still has substantial 6mA events because of the remaining 6mA methyltransferase hsdM (*2*, *28*) (fig. S17, based on 6mASCOPE analysis). We therefore suggest the use of *E. coli* strains with both Dam and hsdM deleted as the plasmid host.

Discussion

Our study cannot exclude the potential presence of authentically high levels of 6mA/A in multicellular eukaryotes in certain samples that we did not examine here. However, our results suggest that a reassessment of 6mA across eukaryotic genomes, using 6mASCOPE to quantitatively estimate the confounding impact of bacterial contamination, is warranted. To facilitate the broad use of 6mASCOPE, we have released a detailed experimental protocol and an automated software package on Zenodo (*40*) and GitHub.

We caution that plasmid 6mA contamination, even from Dam methyltransferase mutant *E. coli*, is possible during genetic manipulation and may have confounded previous characterizations of 6mA enzymes. Lipofection or electroporation, which is used to transfect plasmid DNA directly into the target cells, is more likely to introduce contamination, whereas lentiviral transduction would be less affected if the original plasmids are completely removed during viral packaging.

Our 4mC result suggests that similar caution should be exercised when studying 4mC in eukaryotes by means of SMRT sequencing, which has found 4mC in several eukaryotes [see (*41*)], despite SMRT sequencing being prone to making false positive calls (*16*), especially given the lack of evidence for 4mC in mice even when ultrasensitive UHPLC-MS/MS is used (*19*). More broadly, this study will also help to guide rigorous technological development for the detection of other forms of rare DNA and RNA modifications.

Our study has a few limitations: (i) The focus of 6mASCOPE is more about quantitatively deconvolving the global 6mA/A level into different species and genomic regions of interests, rather than mapping specific 6mA events in a particular genome. We prioritized this focus because the most controversial 6mA findings to date were those reporting high 6mA/A levels in multicellular eukaryotes. The precise mapping of specific 6mA events in a particular genome would require deeper SMRT sequencing and can be pursued in future work. (ii) For reliable data interpretation, it is important to combine the 6mA/A levels estimated by 6mASCOPE with their confidence intervals, which depend on sequencing depth. However, even with a large number of CCS reads, 6mASCOPE does not precisely differentiate 6mA/A levels below 10 ppm because the confidence interval includes 1 ppm, which is the lowest 6mA/A level in our training dataset (Fig. 2F) (*31*). (iii) Two recent studies reported that ribo-m6A on mRNA can be a source of 6mA on DNA via the nucleotide-salvage pathway (*17*, *18*). 6mA events that are misincorporated via this pathway cannot be distinguished from other 6mA events by SMRT sequencing or 6mASCOPE, and isotope labeling coupled with LC/MS-MS is needed instead (*17*). (iv) For each gDNA sample, the CCS reads analyzed by 6mASCOPE only represent the DNA molecules that were sequenced by SMRT sequencing. Although SMRT DNA polymerases can effectively sequence through diverse genomic regions with very complex secondary structures (*42*), it might miss some DNA molecules with certain unknown properties. (v) Although 6mASCOPE enables quantitative 6mA deconvolution, it could be confounded by other DNA modifications that indirectly influence SMRT DNA polymerase kinetics of adenines or flanking bases (*3*, *25*, *30*), so we suggest combining LC/MS-MS and 6mASCOPE for 6mA quantification and deconvolution of eukaryotic gDNA samples.

REFERENCES AND NOTES

1. M. A. Sánchez-Romero, J. Casadesús, *Nat. Rev. Microbiol.* **18**, 7–20 (2020).
2. G. Fang *et al.*, *Nat. Biotechnol.* **30**, 1232–1239 (2012).
3. J. Beaulaurier, E. E. Schadt, G. Fang, *Nat. Rev. Genet.* **20**, 157–172 (2019).
4. Y. Fu *et al.*, *Cell* **161**, 879–892 (2015).
5. Y. Wang, X. Chen, Y. Sheng, Y. Liu, S. Gao, *Nucleic Acids Res.* **45**, 11594–11606 (2017).
6. S. J. Mondo *et al.*, *Nat. Genet.* **49**, 964–968 (2017).
7. E. L. Greer *et al.*, *Cell* **161**, 868–878 (2015).
8. G. Zhang *et al.*, *Cell* **161**, 893–906 (2015).
9. Z. Liang *et al.*, *Dev. Cell* **45**, 406–416.e3 (2018).
10. T. P. Wu *et al.*, *Nature* **532**, 329–333 (2016).
11. Q. Xie *et al.*, *Cell* **175**, 1228–1243.e20 (2018).
12. C. L. Xiao *et al.*, *Mol. Cell* **71**, 306–318.e7 (2018).
13. Z. Hao *et al.*, *Mol. Cell* **78**, 382–395.e8 (2020).
14. S. Zhu *et al.*, *Genome Res.* **28**, 1067–1078 (2018).
15. K. Douvlataniotis, M. Bensberg, A. Lentini, B. Gylemo, C. E. Nestor, *Sci. Adv.* **6**, eaay3335 (2020).
16. Z. K. O'Brown *et al.*, *BMC Genomics* **20**, 445 (2019).
17. M. U. Musheev, A. Baumgärtner, L. Krebs, C. Niehrs, *Nat. Chem. Biol.* **16**, 630–634 (2020).
18. X. Liu *et al.*, *Cell Res.* **31**, 94–97 (2021).
19. S. Schiffers *et al.*, *Angew. Chem. Int. Ed.* **56**, 11268–11271 (2017).
20. A. Lentini *et al.*, *Nat. Methods* **15**, 499–504 (2018).
21. C. W. Q. Koh *et al.*, *Nucleic Acids Res.* **46**, 11659–11670 (2018).
22. G. Z. Luo *et al.*, *Nat. Commun.* **7**, 11301 (2016).
23. A. M. Wenger *et al.*, *Nat. Biotechnol.* **37**, 1155–1162 (2019).
24. D. Deamer, M. Akeson, D. Branton, *Nat. Biotechnol.* **34**, 518–524 (2016).
25. B. A. Flusberg *et al.*, *Nat. Methods* **7**, 461–465 (2010).
26. A. Tourancheau, E. A. Mead, X. S. Zhang, G. Fang, *Nat. Methods* **18**, 491–498 (2021).
27. J. Beaulaurier *et al.*, *Nat. Biotechnol.* **36**, 61–69 (2018).
28. M. J. Blow *et al.*, *PLOS Genet.* **12**, e1005854 (2016).
29. J. Beaulaurier *et al.*, *Nat. Commun.* **6**, 7438 (2015).
30. E. E. Schadt *et al.*, *Genome Res.* **23**, 129–141 (2013).
31. See supplementary materials.
32. P. H. Oliveira *et al.*, *Nat. Microbiol.* **5**, 166–180 (2020).
33. J. Murgier, C. Everaerts, J. P. Farine, J. F. Ferveur, *Sci. Rep.* **9**, 8873 (2019).
34. W. J. Lee, P. T. Brey, *Annu. Rev. Cell Dev. Biol.* **29**, 571–592 (2013).

35. R. J. Roberts, T. Vincze, J. Posfai, D. Macelis, *Nucleic Acids Res.* **43**, D298–D299 (2015).
36. D. S. Lundberg *et al.*, *Nature* **488**, 86–90 (2012).
37. C. P. Corkum *et al.*, *BMC Immunol.* **16**, 48 (2015).
38. C. Ma *et al.*, *Nat. Cell Biol.* **21**, 319–327 (2019).
39. W. M. Guiblet *et al.*, *Genome Res.* **28**, 1767–1778 (2018).
40. Y. Kong, L. Cao, G. Deikus, Y. Fan, E. Mead, W. Lai, Y. Zhang, R. Yong, R. Sebra, H. Wang, X.-S. Zhang, G. Fang, Code and processed data for "Critical assessment of DNA adenine methylation in eukaryotes using quantitative deconvolution" (2022). doi:10.5281/zenodo.5838427
41. P. Ye *et al.*, *Nucleic Acids Res.* **45**, D85–D89 (2017).
42. E. W. Loomis *et al.*, *Genome Res.* **23**, 121–128 (2013).

ACKNOWLEDGMENTS

We thank P. Hegemann (Humboldt University of Berlin) for the *C. reinhardtii* strains, H. D. Madhani (University of California, San Francisco) for the *S. cerevisiae* strain, F. Wang and B. Yao (Emory University) for the *D. melanogaster* embryos and adults, J. Dong (Rutgers University) for the *A. thaliana* strains, J. Mo (Chinese Academy of Sciences) for help with UHPLC-MS/MS analysis, members of the Fang lab for helpful discussions, and the Department of Scientific Computing at the Icahn School of Medicine at Mount Sinai for computational resources and staff expertise. **Funding:** Supported by the Icahn Institute for Genomics and Multiscale Biology, NIH grants R35 GM139655, R01 HG011095, and R56 AG071291, the Irma T. Hirschl/Monique Weill-Caulier Trust, and the Nash Family Foundation (G.F.). UHPLC-MS/MS analyses of 6mA were supported by Strategic Priority Research Program of the Chinese Academy of Sciences grant XDPB2004 and National Natural Science Foundation of China grant 22021003 (H.W.). **Author contributions:** G.F. conceived the study and supervised the research; Y.K. and G.F. developed the 6mASCOPE method; Y.K. performed all the computational analyses; Y.K., L.C., E.A.M., X.-S.Z., and G.F. designed the experiments; L.C., E.A.M., and X.-S.Z. performed most of the experiments; G.D. and R.S. optimized short-insert PacBio library preparation and performed all PacBio sequencing; Y.F. performed raw PacBio sequencing data processing and quality control; W.L. and H.W. performed the UHPLC-MS/MS analysis; Y.Z. and R.Y. performed glioblastoma sample preparation; X.-S.Z. assisted the characterization of bacterial strains and collected *A. thaliana* samples; Y.K., L.C., Y.F., E.A.M., X.-S.Z., and G.F. analyzed the data; and Y.K. and G.F. wrote the manuscript with additional information inputs from other co-authors. **Competing interests:** Y.K. and G.F. are the co-inventors of a pending patent application based on the method described in this work. **Data and materials availability:** All sequencing data generated in this study have been submitted to NCBI with accession number PRJNA667898. The software supporting all proposed methods is available along with a tutorial at Zenodo (*40*) and at GitHub, www.github.com/fanglab/6mASCOPE.

SUPPLEMENTARY MATERIALS

science.org/doi/10.1126/science.abe7489
Materials and Methods
Supplementary Text
Figs. S1 to S17
Tables S1 and S2
References (*43*–*56*)
MDAR Reproducibility Checklist

29 September 2020; resubmitted 8 September 2021
Accepted 7 December 2021
10.1126/science.abe7489

Critical assessment of DNA adenine methylation in eukaryotes using quantitative deconvolution

Yimeng KongLei CaoGintaras DeikusYu FanEdward A. MeadWeiyi LaiYizhou ZhangRaymund YongRobert SebraHailin WangXue-Song ZhangGang Fang

Science, 375 (6580), • DOI: 10.1126/science.abe7489

Reassessment of DNA 6mA in eukaryotes

Certain forms of chemical modifications to DNA play important roles across the kingdoms of life; some forms have been widely studied and others are relatively new. DNA N-methyldeoxyadenosine (6mA), which was recently reported to be prevalent across eukaryotes, created excitement for a new dimension to study biology and diseases. However, some studies have highlighted confounding factors, and there is an active debate over 6mA in eukaryotes. Kong *et al*. describe a method for quantitative 6mA deconvolution and report that bacterial contamination explains the vast majority of 6mA in DNA samples from insects and plants; the method also found no evidence for high 6mA levels in humans (see the Perspective by Boulias and Greer). This work advocates for a reassessment of 6mA in eukaryotes and provides an actionable approach. —DJ