Semi-conservative transmission of DNA N^6-adenine methylation in a unicellular eukaryote

Yalan Sheng[1,2]†‡, Yuanyuan Wang[1,2]†, Wentao Yang[3]†, Xue Qing Wang[3], Jiuwei Lu[4], Bo Pan[1,2], Bei Nan[1,2], Yongqiang Liu[1,2], Fei Ye[1,2], Chun Li[5], Jikui Song[4], Yali Dou[3], Shan Gao[1,2*], Yifan Liu[3*]

[1]MOE Key Laboratory of Evolution and Marine Biodiversity and Institute of Evolution and Marine Biodiversity, Ocean University of China, Qingdao 266003, China

[2]Laboratory for Marine Biology and Biotechnology, Qingdao Marine Science and Technology Center, Qingdao 266237, China

[3]Department of Biochemistry and Molecular Medicine, Keck School of Medicine, University of Southern California, Los Angeles, CA 90033, USA

[4]Department of Biochemistry, University of California Riverside, Riverside, CA 92521, USA

[5]Division of Biostatistics, Department of Preventive Medicine, University of Southern California, Los Angeles, CA 90033, USA

†These authors contributed equally to this work

‡Present address: Guangzhou Key Laboratory of Subtropical Biodiversity and Biomonitoring, Guangdong Provincial Key Laboratory for Healthy and Safe Aquaculture, School of Life Sciences, South China Normal University, Guangzhou 510631, China

*Correspondence: shangao@ouc.edu.cn (S.G.); Yifan.Liu@med.usc.edu (Y.L.)

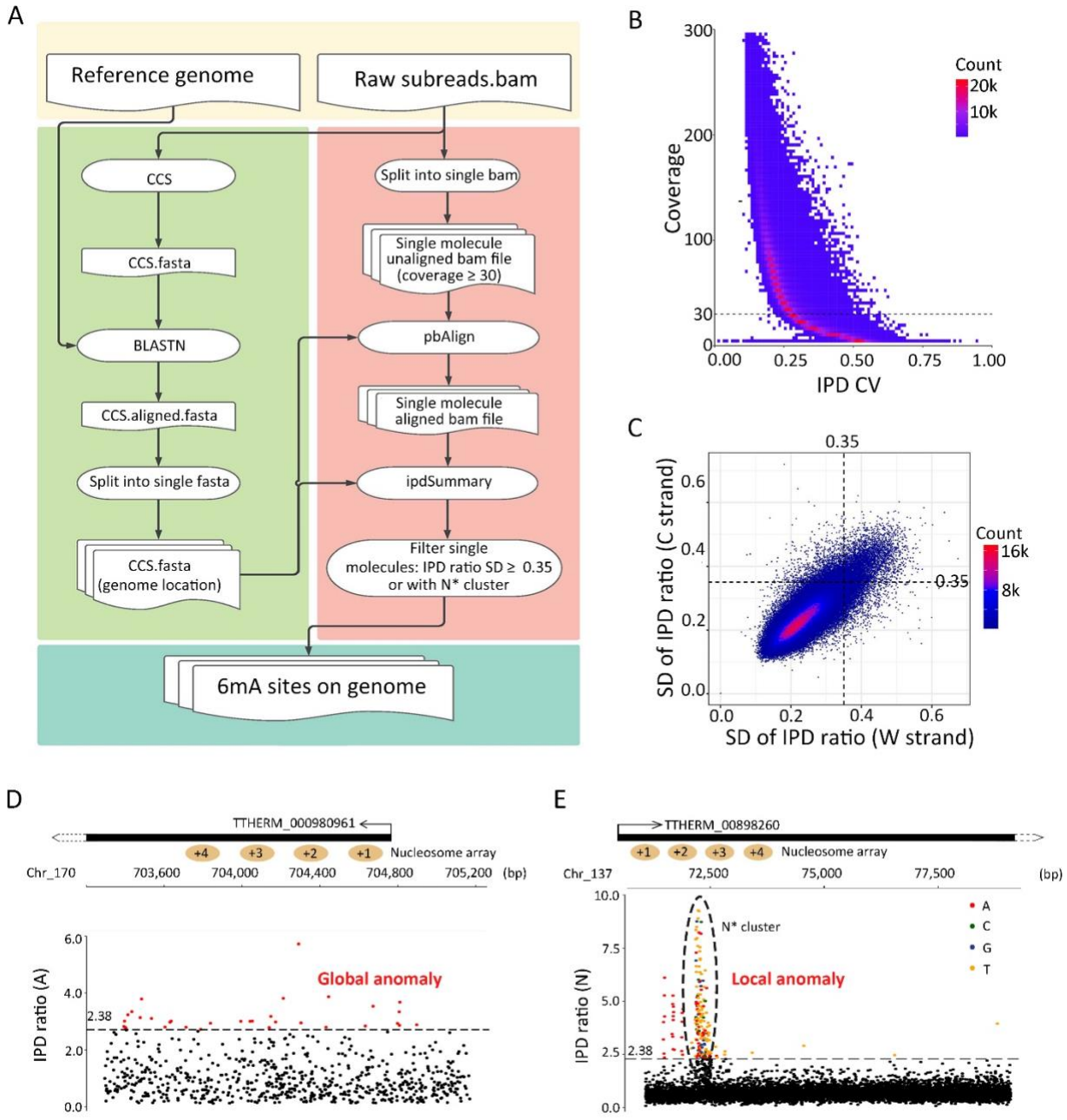


Figure S1. 6mA detection by SMRT CCS.

1. 6mA detection by SMRT CCS.

   A. The bioinformatic pipeline for 6mA calling. See methods for details.

   B. Relationship between IPD coefficients of variance (CV=$\frac{tError}{tMean}$, averaged for all adenine sites of a SMRT CCS read) and effective coverage (number of passes for CCS, averaged for all adenine sites in the same read). Read distribution density (count) is plotted as a heat map. A 30× threshold for effective coverage is used to remove reads with noisy IPD values.

   C. Average standard deviation (SD) of IPD ratios (IPDr) for SMRT CCS reads with high effective coverage (≥30×). SD values are calculated for IPDr of unmodified adenine sites from W and C, respectively. Read distribution density (count) is plotted as a heat map. The SD threshold (≤0.35), applied to both W and C, is used to remove reads with global anomalies in IPDr.

   D. IPDr for all A sites in a SMRT CCS read with global anomalies.

   E. IPDr for all sites in a SMRT CCS read with local anomalies. Note the high IPDr G/C/T sites (as well as A sites) in the N* cluster.

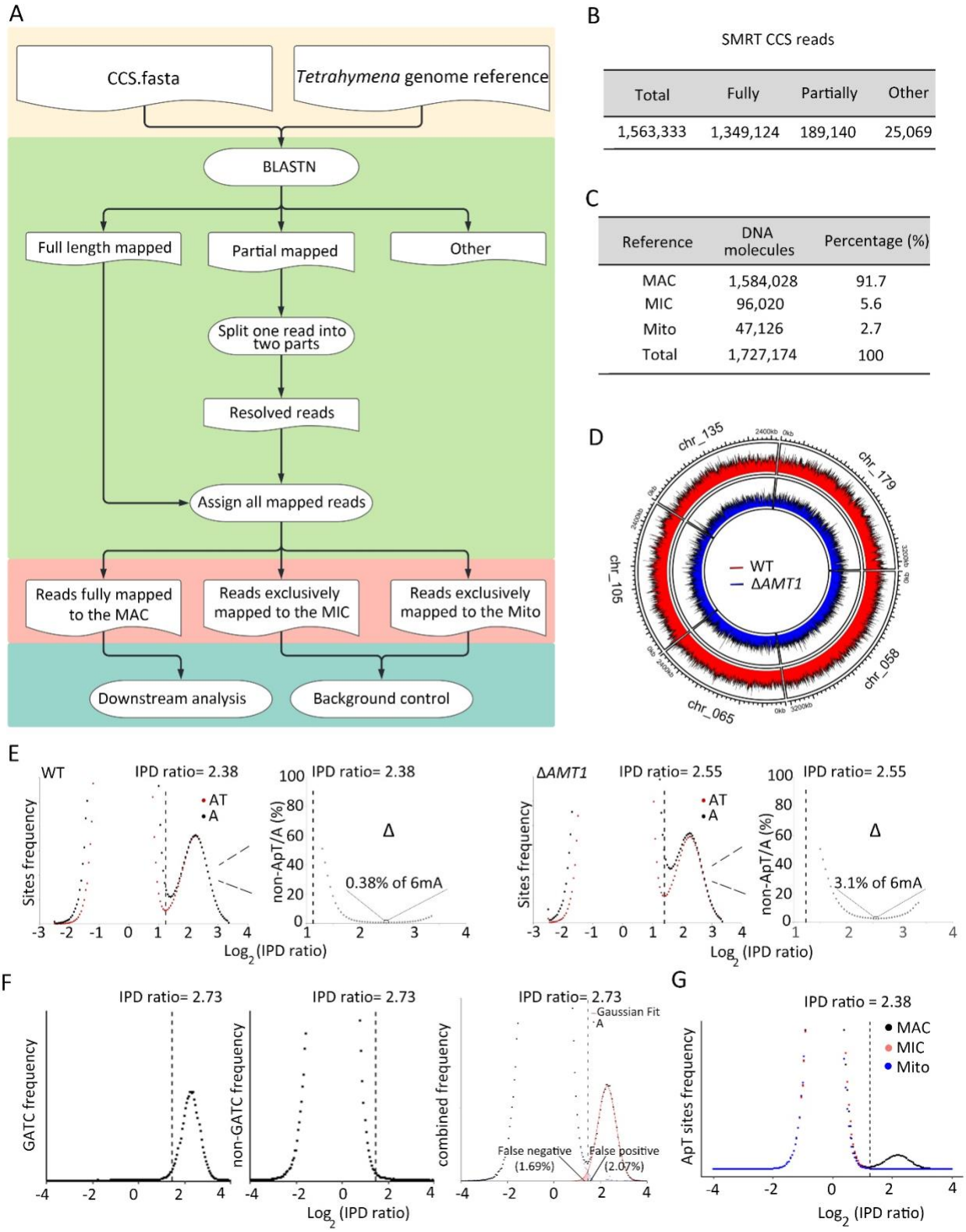
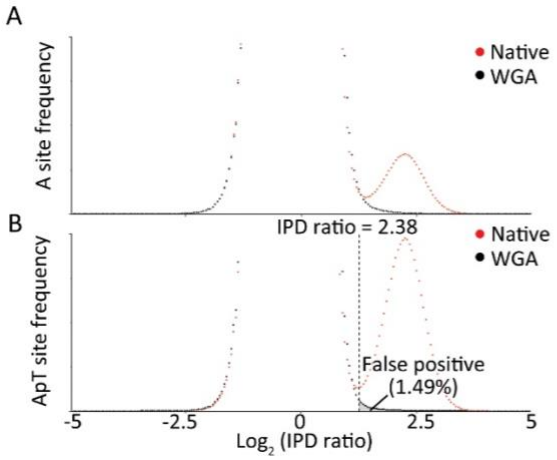


Figure S2. Comparing 6mA in *Tetrahymena* macronucleus (MAC), micronucleus (MIC), and mitochondrion (Mito).

2. Comparing 6mA in *Tetrahymena* macronucleus (MAC), micronucleus (MIC), and mitochondrion (Mito).

   A. The bioinformatic pipeline for mapping CCS reads back to the *Tetrahymena* reference genomes.
   B. Most SMRT CCS reads were either fully or partially mapped to a single locus of *Tetrahymena* reference genomes (MAC, MIC, and Mito) with high confidence. A small percentage of reads were not mapped back with high confidence (Other).
   C. Classification of DNA molecules (after resolving chimeric reads) according to whether they are mapped back to MAC, MIC, or Mito.
   D. Distribution of SMRT CCS reads from WT and Δ*AMT1* cells mapped back to five longest MAC chromosomes. Note the relatively even coverage from telomere to telomere, supporting high quality assembly of these chromosomes.
   E. Predominant, if not exclusive, occurrence of 6mA at the ApT dinucleotide in WT (left) and Δ*AMT1* cells (right). IPDr distributions for all A sites (black) and A sites at the ApT dinucleotide (red) are plotted. The differential between the two curves (Δ), represented as percentage of non-ApT sites relative to all A sites, is also plotted (only to the right of the 6mA calling threshold); the minimum value of the differential curve is indicated.
   F. IPDr distributions for adenines in GATC sites (methylated by *dam*, left), non-GATC sites (not methylated, middle), and all A sites (right) in SMRT CCS reads of plasmid DNA (Abdulhay et al. 2020). In the IPDr distribution of all A sites (right), the 6mA peak and the unmodified A peak are deconvoluted. Note the low false positive and false negative rates of 6mA calling.
   G. IPDr distributions of all A sites in reads mapped back to *Tetrahymena* MAC, MIC, or Mito. The IPDr threshold of 2.38, set for calling 6mApT in MAC, is indicated.



C

Group	Native	WGA
CCS read	1,584,028	686,605
ApT site	992,618,784	385,054,618
6mApT site	18,750,607	115,175
6mApT/ApT (%)	1.89	0.030

Figure S3. Evaluating 6mA background level with whole genome amplification (WGA).

3. Evaluating 6mA background level with whole genome amplification (WGA).

   A. IPDr distributions for all A sites: native *Tetrahymena* genomic DNA (Native: red) and whole genome amplification of *Tetrahymena* genomic DNA (WGA: black). Note that WGA effectively removes all base modifications while preserving the sequence information.
   B. IPDr distributions for A sites at the ApT dinucleotide: native *Tetrahymena* genomic DNA (Native: red) and whole genome amplification of *Tetrahymena* genomic DNA (WGA: black). False positive rate is calculated as the ratio of 6mApT calls in the WGA and Native samples (IPDr threshold: 2.38).
   C. Summary of SMRT CCS results of the Native and WGA samples.

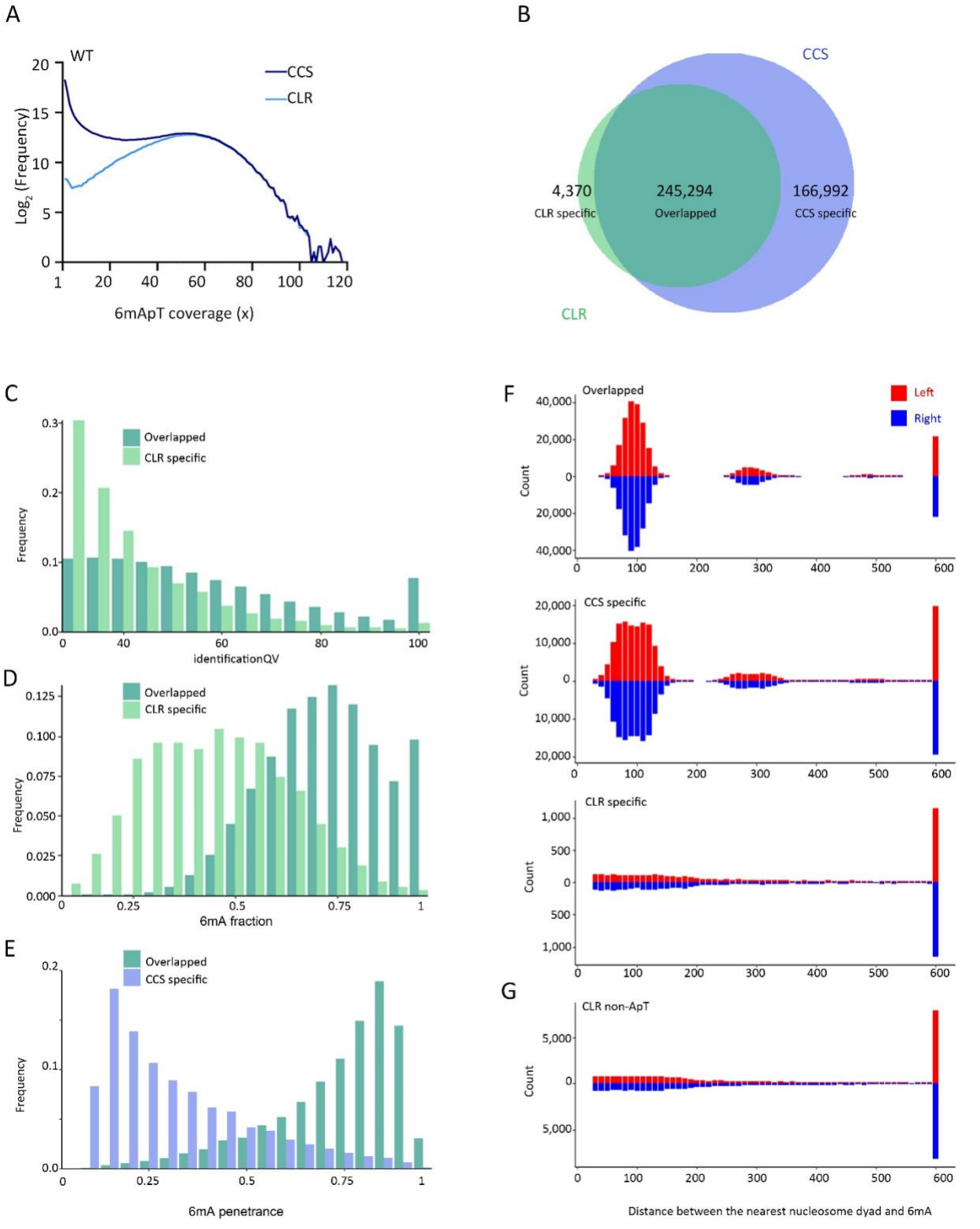
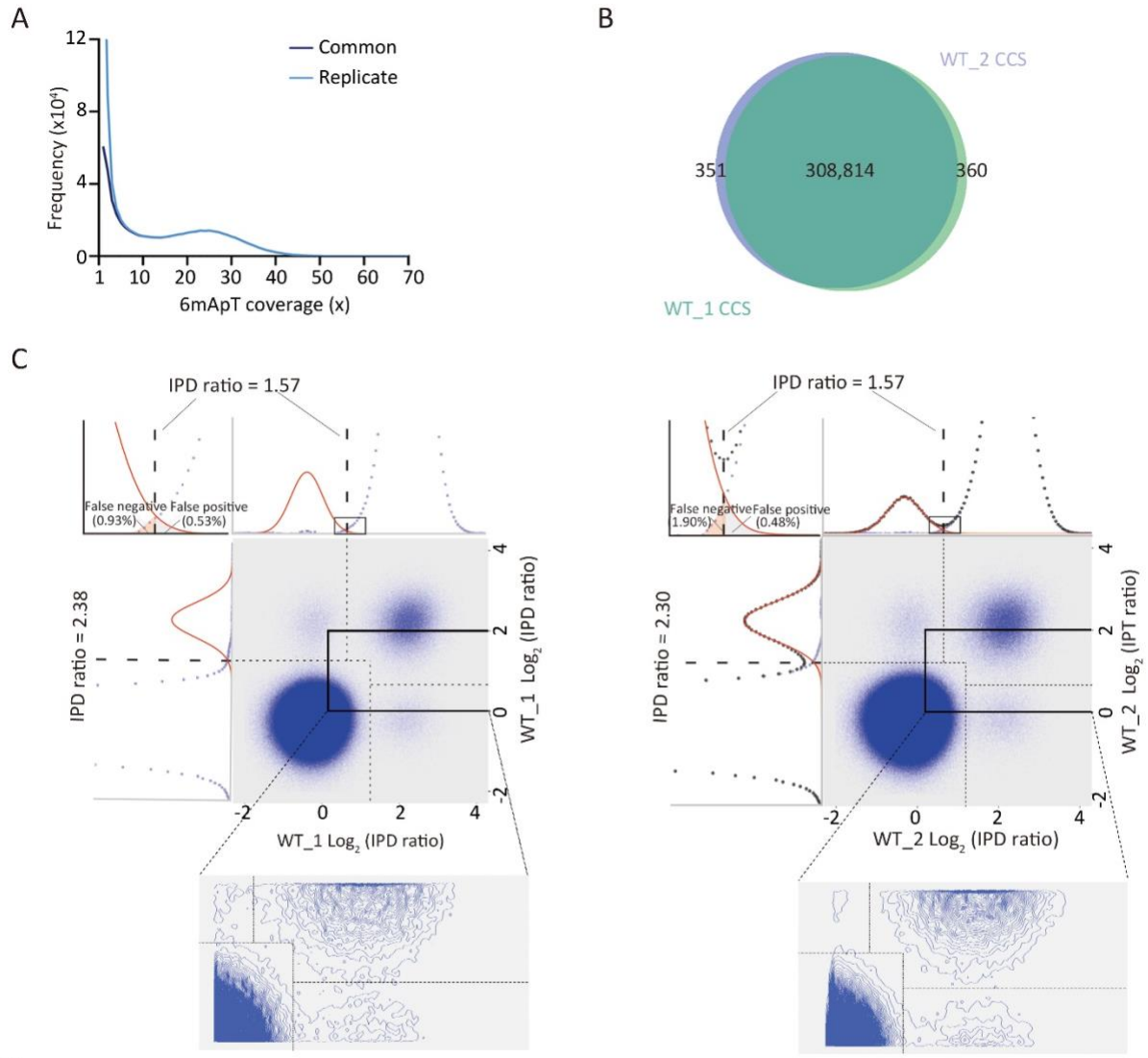


Figure S4. Comparing CCS and CLR results for WT *Tetrahymena* cells.

4. Comparing CCS and CLR results for WT *Tetrahymena* cells.

   A. CCS and CLR results converge at genomic positions with high 6mApT coverage. For the CCS data (dark blue), x-axis: 6mApT coverage for a genomic position (i.e., the number of reads in which 6mApT has been called in this genomic position); y-axis: number of genomic positions (log2) with the 6mApT coverage indicated in x-axis. We examined whether 6mApT genomic positions called by CCS were also called by CLR in our previous study (Wang et al. 2017). For the CLR data (light blue), 6mApT genomic positions called by both CCS and CLR are plotted. At high 6mApT coverage, CLR calls converge with CCS calls.
   B. Overlap between high confidence 6mApT genomic positions called by CLR (identification Qv≥ 30) and CCS (6mApT coverage≥10×).
   C. High confidence 6mApT genomic positions called by both CLR and CCS on average have better identification Qv scores (a CLR metrics) than those called only by CLR.
   D. High confidence 6mApT genomic positions called by both CLR and CCS on average have higher 6mA fractions (a CLR metrics) than those called only by CLR.
   E. High confidence 6mApT genomic positions called by both CLR and CCS on average have higher 6mA penetrance (a CCS metrics) than those called only by CCS.
   F. Distributions of the distance between 6mA and its nearest nucleosome dyad (to 6mA's left or right, plotted separately). Note that high confidence 6mApT genomic positions called by both CLR and CCS show prominent peaks at ~100bp and 300bp, consistent with their enrichment in linker DNA. For CCS-alone genomic positions, a similar, albeit slightly diffusive, pattern emerges. However, CLR-alone genomic positions show no periodic distribution, indicating no linker DNA enrichment.
   G. Distributions of the distance between 6mA and its nearest nucleosome dyad (to 6mA's left or right, plotted separately) for non-ApT genomic positions called only by CLR. There is no periodic distribution, indicating no linker DNA enrichment.



D

Single molecule	WT_1		WT_2	
	Number	Percentage (%)	Number	Percentage (%)
ApT	992,618,784	100	871,635,758	100
6mApT Sites	18,750,607	1.89	17,607,170	2.02
full-6mApT	8,373,723	89.32	7,872,065	89.41
hemi-C-6mApT	998,137	5.32	925,862	5.25
hemi-W-6mApT	1,005,024	5.36	937,178	5.32
Total 6mApT	18,750,607	100	871,635,758	100

Figure S5. Comparing SMRT CCS replicate results for WT *Tetrahymena* cells.

5. Comparing SMRT CCS replicate results for WT *Tetrahymena* cells.

   A. 6mApT genomic positions called in SMRT CCS replicate results converge at high 6mApT coverage. For one of the two replicate datasets (Replicate: light blue, corresponding to WT_1 dataset), x-axis: 6mApT coverage for a genomic position; y-axis: number of genomic positions with the 6mApT coverage indicated in x-axis. We then checked how many of these 6mApT genomic positions were called in both replicate datasets (Common: dark blue); x-axis: for 6mApT genomic positions called in both datasets, the 6mApT coverage in WT_1 dataset; y-axis: number of 6mApT genomic positions called in both datasets, with the corresponding 6mApT coverage indicated in x-axis.
   B. Overlap between 6mApT genomic positions called in SMRT CCS replicate results (6mApT coverage ≥ 10×).
   C. Demarcation of the four methylation states of ApT duplexes in SMRT CCS replicate results. For details, see Fig. 2C: left. Inset: contour plots of the zoomed-in regions, demonstrating that these two IPDr thresholds cut through the valleys separating the four peaks, thus representing an optimal scheme for demarcation.
   D. Statistics for 6mApT sites called in SMRT CCS replicate results.

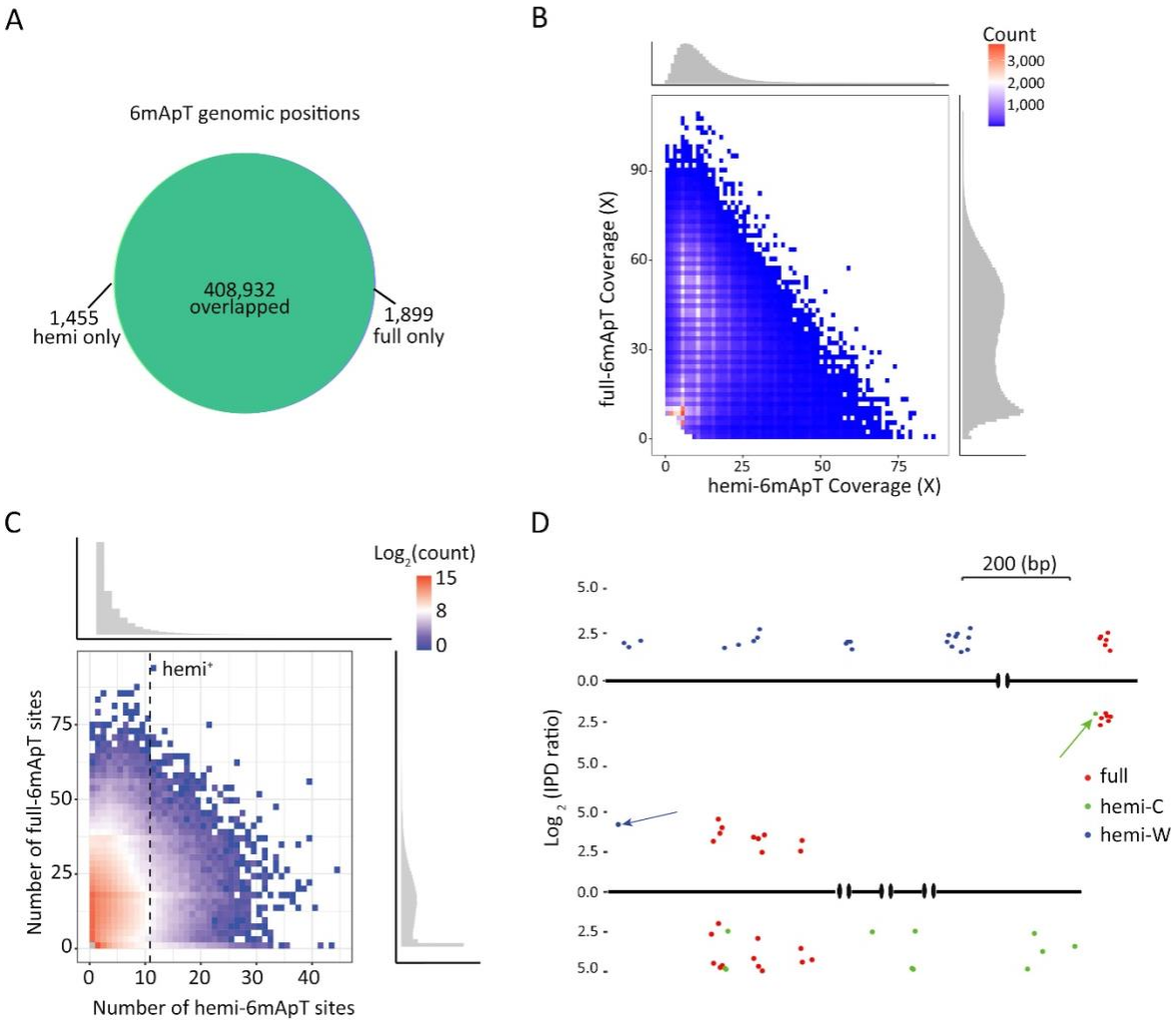
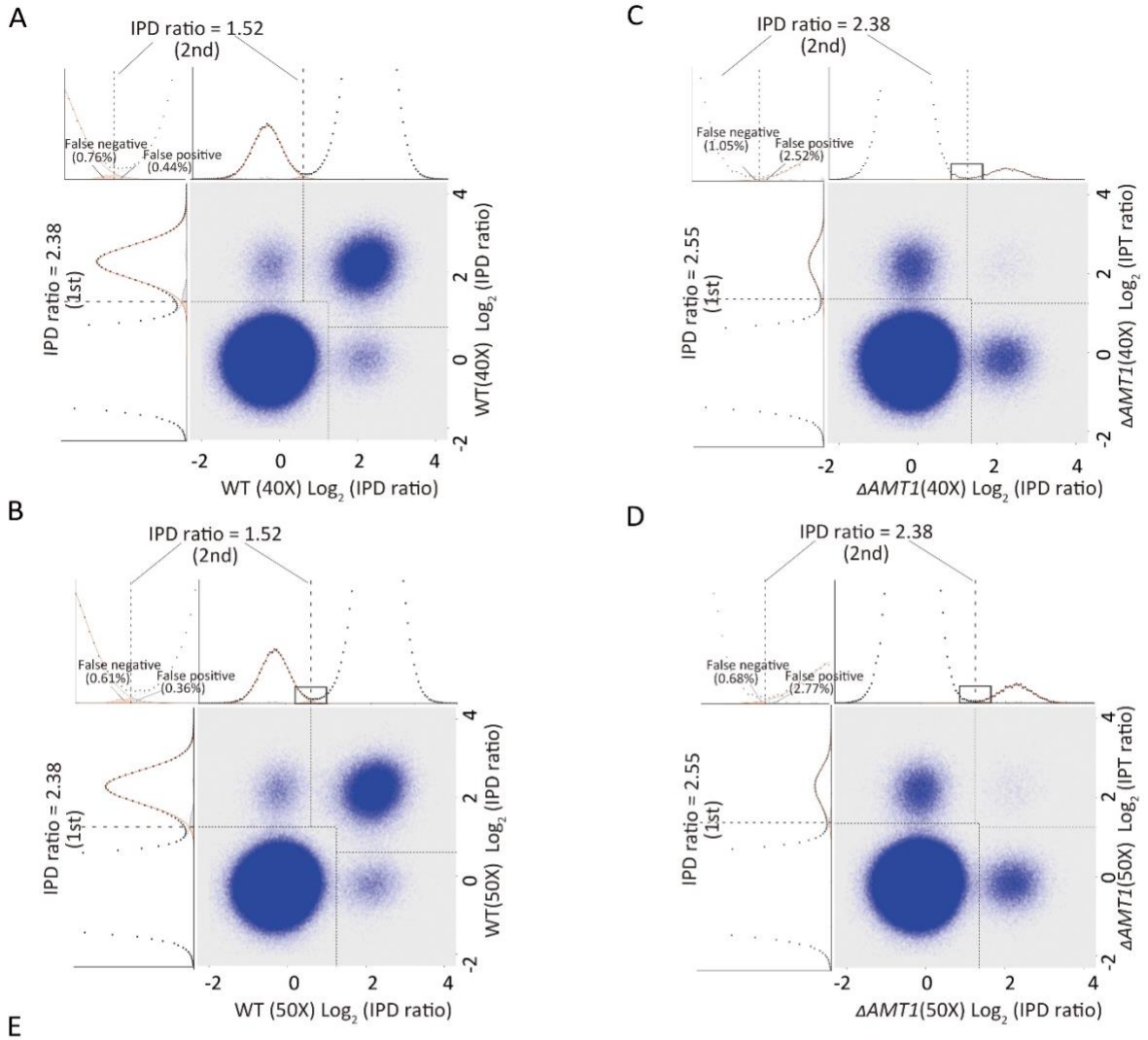


Figure S6. Hemi^{+} molecules in WT *Tetrahymena* cells.

6. Hemi^{+} molecules in WT *Tetrahymena* cells.

   A. 6mApT genomic positions (6mApT coverage ≥10×) with reads supporting hemi and/or full methylation.
   B. Distribution of 6mApT genomic positions (6mApT coverage ≥10×) according to the number of reads supporting hemi-6mApT calls (x-axis: hemi-6mApT coverage) and full-6mApT calls (y-axis: full-6mApT coverage) mapped to these genomic positions.
   C. Distribution of DNA molecules according to the numbers of hemi-6mApT (x-axis) and full-6mApT calls (y-axis) they contain. The distribution density is plotted as a heat map, with the threshold for hemi+ molecules indicated (W+C ≥11, dashed line). The read distributions according to separate counts of full-6mApT (right) and hemi-6mApT (top) are also plotted.
   D. Two hemi^{+} molecules with many hemi-6mApT on one strand, but one hemi-6mApT on the other strand (arrow).



E

Coverage(X)	WT (1st)		*ΔAMT1* (1st)		WT (2nd)		*ΔAMT1* (2nd)	
	FP(%)	FN(%)	FP(%)	FN(%)	FP(%)	FN(%)	FP(%)	FN(%)
30	1.12	1.93	2.84	3.59	0.93	0.53	1.67	4.03
40	0.98	1.81	2.45	2.69	0.76	0.44	1.05	2.52
50	0.84	1.84	2.06	2.55	0.61	0.36	0.68	2.77

F

Coverage(x)	WT		*ΔAMT1*	
	full(%)	hemi(%)	full(%)	hemi(%)
30	89.32	10.68	3.03	96.97
40	89.60	10.40	3.18	96.82
50	89.63	10.37	3.18	96.82

G

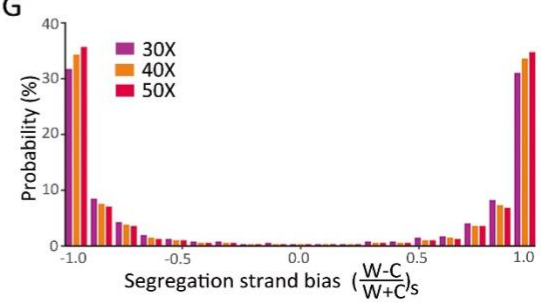
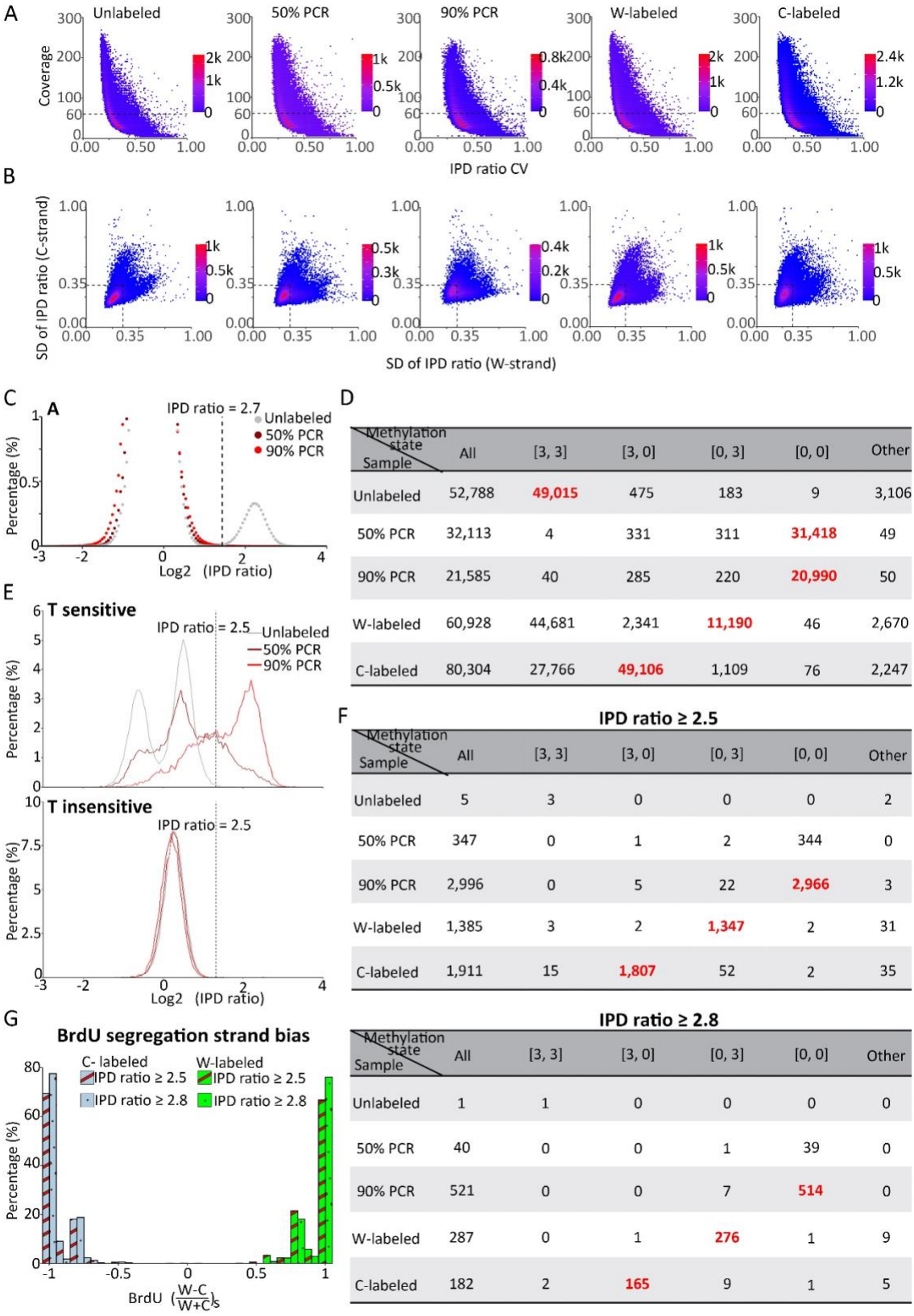


Figure S7. 6mA calling by varying CCS passes.

7. 6mA calling by varying CCS passes.

   A. Demarcation of the four methylation states of ApT duplexes in WT cells and CCS reads with ≥40× passes. For details, see Fig. 2C: left.
   B. Demarcation of the four methylation states of ApT duplexes in WT cells and CCS reads with ≥50× passes. For details, see Fig. 2C: left.
   C. Demarcation of the four methylation states of ApT duplexes in Δ*AMT1* cells and CCS reads with ≥40× passes. For details, see Fig. 2C: right.
   D. Demarcation of the four methylation states of ApT duplexes in Δ*AMT1* cells and CCS reads with ≥50× passes. For details, see Fig. 2C: right.
   E. Comparing the IPDr thresholds (1st and 2nd) for CCS reads with ≥30×, ≥40×, and ≥50× passes.
   F. Comparing the hemi-6mApT and full-6mApT percentages for CCS reads with ≥30×, ≥40×, and ≥50× passes.
   G. Comparing the segregation strand bias distribution for CCS reads with ≥30×, ≥40×, and ≥50× passes.



D

Sample \ Methylation state	All	[3, 3]	[3, 0]	[0, 3]	[0, 0]	Other
Unlabeled	52,788	**49,015**	475	183	9	3,106
50% PCR	32,113	4	331	311	**31,418**	49
90% PCR	21,585	40	285	220	**20,990**	50
W-labeled	60,928	44,681	2,341	**11,190**	46	2,670
C-labeled	80,304	27,766	**49,106**	1,109	76	2,247

F

IPD ratio ≥ 2.5

Sample \ Methylation state	All	[3, 3]	[3, 0]	[0, 3]	[0, 0]	Other
Unlabeled	5	3	0	0	0	2
50% PCR	347	0	1	2	344	0
90% PCR	2,996	0	5	22	**2,966**	3
W-labeled	1,385	3	2	**1,347**	2	31
C-labeled	1,911	15	**1,807**	52	2	35

IPD ratio ≥ 2.8

Sample \ Methylation state	All	[3, 3]	[3, 0]	[0, 3]	[0, 0]	Other
Unlabeled	1	1	0	0	0	0
50% PCR	40	0	0	1	39	0
90% PCR	521	0	0	7	**514**	0
W-labeled	287	0	1	**276**	1	9
C-labeled	182	2	**165**	9	1	5

Figure S8. In vitro BrdU-labeling.

8. In vitro BrdU-labeling.

   A. Relationship between IPD coefficients of variance (CV=$\frac{tError}{tMean}$, averaged for all guanine sites of a SMRT CCS read) and effective coverage (number of passes for CCS, averaged for all guanine sites in the same read). Read distribution density (count) is plotted as a heat map. A 60× threshold for effective coverage is used to remove reads with noisy IPD values.

   B. Average standard deviation (SD) of IPDr for reads with high effective coverage (≥60×). SD values are calculated for IPDr of guanine sites from W and C, respectively. Read distribution density (count) is plotted as a heat map. The SD threshold (≤0.35), applied to both W and C, is used to remove reads with global anomalies in IPDr.

   C. IPDr distributions of A sites in the ApT dinucleotide from filtered reads of unlabeled and W&C-labeled DNA (50% and 90% BrdUTP, respectively). The IPDr threshold of 2.7 for calling 6mApT is indicated.

   D. Comparing reads with different methylation states at the three GATC sites. [3, 3]: all full methylation; [3, 0]: all hemi-W; [0, 3]: all hemi-C; [0, 0]: all unmethylated; Other: mixed methylation states. Our analyses were focused on target groups with expected methylation states (red).

   E. IPDr distributions of sensitive (top) and insensitive T positions (bottom). IPDr values are from filtered reads of unlabeled and W&C-labeled DNA (50% and 90% BrdUTP PCR, respectively). The IPDr threshold of 2.5 for calling BrdU is indicated.

   F. Counts of $BrdU^+$ molecules. BrdU calling threshold was set at 2.5 (top) and 2.8 (bottom), respectively. $BrdU^+$ molecules were defined as DNA molecules with no less than 8 BrdU sites on one strand (W||C ≥ 8). Our analyses were focused on target groups with expected methylation states (red).

   G. Segregation strand biases of BrdU sites in $BrdU^+$ molecules. Segregation strand bias for BrdU was defined as the difference-sum ratio between BrdU sites on W and C: $(\frac{W-C}{W+C})_S$. For both W-labeled and C-labeled samples, BrdU exhibited strong segregation bias toward the newly synthesized strand under either BrdU calling threshold (IPDr=2.5 or 2.8).

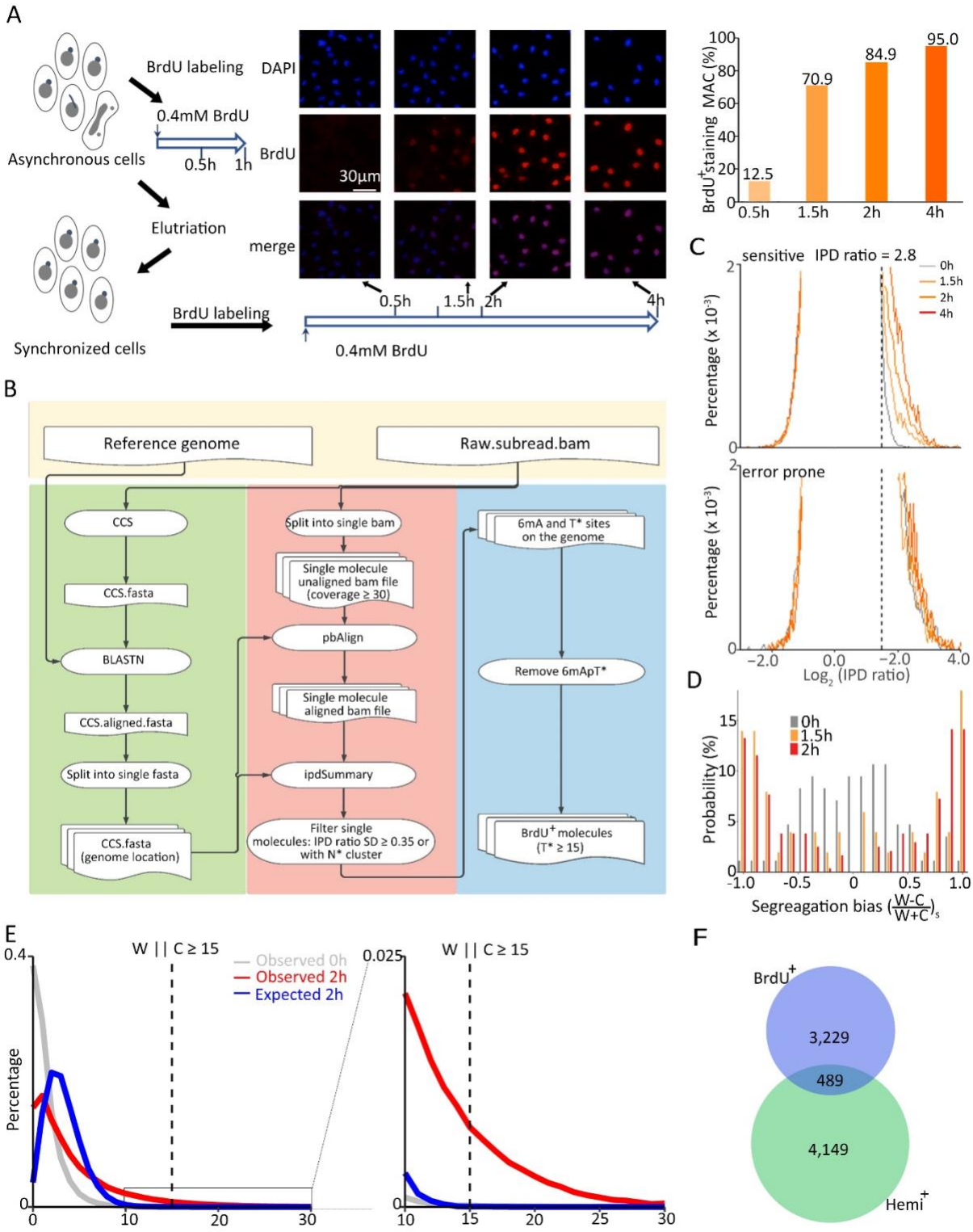


Figure S9. In vivo BrdU-labeling.

9. In vivo BrdU-labeling.

   A. BrdU-labeling of *Tetrahymena* cells. *Tetrahymena* cells were synchronized at G1 phase by centrifugal elutriation and released for growth in the fresh medium with 0.4mM BrdU. *Tetrahymena* samples were taken at 0.5h, 1.5h, 2h, and 4h of labeling. Immunofluorescence (IF) staining showed that the percentage of cells with BrdU signals increased at 1.5h and plateaued at 2h, indicative of highly synchronous progression through S phase.

   B. The bioinformatic pipeline for calling $BrdU^+$ molecules.

   C. IPDr distribution of T sites in the high copy number rDNA of synchronized *Tetrahymena* cells with BrdU-labeling (1.5h, 2h, and 4h) or without (0h). The IPDr threshold for calling BrdU was set at 2.8. Top: T sites mapped to sensitive genomic positions with a more confined IPDr distribution in the unlabeled sample and substantial increases in IPDr in the BrdU-labeled samples, as a result featuring low false positive rates. Bottom: T sites mapped to error-prone genomic positions with a more dispersive IPDr distribution in the unlabeled sample and marginal increases in IPDr in the BrdU-labeled samples, as a result featuring high false positive rates.

   D. Segregation strand biases for BrdU calls in $BrdU^+$ molecules mapped to rDNA. $BrdU^+$ molecules are defined as DNA molecules with a total count of no less than 15 BrdU sites (W+C≥15). Segregation strand bias for BrdU is defined as the difference-sum ratio between BrdU sites on W and C: $(\frac{W-C}{W+C})_S$. Error-prone genomic positions were removed before the tally. Note the strong segregation strand bias for BrdU-labeled samples (1.5h and 2h), but not the unlabeled sample (0h).

   E. Distribution of DNA molecules in genomic DNA samples of synchronized *Tetrahymena* cells with BrdU-labeling (Observed 2h) or without (Observed 0h), according to the number of BrdU calls they contain in one strand (pick the strand with the higher number). For simulation of the background noise (Expected 2h), BrdU sites in the BrdU-labeled sample were randomly redistributed among T sites in all DNA molecules, and the distribution was recalculated. The threshold for calling $BrdU^+$ molecules (with no less than 15 BrdU sites on one strand: W||C ≥15)

is indicated. Note the scarcity of $BrdU^+$ molecules in the distributions of Observed 0h and Expected 2h, in contrast to the substantial presence of $BrdU^+$ molecules in Observed 2h. This is consistent with clustering of BrdU sites at the levels of individual DNA molecules and one of their strands, and a more random distribution of the background noise.

F. Overlap between $hemi^+$ and $BrdU^+$ molecules, from G1-synchronized cells labeled by BrdU for 1.5h.

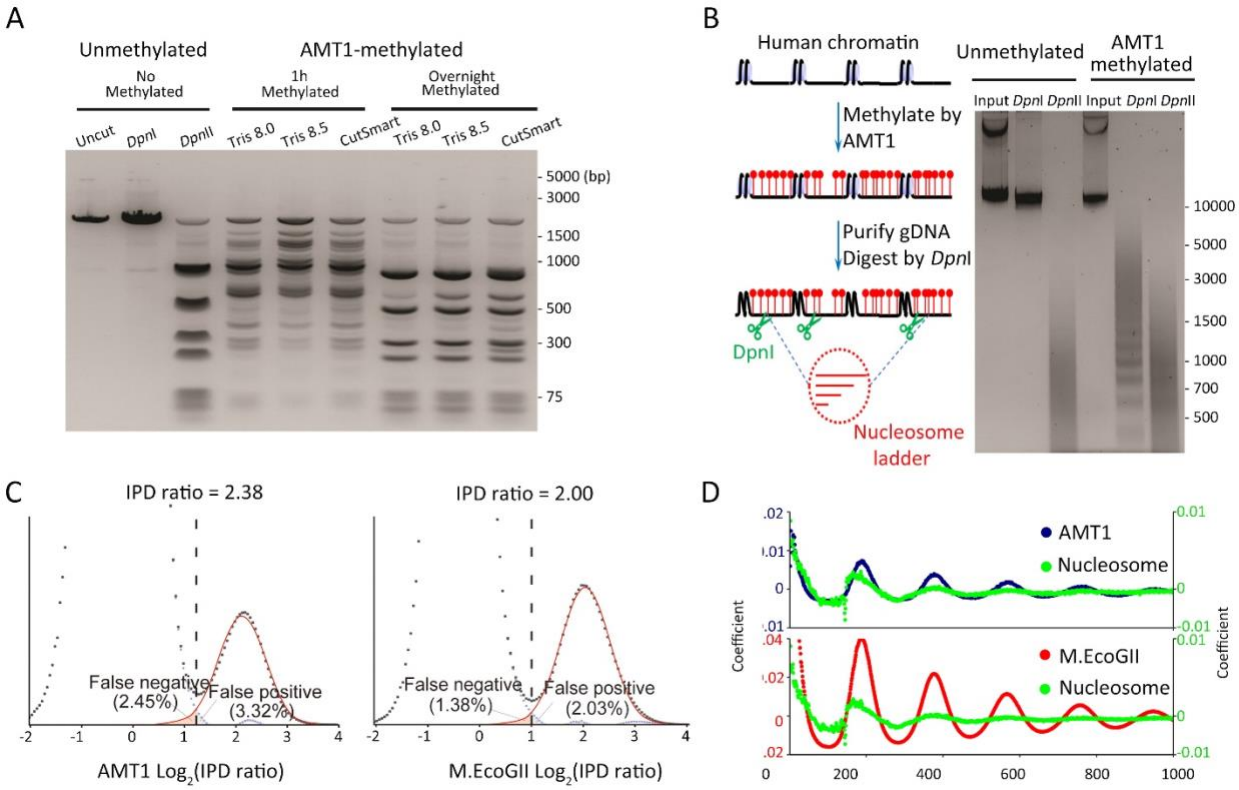


Figure S10. Additional characterization of in vitro methyltransferase activities of AMT1 complex.

10. Additional characterization of in vitro methyltransferase activities of AMT1 complex.

    A. MTase activity of AMT1 complex on linearized plasmid DNA. Three different buffers were tested. Methylation progress was monitored by DpnI cleavage, occurring only at methylated GATC sites. DpnII cleavage, occurring only at unmethylated GATC sites, was used as a control. Note that under all conditions, a substantial fraction of GATC sites were methylated by AMT1 complex in 1h, and almost all overnight, indicative of its robust MTase activity. Additional details are available in Supplemental Methods “In vitro methylation of plasmid DNA”.
    B. MTase activity of AMT1 complex on human chromatin. Methylation progress was monitored by DpnI digestion, with DpnII digestion as a control.
    C. Deconvolution of the 6mA peak and the unmodified A peak for IPDr distributions of all A sites, in human chromatin methylated by AMT1 complex (left) or M.EcoGII (right).
    D. Autocorrelation of 6mA distributions at the ensemble level in human chromatin methylated by AMT1 complex and M.EcoGII, respectively. For comparison, autocorrelation of nucleosome distribution in human chromatin was also plotted.

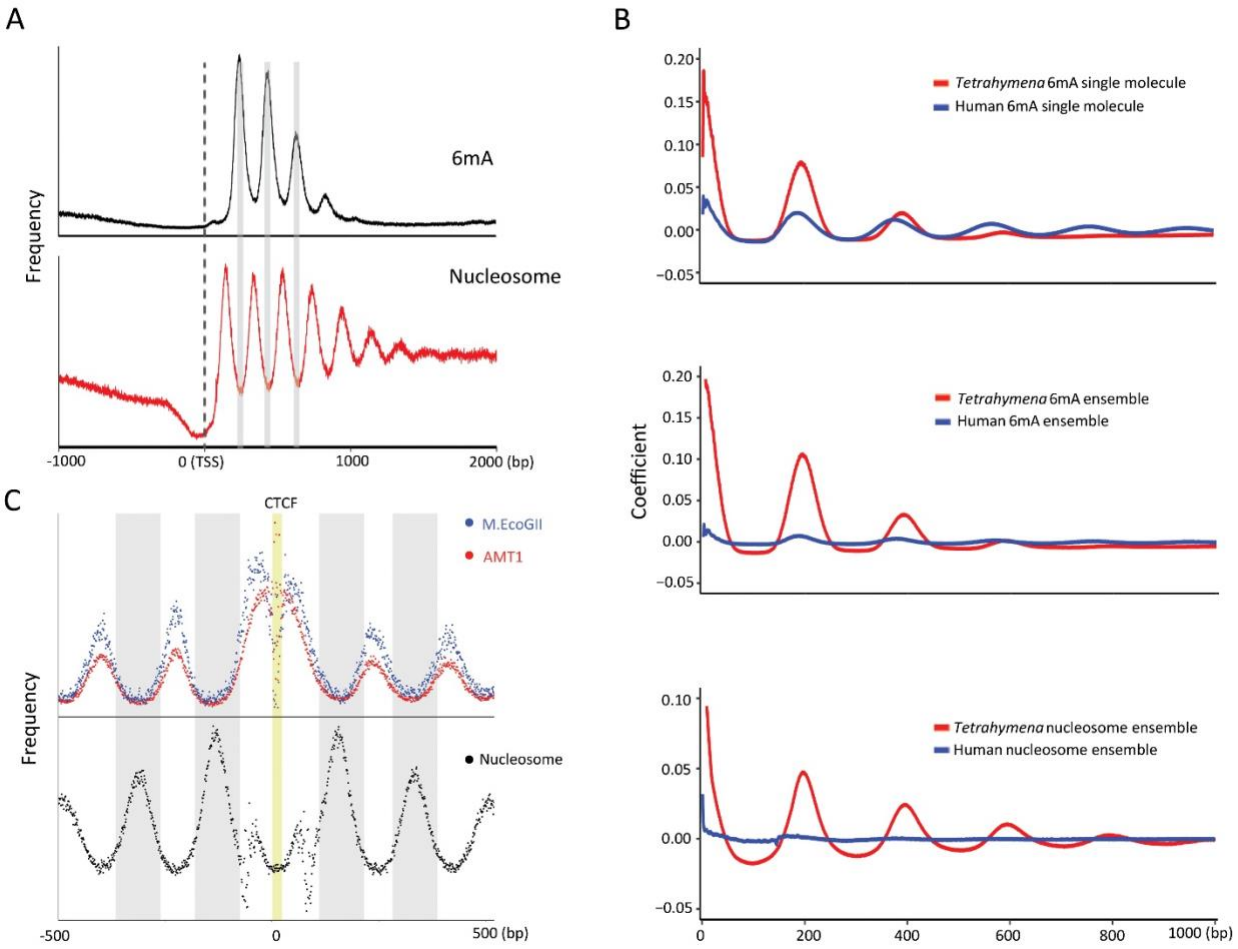


Figure S11. Additional characterization of the 6mA-nucleosome relationship.

11. Additional characterization of the 6mA-nucleosome relationship.
    A. 6mA and nucleosome distributions relative to TSS in *Tetrahymena* MAC. Pol II-transcribed genes are aligned to TSS; x-axis: the distance downstream (2000 bp) or upstream (-1000 bp) of TSS; y-axis: the aggregated count of 6mA sites (top) or nucleosome dyads (bottom). Note the peak-trough correspondence between the two distributions (gray bars), indicating 6mA enrichment in linker DNA. Additional details are available in Supplemental Methods “Gene level analyses” and “Nucleosome distributions in *Tetrahymena* and human”.
    B. Comparing 6mA and nucleosome distributions in *Tetrahymena* cells and in vitro methylated human chromatin. From top to bottom: autocorrelation of 6mA distribution at the single molecule level, 6mA distribution at the ensemble level, and nucleosome distribution at the ensemble level. Additional details are available in Supplemental Methods “Nucleosome distributions in *Tetrahymena* and human”.
    C. Anti-correlation between 6mA (top) and nucleosome distributions (bottom) around CTCF-binding sites (yellow highlight) of in vitro methylated human chromatin. Note the peak-trough correspondence between the two distributions (gray bars), indicating 6mA enrichment in linker DNA. Additional details are available in Supplemental Methods “Nucleosome distributions in *Tetrahymena* and human”.

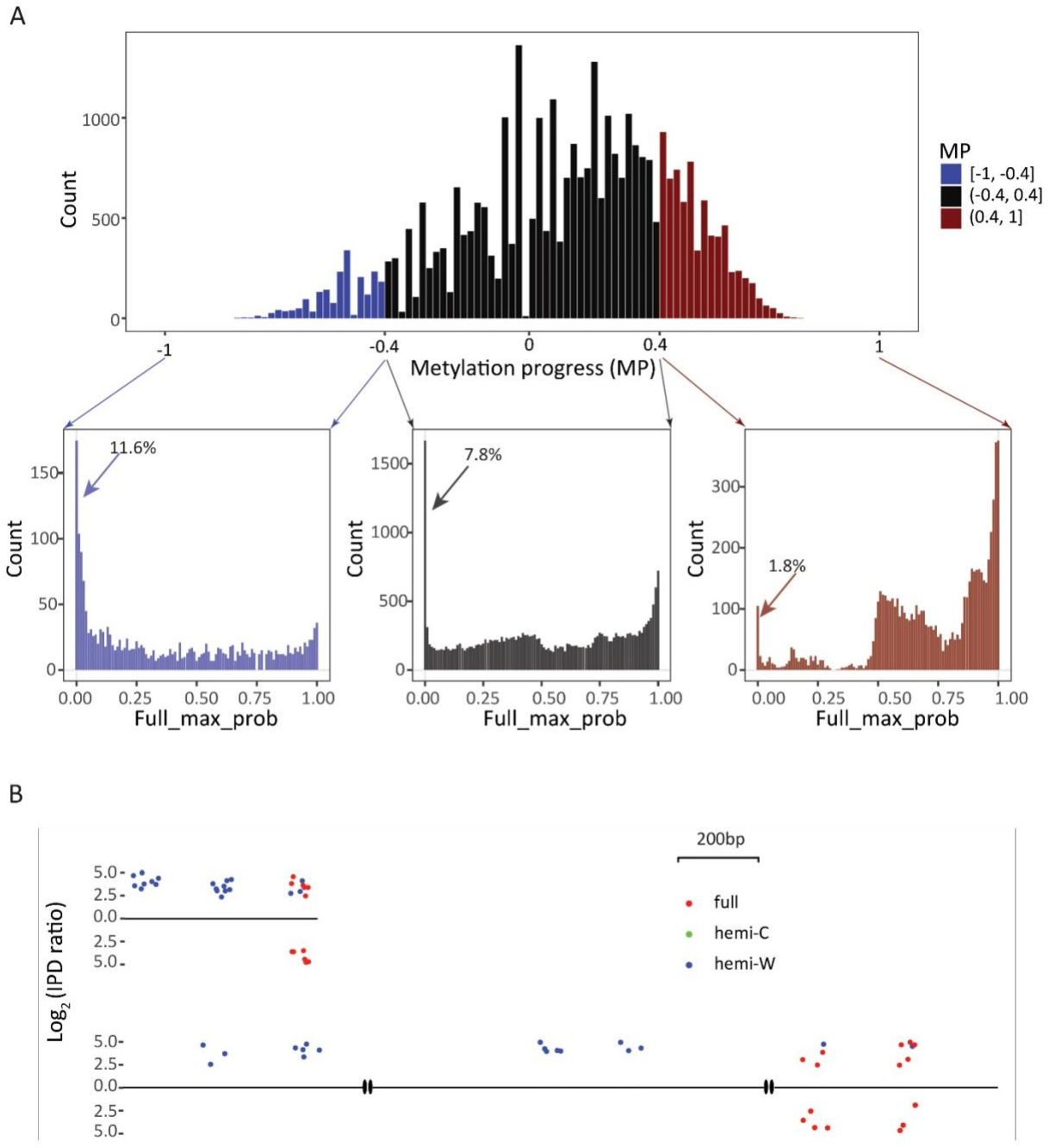


Figure S12. Full-6mApT congregation in DNA molecules undergoing hemi-to-full conversion in WT *Tetrahymena* cells.

12. Full-6mApT congregation in DNA molecules undergoing hemi-to-full conversion in WT *Tetrahymena* cells.
    A. Full-6mApT congregation in DNA molecules undergoing hemi-to-full conversion. Methylation progress (MP) is defined for a DNA molecule as the difference-sum ratio between the number of full-6mApT and hemi-6mApT: $(\frac{full-hemi}{full+hemi})$. DNA molecules are divided into three groups according to their MP values (top). Their max inter-full distances are calculated, and their distributions are plotted according to their probabilities in permutated simulations (bottom). The percentage of DNA molecules with probabilities no greater than 0.01 is labeled. Note its decrease with methylation progress.
    B. DNA molecules with congregated full-6mApT intermixed with hemi-6mApT.

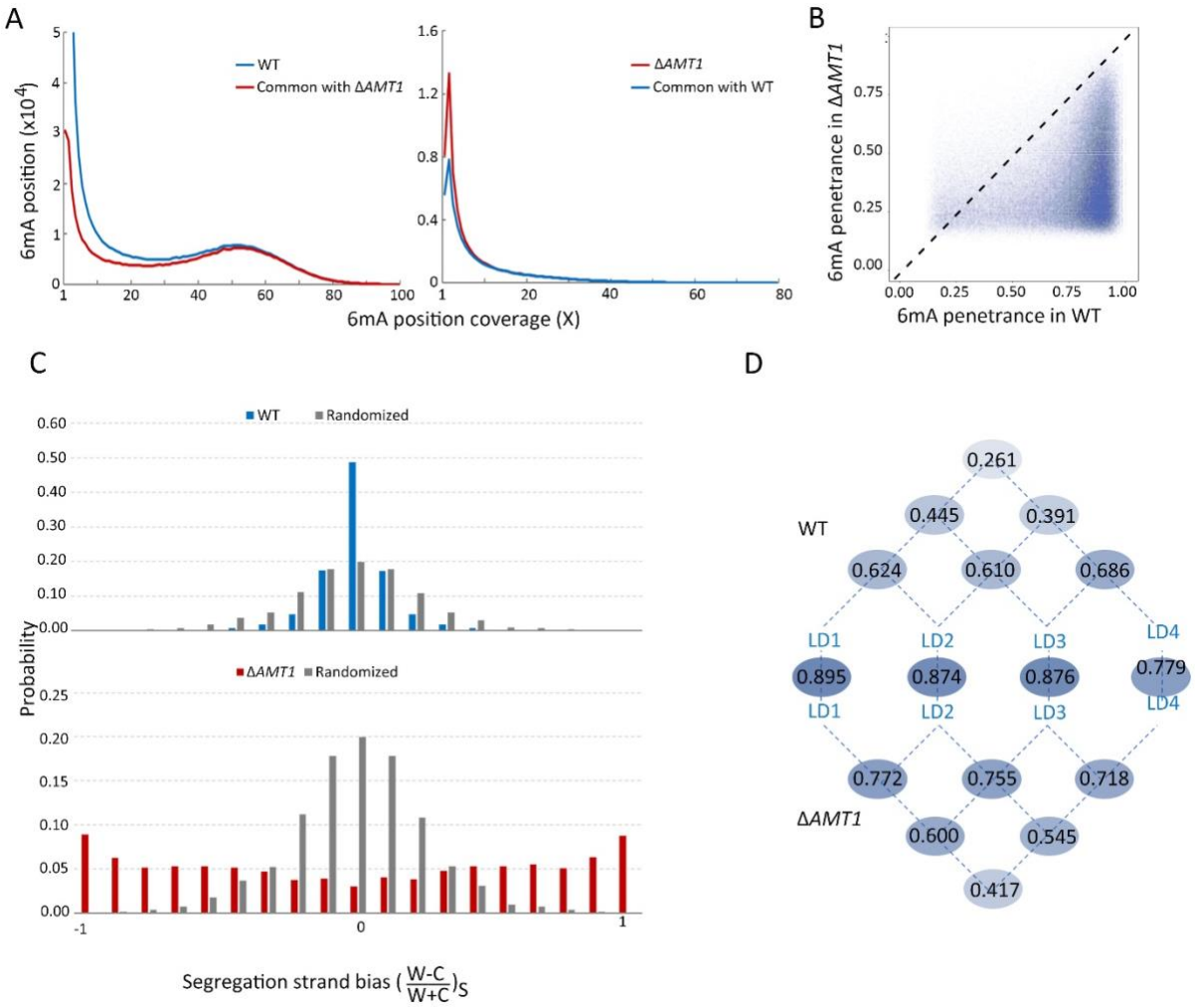


Figure S13. Comparison of 6mA in WT and Δ*AMT1* cells.

13. Comparison of 6mA in WT and Δ*AMT1* cells.
    A. Overlap between methylated ApT positions in the MAC genome of WT and Δ*AMT1* cells. Note their convergence with increasing 6mA coverage.
    B. Comparing 6mA penetrance of individual genomic positions present in both WT and Δ*AMT1* cells. Minimal 6mA coverage: 10. Note that most positions are below the diagonal line, indicative of reduced 6mA penetrance in Δ*AMT1* cells.
    C. 6mA segregation strand bias. In Δ*AMT1* cells, DNA molecules exhibited much stronger 6mA segregation strand bias than the randomized control (bottom), due to slow *de novo* methylation. In WT cells, DNA molecules exhibited even less 6mA segregation strand bias than the randomized control (top), due to quick restoration of full methylation. Note that both hemi and full-6mApT were counted here (hemi once for the corresponding strand; full twice, one for W and one for C). This is different from calculating segregation strand bias for hemi^{+} molecules, in which only hemi-6mApT was counted.
    D. 6mA levels of individual linker DNAs (LDs) in the gene body in WT and Δ*AMT1* cells are strongly correlated. For each LD, 6mA level is quantified by sum of penetrance values for all methylated ApT positions within. LD1 is defined to be in between the +1 and +2 nucleosome (the first and second nucleosome downstream of TSS, belong to the canonical nucleosome array in the gene body); LD2-4 are defined iteratively further downstream. Spearman's rank correlation coefficients were calculated for all possible pairs between LD1-4 within WT and Δ*AMT1* cells, respectively. They were also calculated between equivalent LDs from WT and Δ*AMT1* cells.

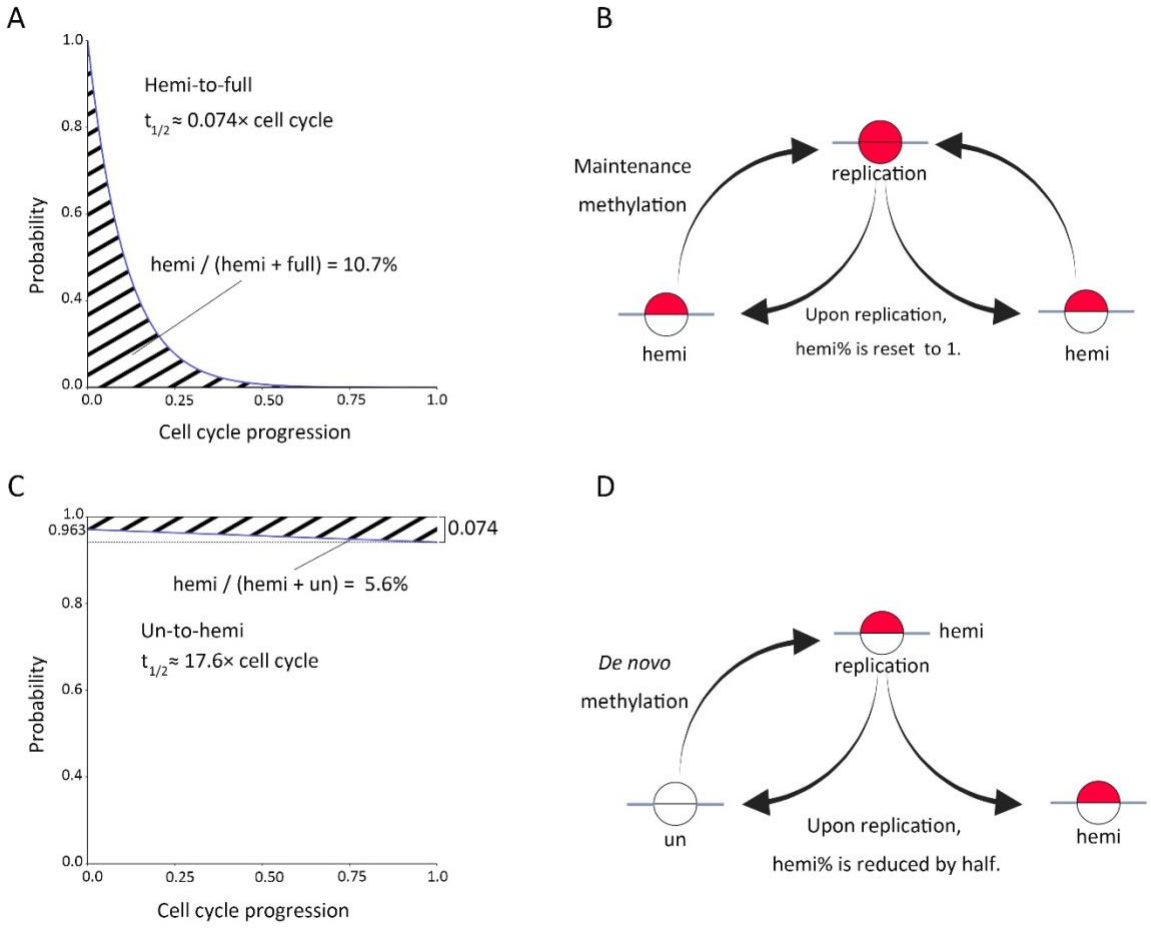


Figure S14. *De novo* and maintenance methylation models in *Tetrahymena*.

14. *De novo* and maintenance methylation models in *Tetrahymena*.
    A. An exponential decay model for maintenance methylation in WT cells. Upon DNA replication (t=0), all methylatable sites are reset as hemi-methylation; with cell cycle progression, hemi sites decrease exponentially as they are converted to full methylation, until the next round of DNA replication (t=1). Additional details are available in Supplemental Methods “Modeling 6mA *de novo* and maintenance methylation in *Tetrahymena*”.
    B. Upon DNA replication, all full methylation sites are split into hemi-methylation sites, as dictated by semi-conservative 6mA transmission.
    C. An exponential decay model for *de novo* methylation in Δ*AMT1* cells. Upon DNA replication (t=0), the hemi-methylation percentage is halved; with cell cycle progression, the count for unmethylated ApT duplexes decreases exponentially, while the hemi count increases, as *de novo* methylation converts the former to the latter, until the next round of DNA replication (t=1). Additional details are available in Supplemental Methods “Modeling 6mA *de novo* and maintenance methylation in *Tetrahymena*”.
    D. Upon DNA replication, the count for hemi-methylation sites stays the same, while the count for total methylatable sites is doubled. Therefore, the hemi-methylation percentage is halved.

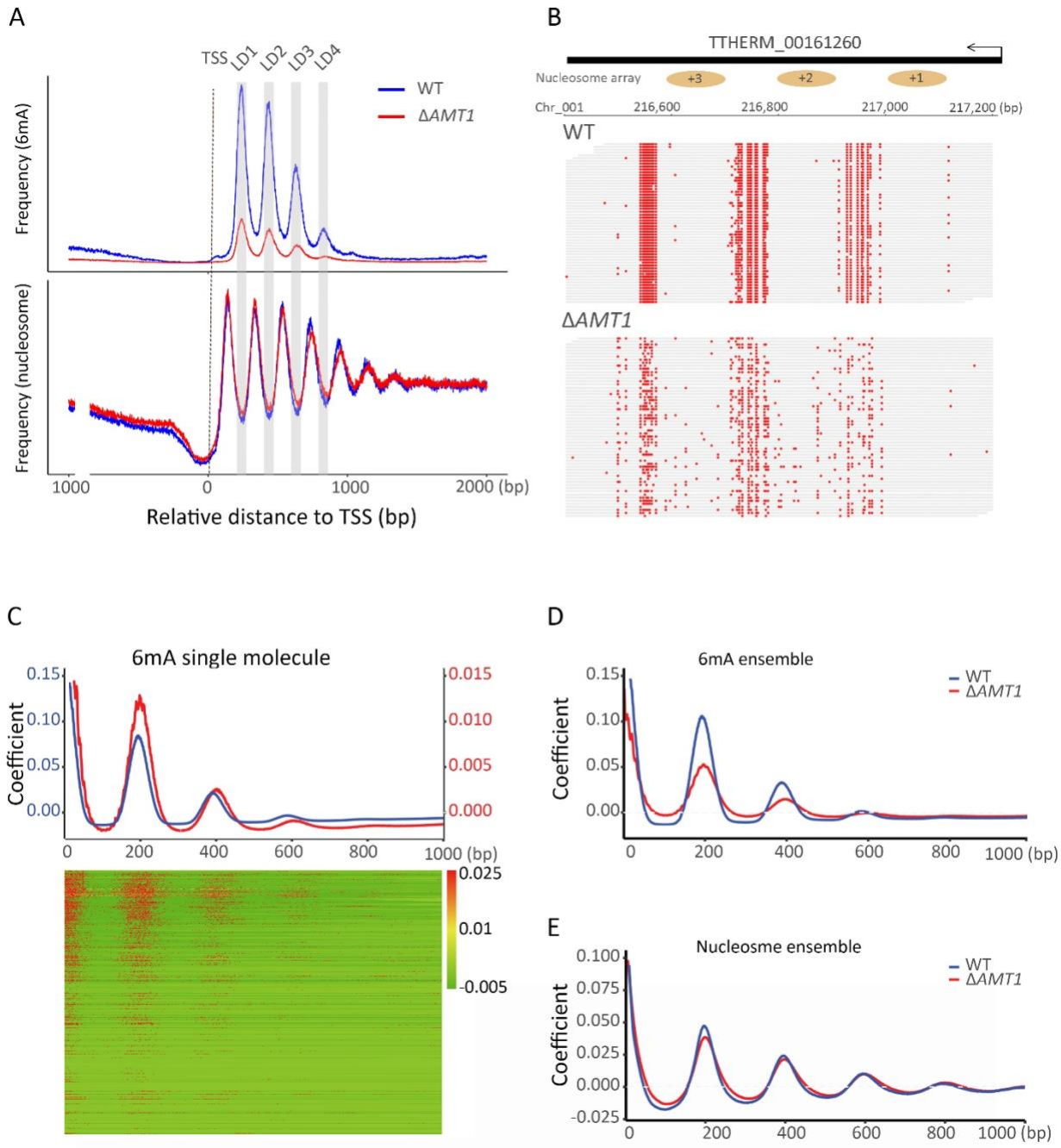


Figure S15. Dispersion of 6mA in Δ*AMT1* cells.

15. Dispersion of 6mA in Δ*AMT1* cells.
    A. 6mA and nucleosome distributions along the gene body in WT and Δ*AMT1* cells. Pol II-transcribed genes are aligned to TSS; x-axis: distance upstream (-1000bp) or downstream (2000bp) of TSS; y-axis: cumulative counts of 6mA sites (top) and nucleosome dyads (bottom). Additional details are available in Supplemental Methods “Gene level analyses” and “Nucleosome distributions in *Tetrahymena* and human”.
    B. 6mApT sites in DNA molecules from WT (top) and Δ*AMT1* cells (bottom). Note 6mA dispersion in the latter.
    C. Periodic 6mA distribution at the single molecule level in Δ*AMT1* cells (red). Autocorrelation between 6mA sites (distance ≤1kb) was calculated for individual DNA molecules, ranked by their median absolute deviations, and plotted as heat maps (bottom) and aggregated correlograms (top, along with the WT curve, blue).
    D. Periodic 6mA distributions at the ensemble level in WT and Δ*AMT1* cells.
    E. Periodic nucleosome distributions at the ensemble level in WT and Δ*AMT1* cells.

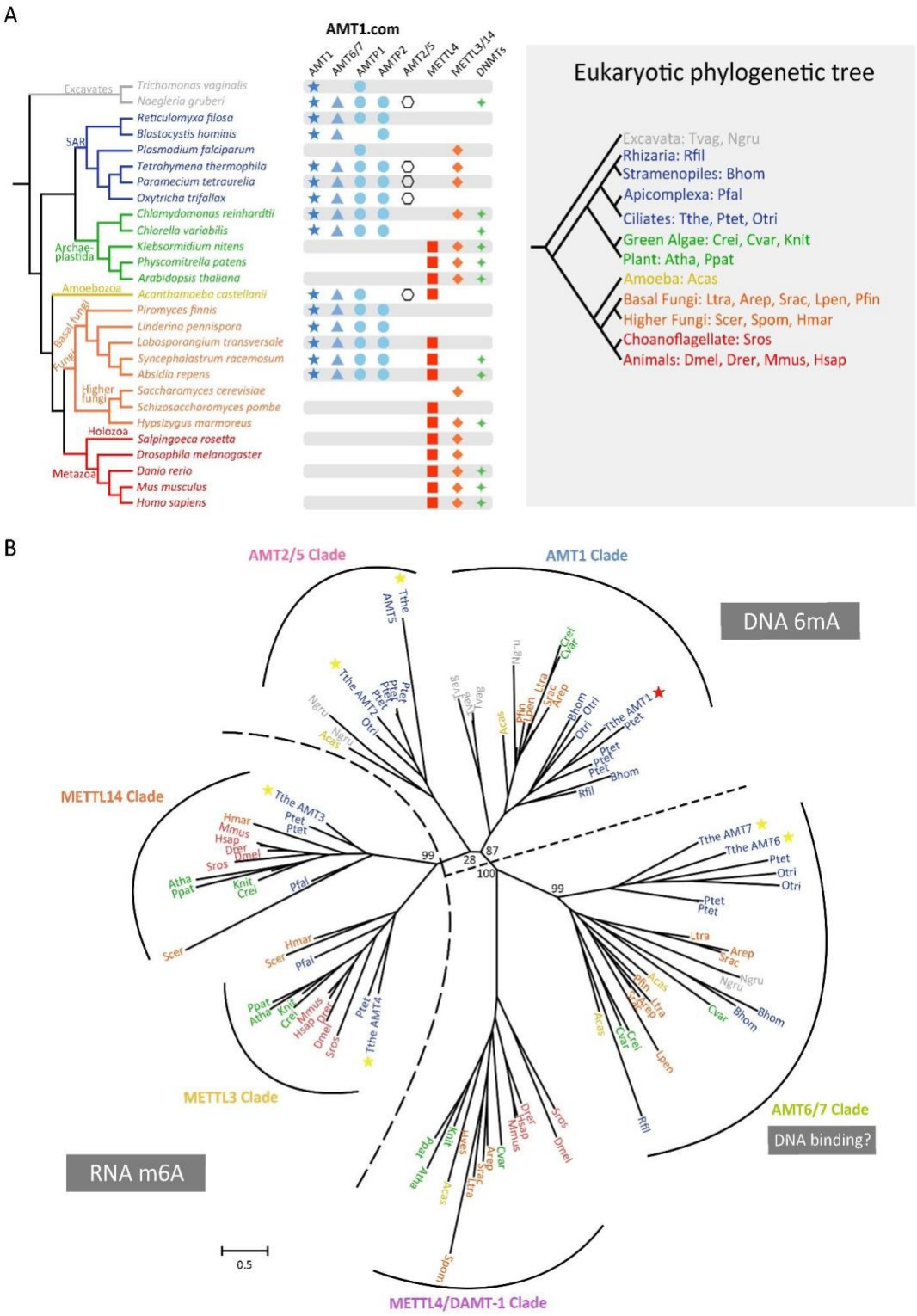


Figure S16. Phylogenetic distribution of MT-A70 MTases in eukaryotes.

16. Phylogenetic distribution of MT-A70 MTases in eukaryotes.
    A. Left: distribution of MT-A70 family MTases (for 6mA and m6A) and DNMT family MTases (for 5mC) in main eukaryote groups. Right: major branches of eukaryotic evolution.
    B. Phylogenetic tree of MT-A70 MTases. Note that members in AMT1 and AMT2/5 clades generally contain the DPPW motif critical for catalysis, while members in AMT6/7 and METTL4/DAMT-1 clade lack it. AMT7 (and possibly AMT6) is the heterodimeric partner for AMT1, likely providing a target recognition domain for binding the double-stranded DNA substrate. This situation is analogous to the heterodimeric RNA m6A MTase METTL3 and METTL14 (3,4). Plants, fungi, and animals have only members of METTL4/DAMT-1 clade, but not AMT1, AMT2/5, and AMT6/7 clades. METTL4/DAMT-1 clade members therein are more likely to be involved in DNA binding than 6mA deposition. Additional details are available in Supplemental Methods “Phylogenetic analysis”.

Reference	6mApT/ApT (%)	6mApA/ApA (%)	6mApC/ApC (%)	6mApG/ApG (%)
MAC	1.89	0.027	0.059	0.081
MIC	0.017	0.021	0.061	0.090
Mito	0.014	0.014	0.054	0.070

Table S1. 6mA levels in MAC, MIC, or mitochondrion.

Percentage of 6mApT/ApA/ApC/ApG sites called in MAC, MIC, or mitochondrion (based on the IPDr threshold of 2.38), relative to all ApT/ApA/ApC/ApG sites.

Site	6mA (%)	Unmodified A (%)
GATC	97.269	2.731
non-GATC	0.056	99.944

Table S2. 6mA levels in plasmid DNA.

Percentage of 6mA in GATC sites (*dam* sites) and non-GATC sites in a published dataset of plasmid DNA sequenced by SMRT CCS (Abdulhay et al. 2020). GATC sites are the positive control, as most of them should be fully methylated. Non-GATC sites are the negative control, with no expected methylation.

Single molecule	AMT1		M.EocGII	
	Number	Percentage (%)	Number	Percentage (%)
6mA	20,471,535	100	83,193,941	100
6mApC Sites	1,227,537	6.00	15,839,502	19.04
6mApG Sites	1,220,911	5.96	35,572,986	42.76
6mApA Sites	667,369	3.26	13,403,712	16.11
6mApT Sites	17,355,718	84.78	18,377,741	22.09
full-6mApT	7,402,199	85.30	2,355,107	25.63
hemi-C-6mApT	1,271,027	7.32	6,839,754	37.22
hemi-W-6mApT	1,280,113	7.38	6,827,773	37.15
Total 6mApT	17,355,718	100	18,377,741	100

Table S3. 6mA statistics for in vitro methylation of human chromatin.

Top: the count and percentage of 6mApT/ApA/ApC/ApG sites called in human chromatin of in vitro methylated by AMT1 complex and M.EcoGII. Bottom: the count and percentage of full and hemi-6mApT sites.

Supplemental Methods

In vitro reconstitution of AMT1 complex

The DNA sequences encoding the *Tetrahymena* AMT1, AMT7, AMTP1 (1-240 aa, truncating the C-terminal low complexity region that may interfere with overexpression and purification) and AMTP2 proteins were each codon optimized for *E. coli* expression and synthesized. AMT1 and AMT7 were inserted in tandem into an in-house bacterial expression vector, in which a His_6-MBP tag was fused to the AMT1 sequence via a TEV protease cleavage site. AMTP1 (1-240) and AMTP2 were cloned to a modified pRSF-Duet vector for co-expression, with AMTP2 preceded by an N-terminal His_6-SUMO tag and a ubiquitin-like protease 1 (ULP1) cleavage site. BL21(DE3) RIL cells harboring the expression plasmids were grown at 37°C and induced by addition of 0.2 mM isopropyl β-D-1-thiogalactopyranoside (IPTG) when the cell density reached A_{600} of 1.0. The cells continued to grow at 16°C overnight. Subsequently, the cells were harvested and lysed in buffer containing 50mM Tris-HCl (pH 8.0), 1M NaCl, 25mM Imidazole, 10% glycerol and 1mM PMSF. The fusion proteins were purified through nickel affinity chromatography, followed by removal of His_6-MBP and His_6-SUMO tags by TEV and ULP1 cleavage, ion-exchange chromatography on a Heparin column (GE Healthcare), and size-exclusion chromatography on a 16/600 Superdex 200pg column (GE Healthcare). The purified proteins were concentrated in 20mM Tris-HCl (pH 7.5), 100mM NaCl, 5% glycerol and 5mM DTT, and stored at -80°C.

In vitro methylation of plasmid DNA and short dsDNA

1µg of pUC19 plasmid generated from *dam-/dcm-* competent *E. coli* (NEB) was linearized by *EcoR*I and methylated with AMT1 complex in a 20-µL reaction (20mM Tris-HCl (pH 8.0 or 8.5) or 1×CutSmart Buffer (NEB), 160µM SAM, 2mM EDTA, 0.5mM EGTA). Reactions were performed for 1h at 37°C or overnight at 30°C, quenched with addition of 1% SDS, and incubated with proteinase K for 1h at 50°C. Methylated plasmid was digested with *Dpn*I. Digestion products of methylated samples and the unmethylated control were resolved on 1% agarose gel.

In vitro methylation of short dsDNA was performed in triplicate at room temperature for 30 min before being quenched by cold AdoMet. The reaction mixtures were spotted on Amersham Hybond-XL paper (GE healthcare), followed by sequential washes by ammonium bicarbonate, Milli Q water and Ethanol. Subsequently, the paper was soaked in ScintiVerse cocktail (Thermo fisher), and the radioactivity was detected with a Beckman LS6500 counter.

SMRT CCS data analysis

Smaller insert size favors accurate calling of 6mA as well as regular bases at the single molecule level, which is dependent on the number of CCS passes of individual DNA molecules. Larger insert size (generating more sequenced bp per SMRT Cell given the fixed upper limit for total reads) favors sensitive calling of low penetrance 6mA positions at the ensemble level, which is dependent on the overall sequencing coverage of the target genome. Our analyses showed that at ≥ 30 passes, 6mA calling accuracy reached a plateau. We therefore decided to use intermediate sized inserts (3-5 kb), enabling most CCS reads to have ≥ 30 passes, given the raw read length distribution on the Sequel II platform (N50 >100 kb). This allowed us to achieve a balance between 6mA calling accuracy and sequencing coverage.

We removed DNA molecules with global dispersion of IPD ratios (IPDr) for unmodified adenine sites: IPDr standard deviation (SD) ≥ 0.35. We also removed DNA molecules with local dispersion of IPDr, referred to as N* clusters. N = G, C, T; N*: IPDr ≥ 2.8; N* cluster: inter-N* distances ≤ 25bp, N* count ≥ 4, on the same strand. Finally, an IPDr threshold was determined by peak deconvolution (IPDr=2.38 for Replicate 1 of WT cells) for calling bulk 6mA in the remaining DNA molecules**.**

6mA sites are associated with significantly increased IPDr; at sufficient abundance, 6mA sites form a peak distinct from unmodified adenines in the IPDr distribution. While PCR-amplified DNA was often included as the negative control for 6mA calling with SMRT CLR (Continuously Long Read), its usage is no longer necessary for SMRT CCS (aka PacBio Hifi sequencing), which allows highly accurate and sensitive calling of 6mA at the

single molecule level and is quickly emerging as the gold-standard for 6mA detection. Indeed, the background level for 6mA detection by SMRT CCS can be very low (<100 ppm) (Kong et al. 2022). Here, the background noise level is defined as the percentage of false 6mA calls, relative to ALL adenine sites (with or without the modification). The false positive rate is defined as the percentage of false 6mA calls, relative to ALL 6mA calls (true or false); low false positive rate means high accuracy. The false negative rate is defined as the percentage of missed 6mA, relative to ALL 6mA (called or missed); low false negative rate means high sensitivity.

We deconvoluted the 6mA peak and the unmodified A peak of the ApT dinucleotide by fitting the former with a Gaussian distribution. We set the 6mA calling threshold at the intersection of the two peaks (IPDr=2.38, for WT *Tetrahymena*), keeping both false positive and false negative rates at low levels (FP: 1.93%, FN: 1.12%). For the same dataset, increasing the threshold (e.g., IPDr=3.03) will decrease FP but increase FN (FP: 0.90%, FN: 10.15%), while decreasing the threshold (e.g., IPDr=2.00) will increase FP but decrease FN (FP: 4.87%, FN: 0.19%). Furthermore, as the threshold is determined by the relative abundance of 6mA and unmodified A, it can vary across different datasets (higher threshold for lower 6mA abundance). This is also critical for distinguishing hemi-6mApT and full-6mApT in WT *Tetrahymena*. For ApT duplexes with a 6mA already called on one of the two strands, the 6mA peak was predominant over the unmodified A peak on the other strand, opposite to the situation for bulk ApT duplexes. This caused the 6mA calling threshold on the other strand to shift substantially to the left (IPDr=1.57), which was essentially a change in conditional probability. This left shift reduced the false negative rate, at the cost of increasing the false positive rate. At a higher false negative rate, many full-6mApT duplexes would be mis-classified as hemi-6mApT duplexes; these falsely identified hemi-6mApT duplexes would exhibit no strand bias and overwhelm the signal from true hemi-6mApT duplexes, which only represents a small minority (~10%) of total 6mApT duplexes in WT *Tetrahymena* cells. At a higher false positive rate, fewer hemi-6mApT duplexes would be called (mis-classified as full-6mApT). However, the strand bias of true hemi-6mApT duplexes would not be affected (even though the number of hemi$^+$ molecules would be reduced). Therefore, the left shift of the IPDr threshold is critical for making stringent calls of hemi-6mApT duplexes, which in turn is critical for

revealing their strand bias.

We examined whether varying the number of CCS passes (by increasing the threshold of minimum number of passes from 30× to 40× and 50×) affected hemi- and full-6mApT calling (Supplemental Fig. S6A-F). The false positive and false negative rates were mostly reduced with increasing CCS passes (Supplemental Fig. S6E), as expected for higher quality reads with reduced coefficients of variance for IPD (Supplemental Fig. S1B and S6A-E). The IPDr thresholds for 6mA calling were stable (1st IPDr unchanged, 2nd IPDr slightly shifted to the left) (Supplemental Fig. S6A-D). The hemi-6mApT and full-6mApT percentages also remained stable (Supplemental Fig. S6F). The segregation strand bias for hemi-6mApT sites in hemi^{+} molecules strengthened slightly with increasing CCS passes (Supplemental Fig. S6G). All these results support that our analysis of hemi^{+} molecules is robust. The non-random nature of hemi-6mApT sites (overlap between hemi- and full methylation containing genomic positions in WT cells, penetrance strand bias for hemi sites in 6mApT genomic positions in Δ*AMT1* cells, as well as segregation strand bias for hemi sites in hemi^{+} molecules of WT cells) strongly argues against that they are generated by background noise or stochastic fluctuations of IPDr.

Mapping CCS reads back to genome references

We aligned CCS reads to the latest *Tetrahymena* genome references for the MAC (Sheng et al. 2020), MIC (Supplemental File_Tet MIC, updating the published MIC reference (Hamilton et al. 2016), with improved assembly of repetitive sequences), and mitochondrion (Brunk et al. 2003). *Tetrahymena* MAC genome reference sequence is a finished telomere-to-telomere (T2T) assembly, which has been published recently (Sheng et al. 2020; Wang et al. 2021). Mapping was performed by BLASTN (Ye et al. 2006); the parameters "-max-target_seqs 2, -max_hsps 2" were used to allow identification and mapping of bipartite reads (with two segment pairs). We focused on fully mapped reads (mapped length ≥ 98% for the only segment or segment pairs combined). Segments with mapped identity ≥ 95% were retained for further analysis. DNA molecules mapped specifically to the MIC rather than the MAC—usually containing MIC-limited internal eliminated sequences (IES) comprising transposable elements and repetitive

sequences—were distinguished by much higher BLASTN alignment scores (Δ ≥ 50) matching the MIC reference genome than the MAC. Some DNA molecules fully mapped to the MAC genome may come from the MIC, as they correspond to genomic regions not interrupted by MIC-limited sequences. Nonetheless, due to the high ploidy (~90×) of the MAC relative to the diploid MIC (Zhou et al. 2022), they only represent a small fraction (< 5%), thus should not significantly affect the analysis results.

Chimeric reads in our datasets probably do not represent issues with the reference genome, as no recurrent breakpoints (covered by more than one read) were found, supporting that these chimeric reads were generated by random ligation of DNA fragments. Mapping reads back to the reference genome revealed relatively smooth and even coverage (Supplemental Fig. S2D), also supporting the correct assembly. Indeed, percentage of chimeric reads varied from sample to sample, strongly suggesting that they were artifacts generated during library preparation. Discrepancy from the MAC reference mostly originated from heterogeneous junctions of MAC-destined sequences, generated by imprecise removal of thousands of MIC-limited sequences. Discrepancy from the MIC reference was mostly found in repetitive regions that are difficult to assemble.

We used reads mapped to the *Tetrahymena* micronuclear and mitochondrial reference sequences as negative controls. As they presumably contain no 6mA, all 6mA calls therein are regarded as false positive. This allowed us to estimate the background noise level (6mApT/ApT: 0.017% and 0.014% for the micronucleus and mitochondrion, respectively; Supplemental Table S1). In contrast, reads mapped to the macronuclear reference genome showed much higher 6mApT/ApT (1.89%). Assuming the same background noise level in reads mapped to the *Tetrahymena* macronuclear reference genome, we deduced that its false positive rate for 6mApT calling was ~1% ($\frac{\text{6mA\% in MIC or Mito}}{\text{6mA\% in MAC}}$). This value was close to what we derived by deconvolution of 6mA and unmodified A peaks (1.93%, Fig 1E).

We also sequenced *Tetrahymena* DNA after whole genome amplification (WGA) and used the result as a negative control. WGA effectively removes all base modifications while preserving the sequence information. Using the same bioinformatic pipeline, we found low background noise level in WGA (6mApT/ApT: 0.030% for reads mapped to the

macronuclear reference; Supplemental Fig. S3C). Assuming the same background noise level in native *Tetrahymena* DNA reads mapped to the macronuclear reference, we deduced that its false positive rate for 6mApT calling was 1.49% ($\frac{\text{6mA\% in WGA}}{\text{6mA\% in Native}}$) (Supplemental Fig. S3B), close to the value derived by deconvolution of 6mA and unmodified A peaks (1.93%, Fig 1E).

The mapping results of representative genomic regions were visualized using Jbrowse (Skinner et al. 2009) or ggplot2 package in R (Wickham 2016).

The SMRT CCS raw data and 6mA calling results can be downloaded from:

https://dataview.ncbi.nlm.nih.gov/object/PRJNA932808?reviewer=tak7a6fv74pn85n6nmi3ekpdqb

The code for base modification calling is available at:

https://gitfront.io/r/user-1129035/86c191cfcd32dc94d0047fa0617509701f714764/Pacbio_m6A_calling/

CCS versus CLR

While SMRT CCS can call 6mA on individual reads/DNA molecules, SMRT CLR only calls 6mA at the ensemble level, by combining different reads covering the same genomic position, and is therefore affected by both sequencing coverage and 6mA homogeneity. 6mApT coverage of a genomic position, as the product of sequencing coverage and 6mA penetrance (6mApT coverage = sequencing coverage × 6mA penetrance), is a good indicator of CLR performance. Indeed, at high 6mApT coverage (i.e., at genomic positions with high sequencing coverage AND high 6mA penetrance), CLR calls converge with CCS calls (Supplemental Fig. S3A, B). In contrast, CLR performs poorly at low 6mApT coverage (i.e., at genomic positions with low sequencing coverage OR low 6mA penetrance) (Supplemental Fig. S3A). Also, as 6mA homogeneity decreased (from high 6mA fraction (>80%), intermediate (20-80%), to low (<20%)), percentage of 6mA called by CLR in non-ApT dinucleotides (regarded as false positive, based on our CCS analysis) increased (0.6%, 9.8%, and 82.7%, respectively) (Wang et al. 2017). False positive 6mA

calls of CLR are not limited to non-ApT dinucleotides. We also compared high confidence ApT genomic positions called by CLR (identification Qv≥ 30) and CCS (6mApT coverage≥ 10) (Supplemental Fig S3B). Genomic positions called by both showed much better scores in identification Qv and 6mA fraction (both CLR metrics) than those called only by CLR (Supplemental Fig S3C, D). In terms of 6mA penetrance (a CCS metric), genomic positions called by both CLR and CCS showed a distribution strongly skewed to the right (high penetrance); in contrast, the distribution of CCS-only calls (regarded as false negative) was skewed to the left (lower penetrance) (Supplemental Fig S3E). Furthermore, unlike 6mApT genomic positions called by CCS, those called only by CLR were no longer enriched in linker DNA (Supplemental Fig S3F). This is very similar to the distribution of 6mA calls at non-ApT dinucleotides (Supplemental Fig S3G) and consistent with the random nature of these false positive calls.

Gene level analyses

Models for well-annotated Pol II-transcribed genes in *Tetrahymena* (15,810 in total) were updated with the latest RNA-seq data (SRR9176464). We calculated the coefficients of variance (CV) for 6mA levels across different DNA molecules covering a gene. We only counted DNA molecules fully covering the gene body. For comparison, we only included genes with high coverage (≥20× overall coverage) and high 6mA levels (ΣP≥2) in both WT and Δ*AMT1* cells. CV ratios between WT and Δ*AMT1* cells were calculated as an indicator of their relative 6mA variability at the gene level.

For composite analysis, *Tetrahymena* genes were aligned to TSS and TES (the gene body is normalized to unit length) and extended in both directions by 0. 5kb. Alternatively, *Tetrahymena* genes are aligned to TSS with upstream (1 kb) and downstream (2 kb) extensions. For 6mA and nucleosome distributions, counts of 6mA sites and MNase-seq fragment centers in specified genomic regions were aggregated.

Nucleosome distributions in *Tetrahymena* and human

Micrococcal nuclease digestion is used to determine nucleosome positioning in *Tetrahymena* and human cells. We have previously published a paper addressing

nucleosome positioning in *Tetrahymena*, in which MNase-seq is performed by paired-end sequencing and at high coverage (Xiong et al. 2016). The reads were mapped back to the latest MAC genome reference (Sheng et al. 2020). For human cells, we aligned SMRT CCS and Illumina reads to the human reference genome HG19 (https://hgdownload.cse.ucsc.edu/goldenpath/hg19/bigZips/hg19.fa.gz). For mapping nucleosome distribution around CTCF-binding sites, CTCF ChIP-seq data for a human leukemia cell line OCI-AML3 were used to call CTCF peaks with MACS2 (Zhang et al. 2008). CTCF motif profile (accession number: MA0139.1) was downloaded from JASPER (Fornes et al. 2020) (http://jaspar.genereg.net/) and used to scan all the CTCF peak regions by FIMO (Grant et al. 2011). High quality CTCF binding sites (motif score ≥ 60) were retained for further analyses. For 6mA distribution, 6mA sites were aggregated for each base in a window of -500 to 500 bp around the high quality CTCF binding sites (aligned to the left boundary of the CTCF binding consensus sequence). For nucleosome distribution, Micro-C data were downloaded from the 4DN data portal (https://data.4dnucleome.org) and processed as previously described (Krietenstein et al. 2020). The raw data adapters were trimmed with Trim Galore and mapped with Bowtie2 (Ramírez et al. 2016) against the human genome HG19 using the following parameters (bowtie2 --local -q --phred33 --threads 12 -x hg19 bowtie2 indexed genome -U trimmed.fq.gz -S output.sam). Multiple mapped reads and reads with low mapping quality (MAPQ < 30) were removed using SAMtools (Li et al. 2009). PCR duplicates were removed with Picard tools (http://broadinstitute.github.io/picard/). Read starts were then shifted by 73 bp to reveal positions of nucleosome dyads, which were aggregated for each base in a window of -500 to 500 bp around the high quality CTCF binding sites.

Modeling 6mA *de novo* and maintenance methylation in *Tetrahymena*

We developed a steady-state model to evaluate kinetics for 6mA *de novo* and maintenance methylation in *Tetrahymena*. Using 6mA data from asynchronously growing *Tetrahymena* cells, we made the following assumptions to simplify the problem: ***1)*** the system is ergodic (i.e., even sampling of the cell cycle); ***2)*** there is no active demethylation; and ***3)*** the free pool of the corresponding MTase(s) is constant (the reaction rate is therefore solely dependent on substrate concentration, i.e., exponential decay).

For 6mA maintenance methylation, we used the WT *Tetrahymena* data to estimate the apparent half-life of hemi-6mApT (Supplemental Fig. S12A, B). Upon DNA replication (t=0), all methylatable sites are reset as hemi-methylation (Supplemental Fig. S12B); with cell cycle progression, hemi sites decrease exponentially as they are converted to full methylation, until the next round of DNA replication (t=1) (Supplemental Fig. S12A). Under these circumstances, the shaded area (A) represents the hemi-6mApT percentage time averaged over the entire cell cycle, corresponding to the observed bulk hemi-6mApT percentage ($\frac{hemi}{hemi + full}$) (Supplemental Fig. S12A and Equation 1).

$$\int_{t=0}^{t=1} e^{-xt} = A \quad (\mathbf{1})$$

$$1 - e^{-x} = Ax \quad (1')$$

$$A: \frac{hemi}{hemi + full}\ (shaded\ area, Supplemental\ Fig. S12A)$$

$$A = 0.107 \text{ (observed)}$$

$$x: kinetic\ parameter\ for\ expoential\ decay;\ \frac{t_{1/2}}{t_{doubling}} = \frac{In\ 2}{x}$$

The equation was solved, yielding the apparent half-life of hemi-6mApT ($t_{1/2}$), relative to the WT *Tetrahymena* cell cycle duration ($t_{doubling}$) (Supplemental Fig. S12A).

For 6mA *de novo* methylation, we used the Δ*AMT1* data to estimate the apparent half-life of unmethylated ApT duplexes (Supplemental Fig. S12C, D). Upon DNA replication (t=0), the hemi-methylation percentage is halved (Supplemental Fig. S12D and Equation 2); with cell cycle progression, the percentage for unmethylated sites decreases exponentially, while the percentage for hemi-methylation sites increases, as *de novo* methylation converts the former to the latter, until the next round of DNA replication (t=1) (Supplemental Fig. S12C). Under these circumstances, the shaded area (A) represents the hemi-6mApT percentage time averaged over the entire cell cycle, corresponding to the observed bulk hemi-6mApT percentage ($\frac{hemi}{hemi + Un}$) (Supplemental Fig. S12C and Equation 3).

$$(1 - y)\, e^{-x} = 1 - 2y \quad (\mathbf{2})$$

$$1 - \int_{t=0}^{t=1} (1 - y)\, e^{-xt} = A \quad (\mathbf{3})$$

$$(1 - e^{-xt})\,(1 - y) = (1 - A)\, x \quad (3')$$

$$A: \frac{hemi}{hemi + Un}\ (shaded\ area, Supplemental\ Fig. S12C)$$

$$A = 0.056 \text{ (observed)}$$

$$y = \frac{hemi}{hemi + un}\ at\ t = 0$$

The equation was solved, yielding the apparent half-life of unmethylated ApT duplexes ($t_{1/2}$), relative to the Δ*AMT1* cell cycle duration ($t_{doubling}$) (Supplemental Fig. S12C).

Phylogenetic analysis

The well-studied 27 taxa covered major branches of eukaryotic evolution were selected to search for homologous proteins (Altschul et al. 1997; Stover et al. 2006; Merchant et al. 2007; Aurrecoechea et al. 2009a; Aurrecoechea et al. 2009b; Fritz-Laylin et al. 2010; Aurrecoechea et al. 2011; Denoeud et al. 2011; Cherry et al. 2012; Swart et al. 2013; Glöckner et al. 2014; Berardini et al. 2015; Mondo et al. 2017; Karimi et al. 2018; Bult et al. 2019; Lock et al. 2019; Ruzicka et al. 2019; Arnaiz et al. 2020; Harris et al. 2020; Larkin et al. 2021). List of species: *Naegleria gruberi*, *Trichomonas vaginalis*, *Reticulomyxa filosa*, *Blastocystis hominis*, *Plasmodium falciparum*, *Paramecium tetraurelia*, *Tetrahymena thermophila*, *Oxytricha trifallax*, *Chlamydomonas reinhardtii*, *Chlorella variabilis*, *Klebsormidium nitens*, *Arabidopsis thaliana*, *Physcomitrella patens*, *Acanthamoeba castellanii* , *Lobosporangium transversale*, *Absidia repens*, *Syncephalastrum racemosum*, *Linderina pennispora*, *Piromyces finnis*, *Saccharomyces cerevisiae*, *Schizosaccharomyces pombe*, *Hypsizygus marmoreus*, *Salpingoeca rosetta*, *Drosophila melanogaster*, *Danio rerio*, *Mus musculus*, and *Homo sapiens*. The AMTs 1–7, AMTP1 and AMTP2 amino acid sequences were queried against the database using PSI-BLAST (Altschul et al. 1997) (maximum E-value = 1e–4), respectively. Retrieved hits were collapsed using CD-HIT (Li and Godzik 2006) (-c 0.97) to remove redundant sequences. Sequences were aligned using MUSCLE (Edgar 2004) and phylogenetic trees were constructed using FastTree (Price et al. 2010) under default parameters.

References

Abdulhay NJ, McNally CP, Hsieh LJ, Kasinathan S, Keith A, Estes LS, Karimzadeh M, Underwood JG, Goodarzi H, Narlikar GJ et al. 2020. Massively multiplex single-molecule oligonucleosome footprinting. *Elife* **9**: e59404.

Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* **25**(17): 3389-3402.

Arnaiz O, Meyer E, Sperling L. 2020. ParameciumDB 2019: integrating genomic data across the genus for functional and evolutionary biology. *Nucleic Acids Res* **48**(D1): D599-D605.

Aurrecoechea C, Barreto A, Brestelli J, Brunk BP, Caler EV, Fischer S, Gajria B, Gao X, Gingle A, Grant G et al. 2011. AmoebaDB and MicrosporidiaDB: functional genomic resources for Amoebozoa and Microsporidia species. *Nucleic Acids Res* **39**(suppl_1): D612-D619.

Aurrecoechea C, Brestelli J, Brunk BP, Carlton JM, Dommer J, Fischer S, Gajria B, Gao X, Gingle A, Grant G et al. 2009a. GiardiaDB and TrichDB: integrated genomic resources for the eukaryotic protist pathogens *Giardia lamblia* and *Trichomonas vaginalis*. *Nucleic Acids Res* **37**(suppl_1): D526-D530.

Aurrecoechea C, Brestelli J, Brunk BP, Dommer J, Fischer S, Gajria B, Gao X, Gingle A, Grant G, Harb OS et al. 2009b. PlasmoDB: a functional genomic database for malaria parasites. *Nucleic Acids Res* **37**(suppl_1): D539-D543.

Berardini TZ, Reiser L, Li D, Mezheritsky Y, Muller R, Strait E, Huala E. 2015. The arabidopsis information resource: Making and mining the "gold standard" annotated reference plant genome. G*enesis* **53**(8): 474-485.

Brunk CF, Lee LC, Tran AB, Li J. 2003. Complete sequence of the mitochondrial genome of *Tetrahymena thermophila* and comparative methods for identifying highly divergent genes. *Nucleic Acids Res* **31**(6): 1673-1682.

Bult CJ, Blake JA, Smith CL, Kadin JA, Richardson JE, the Mouse Genome Database G. 2019. Mouse Genome Database (MGD) 2019. *Nucleic Acids Res* **47**(D1): D801-D806.

Cherry JM, Hong EL, Amundsen C, Balakrishnan R, Binkley G, Chan ET, Christie KR, Costanzo MC, Dwight SS, Engel SR et al. 2012. *Saccharomyces* Genome Database: the genomics resource of budding yeast. *Nucleic Acids Res* **40**(D1): D700-D705.

Denoeud F, Roussel M, Noel B, Wawrzyniak I, Da Silva C, Diogon M, Viscogliosi E, Brochier-Armanet C, Couloux A, Poulain J et al. 2011. Genome sequence of the stramenopile *Blastocystis*, a human anaerobic parasite. *Genome Biol* **12**(3): R29.

Edgar RC. 2004. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinform* **5**(1): 113.

Fornes O, Castro-Mondragon JA, Khan A, Van der Lee R, Zhang X, Richmond PA, Modi BP, Correard S, Gheorghe M, Baranašić D. 2020. JASPAR 2020: update of the open-access database of transcription factor binding profiles. *Nucleic Acids Res* **48**(D1): D87-D92.

Fritz-Laylin LK, Prochnik SE, Ginger ML, Dacks JB, Carpenter ML, Field MC, Kuo A, Paredez A, Chapman J, Pham J et al. 2010. The genome of *Naegleria gruberi* illuminates early eukaryotic versatility. *Cell* **140**(5): 631-642.

Glöckner G, Hülsmann N, Schleicher M, Noegel Angelika A, Eichinger L, Gallinger C, Pawlowski J, Sierra R, Euteneuer U, Pillet L et al. 2014. The genome of the foraminiferan *Reticulomyxa filosa*. *Curr Biol* **24**(1): 11-18.

Grant CE, Bailey TL, Noble WS. 2011. FIMO: scanning for occurrences of a given motif. *Bioinformatics* **27**(7): 1017-1018.

Hamilton EP, Kapusta A, Huvos PE, Bidwell SL, Zafar N, Tang H, Hadjithomas M, Krishnakumar V, Badger JH, Caler EV et al. 2016. Structure of the germline genome of *Tetrahymena thermophila* and relationship to the massively rearranged somatic genome. *Elife* **5**.

Harris TW, Arnaboldi V, Cain S, Chan J, Chen WJ, Cho J, Davis P, Gao S, Grove CA, Kishore R et al. 2020. WormBase: a modern model organism information resource. *Nucleic Acids Res* **48**(D1): D762-D767.

Karimi K, Fortriede JD, Lotay VS, Burns KA, Wang DZ, Fisher ME, Pells TJ, James-Zorn C, Wang Y, Ponferrada V G et al. 2018. Xenbase: a genomic, epigenomic and transcriptomic model organism database. *Nucleic Acids Res* **46**(D1): D861-D868.

Kong Y, Cao L, Deikus G, Fan Y, Mead EA, Lai W, Zhang Y, Yong R, Sebra R, Wang H et al. 2022. Critical assessment of DNA adenine methylation in eukaryotes using quantitative deconvolution. *Science* **375**(6580): 515-522.

Krietenstein N, Abraham S, Venev SV, Abdennur N, Gibcus J, Hsieh TS, Parsi KM, Yang L, Maehr R, Mirny LA et al. 2020. Ultrastructural details of mammalian chromosome architecture. *Mol Cell* **78**(3): 554-565 e557.

Larkin A, Marygold SJ, Antonazzo G, Attrill H, dos Santos G, Garapati PV, Goodman Joshua L, Gramates L S, Millburn G, Strelets VB et al. 2021. FlyBase: updates to the *Drosophila melanogaster* knowledge base. *Nucleic Acids Res* **49**(D1): D899-D907.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R. 2009. The sequence alignment/map format and SAMtools. *Bioinformatics* **25**(16): 2078-2079.

Li W, Godzik A. 2006. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* **22**(13): 1658-1659.

Lock A, Rutherford K, Harris MA, Hayles J, Oliver SG, Bähler J, Wood V. 2019. PomBase 2018: user-driven reimplementation of the fission yeast database provides rapid and intuitive access to diverse, interconnected information. *Nucleic Acids Res* **47**(D1): D821-D827.

Merchant SS Prochnik SE Vallon O Harris EH Karpowicz SJ Witman GB Terry A Salamov A Fritz-Laylin LK Maréchal-Drouard L et al. 2007. The *Chlamydomonas* genome reveals the evolution of key animal and plant functions. *Science* **318**(5848): 245-250.

Mondo SJ, Dannebaum RO, Kuo RC, Louie KB, Bewick AJ, LaButti K, Haridas S, Kuo A, Salamov A, Ahrendt SR et al. 2017. Widespread adenine N^6-methylation of active genes in fungi. *Nature Genetics* **49**(6): 964-968.

Price MN, Dehal PS, Arkin AP. 2010. FastTree 2 – approximately maximum-likelihood trees for large alignments. *PLoS ONE* **5**(3): e9490.

Ramírez F, Ryan DP, Grüning B, Bhardwaj V, Kilpert F, Richter AS, Heyne S, Dündar F, Manke T. 2016. deepTools2: a next generation web server for deep-sequencing data analysis. *Nucleic Acids Res* **44**(W1): W160-W165.

Ruzicka L, Howe DG, Ramachandran S, Toro S, Van Slyke CE, Bradford YM, Eagle A, Fashena D, Frazer K, Kalita P et al. 2019. The Zebrafish Information Network: new support for non-coding genes, richer Gene Ontology annotations and the Alliance of Genome Resources. *Nucleic Acids Res* **47**(D1): D867-D873.

Sheng Y, Duan L, Cheng T, Qiao Y, Stover NA, Gao S. 2020. The completed macronuclear genome of a model ciliate *Tetrahymena thermophila* and its application in genome scrambling and copy number analyses. *Sci China Life Sci* **63**(10): 1534-1542.

Skinner ME, Uzilov AV, Stein LD, Mungall CJ, Holmes IH. 2009. JBrowse: A next-generation genome browser. *Genome Res* **19**(9): 1630-1638.

Stover NA, Krieger CJ, Binkley G, Dong Q, Fisk DG, Nash R, Sethuraman A, Weng S, Cherry JM. 2006. *Tetrahymena* Genome Database (TGD): a new genomic resource for *Tetrahymena thermophila* research. *Nucleic Acids Res* **34**(suppl_1): D500-D503.

Swart EC, Bracht JR, Magrini V, Minx P, Chen X, Zhou Y, Khurana JS, Goldman AD, Nowacki M, Schotanus K et al. 2013. The *Oxytricha trifallax* macronuclear genome: a complex eukaryotic genome with 16,000 tiny chromosomes. *PLOS Biol* **11**(1): e1001473.

Wang G, Wang S, Chai X, Zhang J, Yang W, Jiang C, Chen K, Miao W, Xiong J. 2021. A strategy for complete telomere-to-telomere assembly of ciliate macronuclear genome using ultra-high coverage Nanopore data. *Comput Struct Biotechnol J* **19**: 1928-1932.

Wang Y, Chen X, Sheng Y, Liu Y, Gao S. 2017. N^6-adenine DNA methylation is associated with the linker DNA of H2A.Z-containing well-positioned nucleosomes in Pol II-transcribed genes in *Tetrahymena*. *Nucleic Acids Res* **45**(20): 11594-11606.

Wickham H. 2016. *ggplot2: elegant graphics for data analysis*. O' Reilly Press, USA.

Xiong J, Gao S, Dui W, Yang W, Chen X, Taverna SD, Pearlman RE, Ashlock W, Miao W, Liu Y. 2016. Dissecting relative contributions of cis- and trans-determinants to nucleosome distribution by comparing *Tetrahymena* macronuclear and micronuclear chromatin. *Nucleic Acids Res* **44**(21): 10091-10105.

Ye J, McGinnis S, Madden TL. 2006. BLAST: improvements for better sequence analysis. *Nucleic Acids Res* **34**(Web Server issue): W6-9.

Zhang Y, Liu T, Meyer CA, Eeckhoute J, Johnson DS, Bernstein BE, Nusbaum C, Myers RM, Brown M, Li W. 2008. Model-based analysis of ChIP-Seq (MACS). *Genome Biol* **9**(9): 1-9.

Zhou Y, Fu L, Mochizuki K, Xiong J, Miao W, Wang G. 2022. Absolute quantification of chromosome copy numbers in the polyploid macronucleus of Tetrahymena thermophila at the single-cell level. *J Eukaryot Microbiol* **69**(4): e12907.