

# Direct enzymatic sequencing of 5-methylcytosine at single-base resolution

Received: 7 June 2022

Accepted: 17 March 2023

Published online: 15 June 2023

 Check for updates

Tong Wang<sup>1</sup>, Johanna M. Fowler<sup>2</sup>, Laura Liu<sup>2</sup>, Christian E. Loo<sup>1</sup>, Meiqi Luo<sup>2</sup>, Emily K. Schutsky<sup>1</sup>, Kiara N. Berríos<sup>1</sup>, Jamie E. DeNizio<sup>1</sup>, Ashley Dvorak<sup>3</sup>, Nick Downey<sup>3</sup>, Saira Montermoso<sup>1</sup>, Bianca Y. Pingul<sup>1</sup>, MacLean Nasrallah<sup>4</sup>, Walraj S. Gosal<sup>5</sup>, Hao Wu<sup>6,7</sup> & Rahul M. Kohli<sup>1,2,8</sup> ✉

5-methylcytosine (5mC) is the most important DNA modification in mammalian genomes. The ideal method for 5mC localization would be both nondestructive of DNA and direct, without requiring inference based on detection of unmodified cytosines. Here we present direct methylation sequencing (DM-Seq), a bisulfite-free method for profiling 5mC at single-base resolution using nanogram quantities of DNA. DM-Seq employs two key DNA-modifying enzymes: a neomorphic DNA methyltransferase and a DNA deaminase capable of precise discrimination between cytosine modification states. Coupling these activities with deaminase-resistant adapters enables accurate detection of only 5mC via a C-to-T transition in sequencing. By comparison, we uncover a PCR-related underdetection bias with the hybrid enzymatic-chemical TET-assisted pyridine borane sequencing approach. Importantly, we show that DM-Seq, unlike bisulfite sequencing, unmasks prognostically important CpGs in a clinical tumor sample by not confounding 5mC with 5-hydroxymethylcytosine. DM-Seq thus offers an all-enzymatic, nondestructive, faithful and direct method for the reading of 5mC alone.

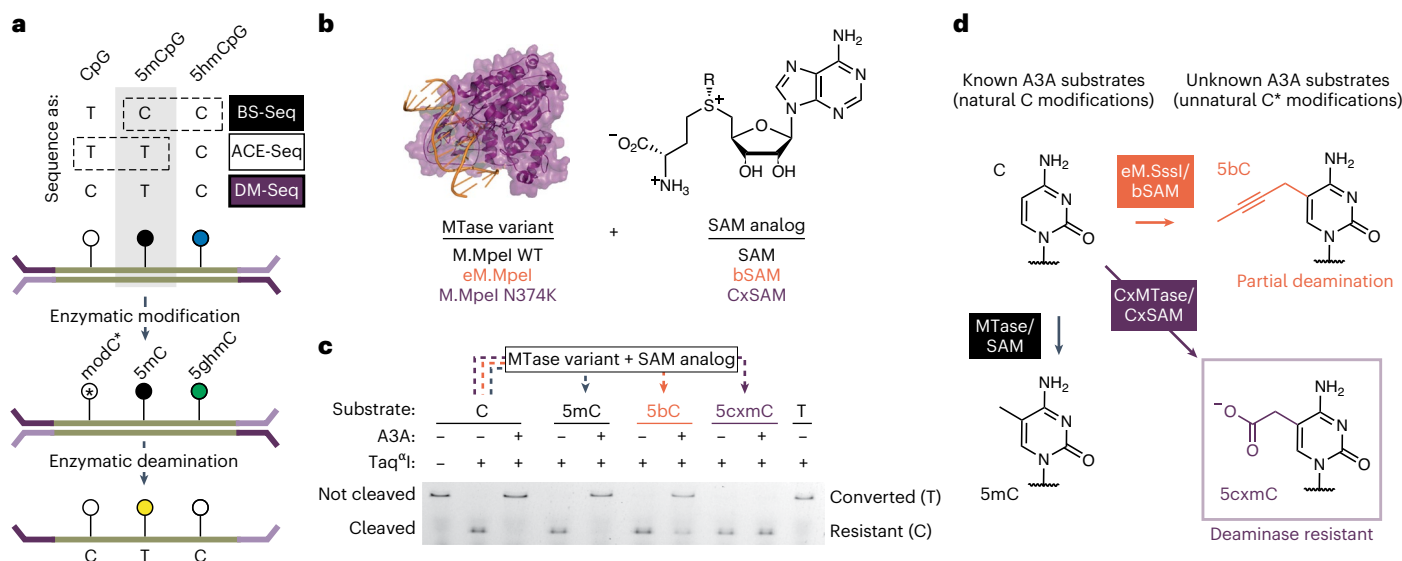
The methylation of cytosine bases in cytosine-guanine (CpG) dinucleotides is critical for diverse biological processes including gene expression, imprinting and the suppression of mobile genetic elements<sup>1</sup>. Given its role in shaping transcriptional programs, the landscape of 5mC can define cell lineages<sup>2</sup>, and dysregulation of cytosine methylation is a hallmark of diseases such as cancer<sup>3,4</sup>.

Mapping 5mC has most commonly been accomplished using chemical deamination methods. In bisulfite-based sequencing (BS-Seq), the reaction of sodium bisulfite with unmodified cytosines results in a C-to-T transition. As 5mC reacts slowly with bisulfite, its presence can be indirectly inferred by bases that remain as a C (ref. 5). BS-Seq, however, poses at least two limitations that impact our ability

to resolve 5mC. First, bisulfite damages DNA<sup>6</sup>. Second, the indirect inference of 5mC has created major challenges. As one striking example, 5-hydroxymethylcytosine (5hmC), the product of TET-mediated oxidation of 5mC, also remains as a C in BS-Seq<sup>7</sup>. Reliance on BS-Seq was a major reason why 5hmC had escaped detection for decades although it accounts for more than 20% of modified cytosines in some cell types<sup>8</sup>. Parsing 5mC and 5hmC is functionally important. For example, in human glioblastomas (GBMs) these modifications can have antagonistic functions on gene expression and impact prognoses<sup>9</sup>.

Several recently developed epigenetic sequencing technologies partially address the limitations associated with BS-Seq. Important advances have come from both bisulfite-dependent and

<sup>1</sup>Graduate Group in Biochemistry and Molecular Biophysics, University of Pennsylvania, Philadelphia, PA, USA. <sup>2</sup>Department of Medicine, University of Pennsylvania, Philadelphia, PA, USA. <sup>3</sup>Integrated DNA Technologies, Inc., Coralville, IA, USA. <sup>4</sup>Department of Pathology, University of Pennsylvania, Philadelphia, PA, USA. <sup>5</sup>Cambridge Epigenetics, Saffron Walden, United Kingdom. <sup>6</sup>Department of Genetics, University of Pennsylvania, Philadelphia, PA, USA. <sup>7</sup>Epigenetics Institute, University of Pennsylvania, Philadelphia, PA, USA. <sup>8</sup>Department of Biochemistry and Biophysics, University of Pennsylvania, Philadelphia, PA, USA. ✉e-mail: [rkohli@pennmedicine.upenn.edu](mailto:rkohli@pennmedicine.upenn.edu)



**Fig. 1 | DM-Seq is enabled by 5cxmC generation.** **a**, Top: sequencing methods for localizing C, 5mC and 5hmC differ in their use of chemical (for example, BS-Seq) or enzymatic (for example, ACE-Seq) deamination, with 5mC signal confounded by either 5hmC or C (boxes with dashed lines). Bottom: proposed workflow for DM-Seq. DM-Seq was envisioned as an all-enzymatic workflow for the direct detection of only 5mC. This goal could be realized by coupling an engineered DNA MTase (MTase\*) with a SAM analog to create a sterically bulky cytosine base that resists deamination by A3A. C\*, modified C generated from MTase and SAM analog; 5ghmC, glucosylated 5hmC. In the proposed workflow, only 5mC is converted to T at CpG sites. **b**, MTase variants and SAM analogs including two candidates for DM-Seq. **c**, Restriction enzyme-coupled assay for assessing A3A deamination of

unnatural cytosine analogs. An oligonucleotide with a single Taq<sup>q</sup>I restriction site (TCGA) is modified by the appropriate MTase variant to create 5mC, 5bC or 5cxmC (dashed lines). Modified DNA is then deaminated by A3A. Taq<sup>q</sup>I only cleaves DNA if C is protected from A3A deamination. Experiment was performed twice with similar results. See Extended Data Fig. 2 for assay schematic and ESI-MS validation of 5bC and 5cxmC substrates. **d**, Summary of reactivity of various cytosine derivatives towards A3A. Left: cytosine (C) is modified to 5mC by WT MTases and SAM. Both C and 5mC are favorable substrates for enzymatic deamination by A3A. Right: structures of 5bC and 5cxmC, with previously uncharacterized reactivity towards A3A. Box: 5cxmC satisfies the criteria required for the DM-Seq strategy: efficient MTase\* transfer and complete protection from A3A deamination.

bisulfite-independent techniques (Extended Data Fig. 1)<sup>10</sup>. For example, resolving 5mC and 5hmC is possible with oxidative bisulfite sequencing (oxBS-Seq), although low-microgram DNA input remains an impractical limitation for many samples<sup>11,12</sup>. The chemical deamination method TET-assisted pyridine borane sequencing (TAPS) can also map modified cytosines<sup>13,14</sup>. TAPS requires efficient oxidation of 5mC by TET enzymes as well as chemical conversion to the nonaromatic nucleobase analog dihydrouracil (DHU). In parallel, methods employing enzymatic rather than chemical deamination have been developed. Enzymes are ideal tools for epigenetic sequencing as they are accurate and nondestructive<sup>10</sup>. APOBEC-coupled epigenetic sequencing (ACE-Seq) was the first technology employing a DNA deaminase, APOBEC3A (A3A), to selectively deaminate unmodified Cs and 5mCs, while leaving protected 5hmCs unconverted<sup>15</sup>. Similarly, enzymatic-methylation sequencing (EM-Seq) utilizes TET and  $\beta$ -glucosyltransferase ( $\beta$ GT) enzymes before enzymatic deamination to achieve a readout akin to bisulfite merging 5mC and 5hmC (ref. 16).

Despite the promise of enzymatic deamination approaches, to date, these methods have been limited by the fact that both C and 5mC are deaminated by A3A. Motivated to develop an accurate and nondestructive pipeline for studying 5mC alone, we envisioned that direct 5mC detection could be achieved if unmodified CpGs could be protected, leaving only 5mC subject to deamination in the CpG context (Fig. 1a). Towards this goal, we considered the possibility of pairing an engineered methyltransferase (MTase\*) with an S-adenosyl-L-methionine (SAM) analog (Fig. 1b) to create a modified cytosine base resistant to A3A deamination, creating a direct strategy for localizing 5mCs within genomes.

Here, we describe how the full DM-Seq approach was realized to directly sequence 5mC alone. Additional comparisons with TAPS uncovered unexpected biases in 5mC detection, likely a result of poor

polymerase amplification of the DHU base. Ultimately, using nanogram quantities of input DNA, we show that DM-Seq outperforms BS-Seq in terms of both sequencing coverage and accurate quantification of 5mC, including at prognostically important CpG sites within a human GBM tumor.

## Results

### 5-carboxymethylcytosine is a candidate modified base for DM-Seq

Having previously exploited two classes of cytosine-modifying enzymes (glucosyltransferases and AID/APOBEC deaminases) to build ACE-Seq<sup>15</sup>, we envisioned that the addition of an MTase\* could enable DM-Seq. In evaluating the feasibility of this idea, we were encouraged for two reasons. First, wild-type (WT) CpG-specific MTases have been applied in sequencing to detect modifications that are significantly more sparse than 5mC (ref. 17). Second, multiple MTases have been engineered to transfer extended alkyl chains that could feasibly render previously unmodified CpGs resistant to enzymatic deamination. One example is a Q142A/N370A mutant of the CpG MTase M.SssI (eM.SssI) that has been shown to utilize multiple SAM analogs, including but-2-ynyl-SAM (bSAM)<sup>18,19</sup>. We also recently discovered a CpG-specific carboxymethyltransferase (CxMTase), M.MpeI N374K, which uses the naturally occurring *Escherichia coli* metabolite, carboxy-S-adenosyl-L-methionine (CxSAM), to form 5-carboxymethylcytosine (5cxmC)<sup>20</sup>.

There would be at least two requirements for DM-Seq to succeed: efficient transfer of the protecting group to unmodified CpGs and complete protection of the newly generated modified base from A3A-mediated deamination. We reasoned that the second requirement was critical to prioritize because transfer could be improved while deamination is difficult to prevent. We therefore focused analysis on two candidate MTase\* and SAM-analog pairs, an M.MpeI Q136A/N374A

variant (eM.MpeI, analogous to eM.SssI) and bSAM, as well as our CxMTase and CxSAM. Using an oligonucleotide containing a single CpG site embedded within a Taq<sup>q</sup>I (TCGA) restriction site, we demonstrated that each MTase\* and SAM-analog pair resulted in efficient conversion of the CpG to a modified CpG (Extended Data Fig. 2), yielding DNA with 5-(but-2-ynyl)-cytosine (5bC) or 5cxmC, respectively. We then subjected the oligonucleotides to enzymatic deamination with A3A and analyzed for deamination by restriction cleavage with Taq<sup>q</sup>I (Fig. 1c). While C and 5mC are readily deaminated by A3A, we show that 5bC can be partially deaminated by A3A, and 5cxmC, which has features of both size and negative charge that can be disfavored by A3A (refs. 21,22), appears to resist enzymatic deamination (Fig. 1d). The promising properties of 5cxmC led us to focus on the CxMTase:CxSAM enzyme:substrate pair for further development.

### DNA carboxymethylation is limited by opposite-strand biases

Having identified our candidate MTase\* and SAM-analog pair, we next aimed to assess the efficiency of DNA carboxymethylation using sequencing. We sheared unmodified 48.5-kilobase (kb) lambda phage genomic DNA (gDNA) and ligated Illumina Y-shaped adapters containing 5mC bases. These samples were subjected to WT M.MpeI or M.MpeI N374K and different SAM substrate conditions. 5mC and 5cxmC generation was assessed by BS-Seq across the 3,113 CpG dyads. For the WT M.MpeI, in the absence of SAM, 0.2% of Cs in the CpG context were detected as modified, while in the presence of SAM, 97.2% of CpGs were modified (Fig. 2a). When M.MpeI N374K was used with SAM, the majority of CpG sites were detected as modified. However, only 48.9% of the CpGs were detected as modified, a disappointing result inconsistent with our previous restriction digestion assay that suggested complete DNA carboxymethylation<sup>20</sup>.

To understand the mechanistic basis for inefficient transfer, we focused on understanding CpGs within the same dyad but on opposite strands, as some MTases, such as DNMT1, are impacted by opposite-strand modified cytosines<sup>23</sup>. We found that M.MpeI N374K transfer of SAM was mostly symmetrically clustered (purple), suggesting no strong influence of the target-strand on the opposite-strand within the same CpG dyad (Fig. 2b). In contrast, with CxSAM, many CpGs were asymmetrically modified (yellow). These data are consistent with a model where the first carboxymethylation event at a CpG dyad is efficient, but a second carboxymethylation event on the opposite-strand is slow (Fig. 2c, bottom).

To understand the impact of the opposite-strand modifications that appeared to be limiting to our initial DM-Seq strategy, we devised an oligonucleotide assay (Extended Data Fig. 3). In this assay, we perform the MTase\* reaction using a fluorophore-labeled top-strand containing a single CpG embedded within a methylation-sensitive HpaII (CCGG) restriction site and duplexed to a chemically synthesized bottom-strand containing either an unmodified CpG or 5mCpG. After strand exchange with an excess of an unmodified bottom-strand, digestion with HpaII can detect top-strand modification. When the duplex was reacted with M.MpeI N374K and SAM, the top-strand CpG could be readily modified opposite either an unmodified CpG or 5mCpG (Fig. 2c). Critically, M.MpeI N374K could fully transfer CxSAM across from a 5mCpG but only partially with an opposite-strand unmodified CpG, recapitulating our sequencing results.

### 5-propynylcytosine adapters enable efficient carboxymethylation

Our biochemistry unveiled that the presence of a hemimethylated CpG is particularly favorable for DNA carboxymethylation. This finding offered us a potential solution to realize our initial objective. In our envisioned workflow, after adapter ligation, a primer complementary to the adapter could initiate synthesis of a copy-strand containing all 5mCs in lieu of unmodified Cs with the resulting hemimethylated duplex favorable for CxMTase activity (Fig. 2d). In the proposed workflow, the

copy-strand is not deaminase-resistant and thus would not be amplified upon library preparation.

This alternative workflow, however, posed new potential challenges. Traditional adapters contain 5mC, which readily converts to T with enzymatic deamination. Our workflow with pre-conversion adapter ligation would require custom adapters resistant to deamination by A3A, leading us to reflect on A3A selectivity further. Biochemical studies suggest that the enzyme discriminates against bulky modifications at the 5-position of cytosine<sup>22</sup>, a feature which is borne out by structural work highlighting a 'steric gate' residue Y130 (orange) abutting the C5-C6 face of the target cytosine (Fig. 2d)<sup>24</sup>. To identify candidate analogs suitable for A3A-resistant adapters, we explored a series of dCTP analogs with increasing steric bulk at C5, including C, 5mC, 5-vinylcytosine (5vC), 5-ethynylcytosine (5eyC) and 5-propynylcytosine (5pyC)<sup>25</sup>, generating duplex DNA with exclusively modified cytosines by PCR. After reacting with A3A, the DNA was reamplified and interrogated for cleavage at a specific Taq<sup>q</sup>I restriction site within the amplicon or by deep sequencing (Extended Data Fig. 4a). Using this assay approach, amplicons with either C or 5mC show resistance to Taq<sup>q</sup>I, indicating deamination (Fig. 2e). The larger 5vC and 5eyC both showed intermediate resistance, while the 5pyC-containing template appeared resistant to A3A. In agreement with the qualitative assay, the C and 5mC templates were fully deaminated, 5eyC and 5vC incompletely deaminated, and the 5pyC substrate showed <1% deamination by sequencing. Integrating across the series of chemically modified cytosines, along with the enzymatically generated 5bC and 5cxmC, our results offer a more complete and structurally informed model for how the hybridization of C5-bond linkages, steric bulk and charge all collaborate to shape selective enzymatic deamination by A3A (Extended Data Fig. 4b,c).

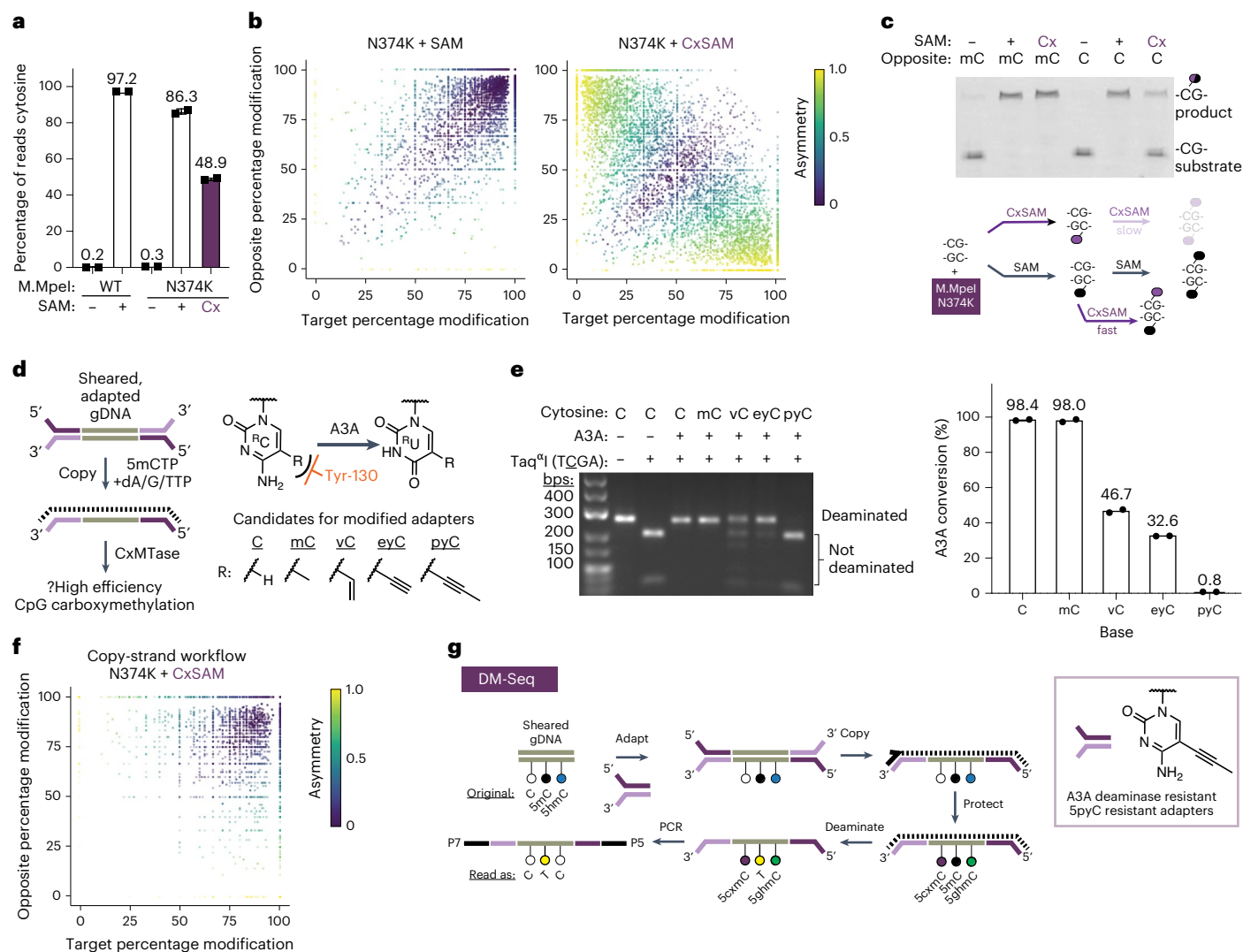
Our results support the candidacy of 5pyC as a suitable modification for use in DNA deaminase-resistant adapters, especially given its synthetic accessibility<sup>26</sup>. We therefore synthesized Illumina TruSeq Y-shaped adapters, with all Cs replaced with 5pyC bases, and validated that ligation was not impacted by the presence of 5pyC bases (Extended Data Fig. 5). We next evaluated the copy-strand workflow using sheared, unmodified lambda phage gDNA. With SAM transfer, the inclusion of the copy step had minimal impact on the efficiency of CpG protection (Extended Data Fig. 6). By contrast, CxSAM transfer improved significantly, and the asymmetric protection evident in the absence of the copy-strand was no longer present (Fig. 2f). The analogous experiment with 5pyC adapter ligation and A3A deamination also showed efficient transfer (Extended Data Fig. 6), providing a roadmap for a new DM-Seq workflow using 5pyC adapters.

### DM-Seq detection of heterogeneous samples

Building on this design, we optimized our pipeline to further improve DNA carboxymethylation efficiency. In the final DM-Seq workflow (Fig. 2g), 5pyC adapters are ligated to sheared gDNA and copied to create a strand exclusively containing 5mCs in place of C. The gDNA is then protected by the CxMTase (acting on unmodified CpGs) and glucosylation by  $\beta$ GT (for 5hmCs). Subsequent deamination by A3A is performed before PCR amplification and sequencing.

We sought to quantify the fidelity of this workflow using three lambda phage gDNA samples: native gDNA as a standard with unmodified CpGs, gDNA methylated at CpG sites with M.SssI and gDNA methylated at CpC sites with the MTase M.CviPI. Sheared gDNA samples were split and then analyzed with either 5mC-containing adapters and BS-Seq or 5pyC-containing adapters and DM-Seq (Fig. 3a). After deamination, amplifiable DNA content was 22-fold more across DM-Seq samples as compared with BS-Seq samples by quantitative PCR (qPCR) (average cycle threshold ( $C_t$ ) = 17.0 versus 12.5; Fig. 3a). Focusing next on the genome-wide comparison of methods (Fig. 3b) for the unmodified lambda phage gDNA, we found a low rate of CpG nonconversion by BS-Seq (0.23%), and a high rate of protection from deamination with DM-Seq (96.7%), validating the efficiency of the copy-strand protocol





**Fig. 2 | The challenge of asymmetric DNA carboxymethylation is overcome with copy-strand synthesis enabled by SpyC adapters.** **a**, Lambda gDNA sequencing experiment. Lambda gDNA was incubated with WT M.Mpel or M.Mpel N374K in the presence of no SAM, SAM or CxSAM. DNA was treated with bisulfite (BS) before PCR and Illumina sequencing ( $n = 2$  independent experiments). Data are presented as mean values  $\pm$  s.d. **b**, Scatter plot showing the relationship between CpGs on opposite strands within the same dyad. Data are filtered for CpGs with at least five sequencing reads. Asymmetry = |percentage target – percentage opposite| / total. **c**, Top: oligonucleotide modification assay (Extended Data Fig. 3). An oligonucleotide with a labeled top-strand containing an unmodified CpG across from either a 5mCpG or unmodified CpG was reacted with M.Mpel N374K and no SAM, SAM or CxSAM. Shown is the denaturing gel after digestion with a modification-sensitive restriction enzyme that reports on the modification status of the top-strand. The experiment was performed twice with similar results. Bottom: model for incomplete transfer. Symmetrical modification proceeds readily with SAM but is slow with CxSAM due to inefficient transfer

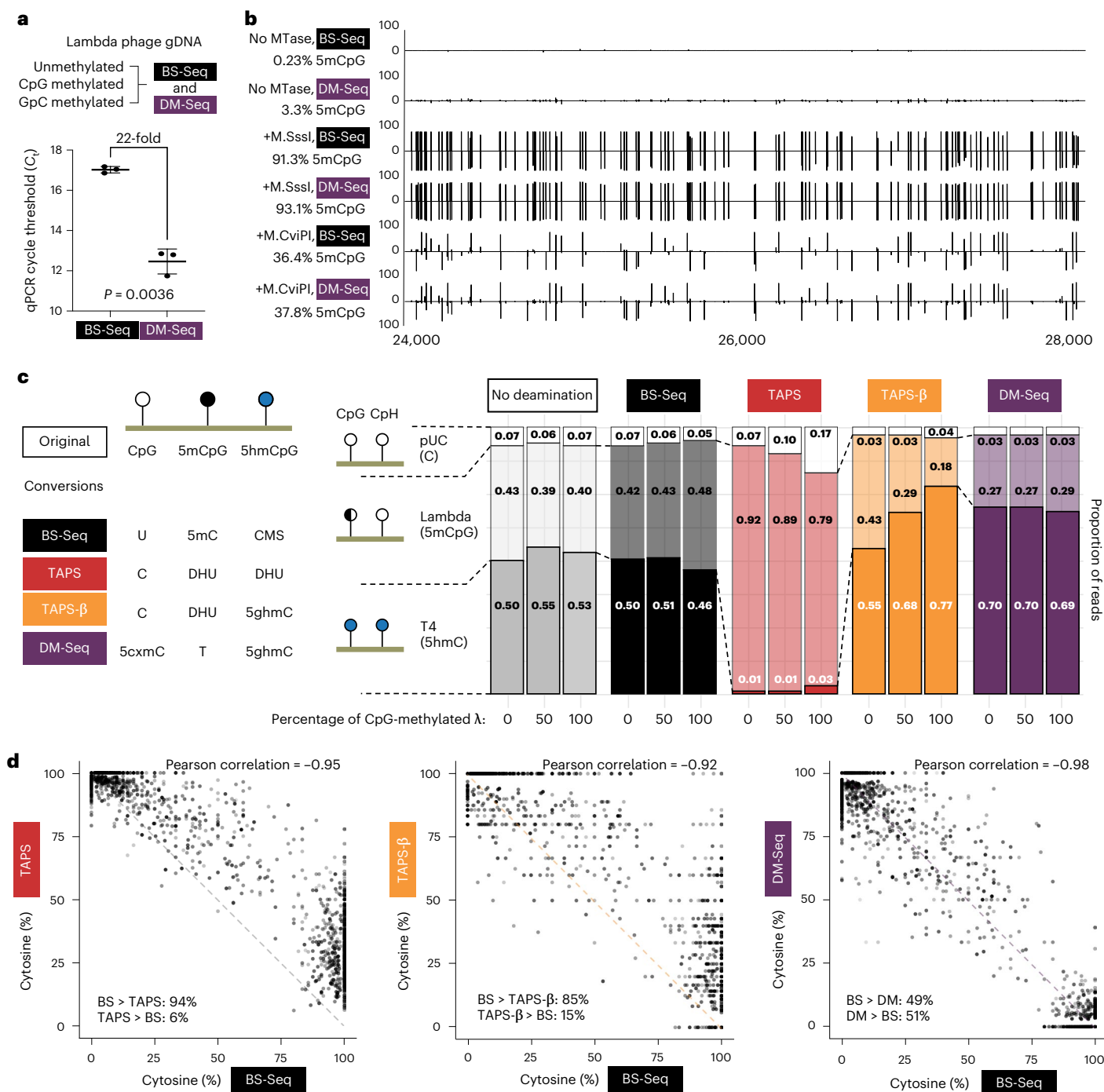
across from a 5xmC. Carboxymethylation is efficient when the opposite-strand contains a 5mC. **d**, Left: envisioned scheme where A3A-resistant adapters (purple) initiate universal copy-strand synthesis. A copy-strand (dotted line) containing 5mCs serves as a favorable substrate for DNA carboxymethylation. Right: structures of unnatural cytosine analogs explored. The 5-position modifications are anticipated to interact with the steric gate Tyr-130 residue in A3A (orange) (Extended Data Fig. 4). **e**, Left: PCR product is generated using modified dCTPs in place of dCTP. dsDNA is deaminated, PCR-amplified and restriction digested to analyze deamination status at a single TCGA Taq<sup>q1</sup> site. Right: deep sequencing of the same PCR products as in gel ( $n = 2$ ). **f**, Scatter plot as in **b** showing improvement of carboxymethylation with copy-strand synthesis. Data correspond to Extended Data Fig. 6. **g**, Full DM-Seq workflow. Sheared gDNA is end-prepped and adapted to A3A-resistant 5pyC adapters. A copy-strand made with 5mCTPs is synthesized before glucosylation and carboxymethylation. A3A deaminates 5mCpGs to Ts which can be detected upon PCR amplification. Box: 5pyC adapters are synthetically accessible and permissive for A3A-dependent sequencing.

for CpG conversion to 5xmCpG. For the gDNA sample treated with M.SssI, 91.3% of CpGs were protected from deamination with BS-Seq, with a comparable level (93.1%) deaminated by A3A in DM-Seq. In the M.CviPI MTase condition, detection of 5mCpG at the genome-wide level was similar for BS-Seq and DM-Seq (Fig. 3b). M.CviPI-treated gDNA provided an added opportunity to compare heterogeneous levels of methylation, as this enzyme is known to have detectable but variable off-target activity at CpCpG sites<sup>27</sup>. At highly methylated sites, we detected 95.4% of GpCpGs as methylated by BS-Seq and 94.5% as

methylated by DM-Seq, while off-target CpCpG showed strong correlation at the level of individual CpG sites and globally (Extended Data Fig. 7a,b).

### Comparison of DM-Seq with TAPS uncovers DHU bias

Given the burgeoning interest in bisulfite-free epigenetic sequencing technologies, we saw an opportunity to directly compare DM-Seq with other methods. In particular, we focused on TAPS with  $\beta$ GT blocking (TAPS- $\beta$ ), a variation of TAPS where 5hmCs are protected from TET



**Fig. 3 | DM-Seq accurately detects 5mCpGs at single-base resolution and is more accurate than TAPS.** **a**, Difference in  $C_t$  between DM-Seq and BS-Seq determined by qPCR.  $P$  value represents paired two-tailed  $t$ -test ( $n = 3$  MTase conditions). Data are presented as mean values  $\pm$  s.d. **b**, Shown is the genome browser view for coordinates 24,000–28,000 in the lambda phage genome for all CpGs. Lambda gDNA was modified with SAM and no MTase, M.SssI (CpG) or M.CviPI (GpC). Numbers on the left represent total efficiency across the entire 48.5-kb genome. **c**, Comparison of multiple deamination-dependent sequencing workflows. At the left is a schematic showing the state of specific DNA modifications after conversion and before library generation (CMS, cytosine methylene sulfonate). A mixture of three sheared DNA samples:

unmodified pUC19 DNA, variably methylated lambda phage (0%, -50% or 100% CpG-methylated) and T4-hmC phage (with all C bases replaced with 5hmC), was subjected to no deamination, BS-Seq, TAPS, TAPS- $\beta$  or DM-Seq workflows. Plotted is the distribution of reads mapping to each genome under each condition, with the read fraction listed. **d**, Correlation of BS-Seq to TAPS, TAPS- $\beta$  and DM-Seq on an M.CviPI GpC-methylated substrate. The dashed line shows the readout if BS-Seq signal inversely correlates with TAPS, TAPS- $\beta$  or DM-Seq as anticipated, with skew between methods suggested by asymmetrical distribution around this line. In the bottom corner of each plot, for sites where the two methods are not in agreement, the percentage of sites where one method detects a higher level of modification than the other method is given.

oxidation, resulting in the chemical reductive deamination of only 5mCs to DHU<sup>14</sup>. To this end, we combined unmethylated pUC19 plasmid, CpG-methylated lambda gDNA and T4-hmC phage gDNA to quantify the

behavior of C, 5mC and 5hmC in a single mixture. To assess correlations between methods, we included 0% methylated, -50% CpG-methylated, 100% CpG-methylated or 100% GpC-methylated lambda substrate.

The ~50% CpG-methylated substrate was made by combining 0% methylated and 100% CpG-methylated lambda DNA. We then subjected these substrates to five possible conditions: no deamination, BS-Seq, TAPS, TAPS- $\beta$  or DM-Seq.

We first examined the relative proportion of reads mapping to each of the three DNA genomes (Fig. 3c) and noted two unexpected trends unique to the borane-based technologies. First, and most strikingly, TAPS shows nearly complete obliteration of mapping of the T4-hmC phage (dark red, 1–3% of reads compared with ~50% in the no deamination control). Notably, in the T4-hmC phage, all of the Cs including those in non-CpG (CpH) contexts are modified 5hmC bases, thus resulting in a high density of DHU. These findings contrast with TAPS- $\beta$ , where the 5hmCs are protected from conversion to DHU and T4-hmC reads are detectable. Second, unlike BS-Seq and DM-Seq, both TAPS and TAPS- $\beta$  show depletion of the lambda gDNA reads as a function of the level of CpG methylation, with 43% of the total TAPS- $\beta$  reads mapping to the lambda genome when unmethylated but only 18% when fully methylated.

We next focused on the GpC-methylated gDNA (Fig. 3d). While, DM-Seq and BS-Seq strongly correlate (Pearson coefficient =  $-0.98$ ), TAPS and TAPS- $\beta$  showed weaker correlations with BS-Seq and additionally showed consistent skew in the data. While DM-Seq is equally likely to detect that an individual CpG is more or less methylated relative to BS-Seq, for TAPS and TAPS- $\beta$ , 5mC levels are lower than those detected by BS-Seq at the vast majority of sites (94% or 85% of individual CpGs sites, respectively) (Fig. 3d and Extended Data Fig. 7c). Importantly, we note that while underestimation of 5mC could be explained partially by incomplete conversion efficiencies (discussed below), decreased coverage of the T4-hmC and lambda genomes with more highly modified DNA cannot be.

We sought to further investigate the mechanisms responsible for 5mC detection bias in TAPS. We considered that DHU-containing DNA might be less efficiently amplified, as this nonplanar, nonaromatic analog of uracil has been shown to stall numerous polymerases including the PacBio SMRT polymerase<sup>13,28</sup> and to disrupt base stacking in nucleic acids<sup>29,30</sup>. With CpG-methylated lambda gDNA, while BS-Seq and DM-Seq detected 96.5% and 98.1% modification, respectively, TAPS and TAPS- $\beta$  reported ~84% methylation (Extended Data Fig. 7d). However, with the mixed unmethylated and methylated lambda gDNA sample, while BS-Seq and DM-Seq correlated with one another (64.0% versus 59.3%), both TAPS and TAPS- $\beta$  significantly underestimated the methylation levels (20.3% and 19.9%). The fact that 5mC was underestimated much more significantly with the mixed lambda gDNA sample than with the fully methylated lambda sample is consistent with preferential amplification of the unmethylated lambda gDNA relative to the modified, DHU-containing lambda gDNA. To further understand both read depletion and 5mC detection accuracy in TAPS, we additionally used our samples with fully methylated lambda gDNA and performed TET oxidation followed by BS-Seq or A3A treatment (to measure 5-formylcytosine (5fC)/5-carboxylcytosine (5caC) levels) or borane-mediated deamination (Extended Data Fig. 8). This analysis supports the conclusion that the observed bias in amplification and 5mC detection requires both TET-mediated oxidation of 5mC and borane-mediated deamination, consistent with DHU as the source of bias.

Although our conversion efficiency values are better than those published by other groups who have attempted to replicate TAPS<sup>31</sup>, our values (range 84–92% across experiments) remained lower than those reported by the group that initially advanced TAPS (>97%)<sup>13</sup>. We reasoned that if bias was related to DHU generation and not incomplete conversion efficiency, there should also be evidence of bias in published datasets. We elected to further investigate an established matched mouse embryonic stem cell (ESC) dataset where TAPS was optimized and BS-Seq was also performed<sup>13</sup>. When we reanalyzed nonoverlapping 1-kb bins across the genome, a modest correlation between TAPS and BS-Seq could be observed (Pearson correlation

$-0.745$ ; Extended Data Fig. 9a). Consistent with our data, TAPS underestimates modification levels at 66.5% of these bins relative to BS-Seq (Extended Data Fig. 9b). We further considered whether bias was a function of modification percentage, as we observed with lambda gDNA samples. Indeed, lowly modified 1-kb bins have an equal probability of being underestimated or overestimated by TAPS, while TAPS detects ~21% lower levels on average than BS-Seq with bins where >90% of CpGs are modified (Extended Data Fig. 9c). Finally, we considered whether methylated regions in *cis* would be especially prone to TAPS bias in ESCs. Indeed, when examining an established set of imprinting control regions (ICRs)<sup>15</sup>, we find that TAPS detected lower levels of modification at 28 of 29 ICRs, with an average of 41.9% CpG modification detected by BS-Seq and 31.6% by TAPS (Extended Data Fig. 9d, 9e). Interestingly, the level of deviation between BS-Seq and TAPS increased as a function of CpG density at the ICRs (Pearson coefficient =  $-0.65$ ). These data show that TAPS modification bias is reproduced in an existing mammalian dataset generated by an independent research group, with multiple trends consistent with DHU being responsible for the observed bias.

### DM-Seq is superior to BS-Seq in characterizing a human tumor

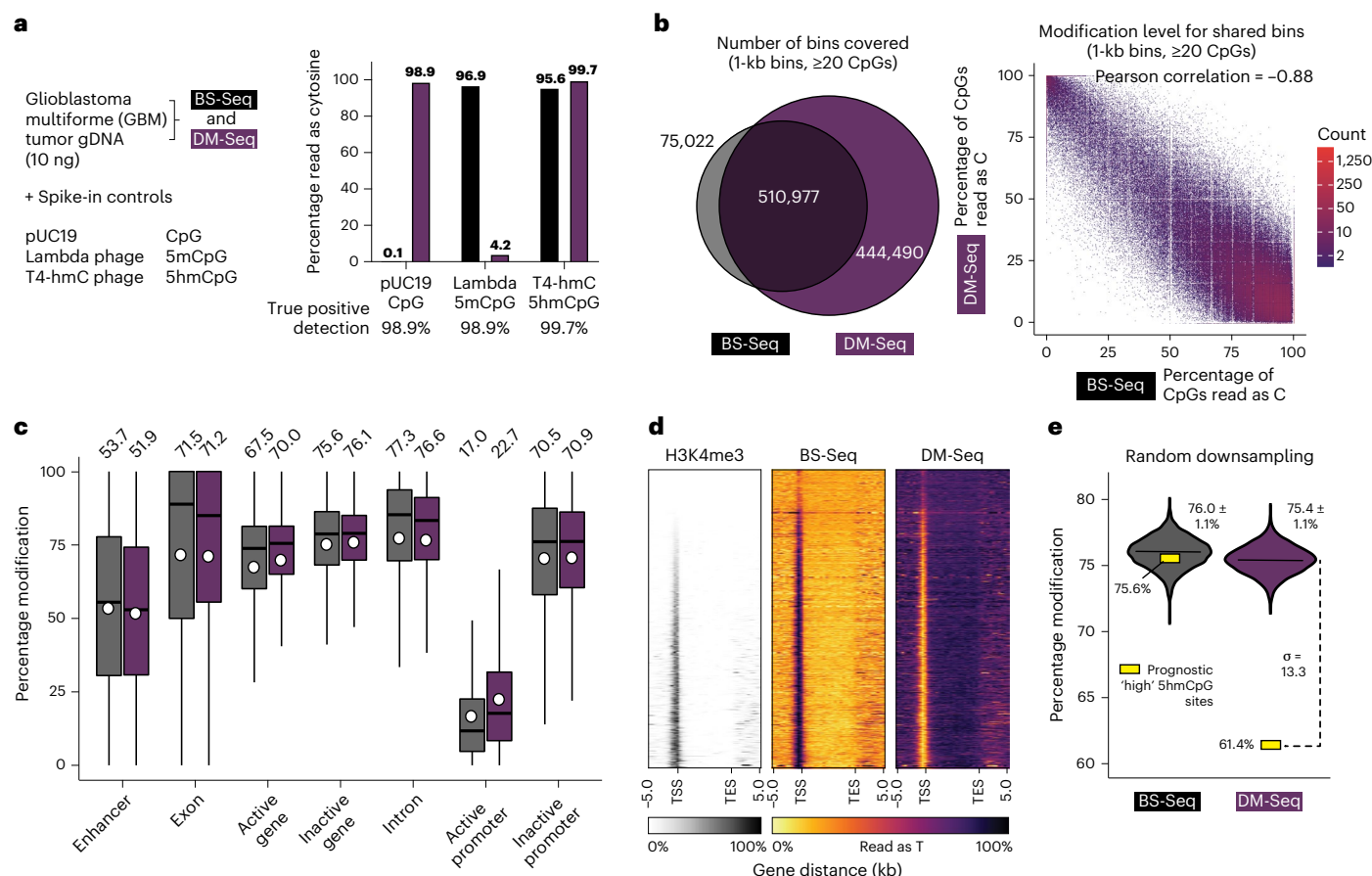
We chose to apply DM-Seq on a human GBM sample because this cancer has been extensively characterized for its heterogeneous cytosine methylation patterns<sup>32,33</sup>. Although mammalian brain tissue is typically enriched with 5hmC, two independent studies utilizing oxBS-Seq concluded that 5hmC is highly depleted in GBMs<sup>9,34</sup>. GBMs thus offer a complex mammalian genome where DM-Seq and BS-Seq can be directly compared with limited interference from 5hmC, which BS-Seq cannot parse. At the same time, despite relatively low overall abundance of 5hmC, 5hmC at a limited set of CpG sites has been implicated as an important disease biomarker<sup>9</sup>, offering the possibility that direct detection of 5mC with DM-Seq could provide a prognostically relevant signal through the accurate sequencing of 5mC alone.

We obtained gDNA from a surgically resected human GBM and added our three spike-in controls to validate DM-Seq efficiency (Fig. 4a). We used >35–100-fold less DNA input than previously used to characterize GBM by oxBS-Seq<sup>9,34</sup>. The sample was sheared, evenly split and processed by either DM-Seq or BS-Seq pipelines. Overall, the nondestructive nature of DM-Seq was evident in the generated libraries, with a 2.8 $\times$  greater library yield and a greater average library size (447 versus 346 base pairs (bp)) (Extended Data Fig. 10). Despite the higher yield with DM-Seq, for rigorous comparison of the two methods, we normalized the libraries and aimed to sequence equally, targeting  $\sim 1\times$  coverage on a single Illumina NextSeq run (223,862,027 reads for DM-Seq and 223,430,253 reads for BS-Seq).

Analysis of the BS-Seq spike-ins showed accurate conversion of unmodified Cs as Ts while 5mC and 5hmC were detected as Cs (Fig. 4a). With DM-Seq, CpGs in the unmodified pUC DNA sample were 98.9% sequenced as cytosine, validating efficient generation of 5cxmC and its protection from deamination by A3A. Efficient deamination of 5mC was confirmed, with 95.8% of the CpGs in the *in vitro* methylated lambda gDNA sample reading as T relative to 96.9% expected based on BS-Seq. The T4-hmC sample was protected from deamination, with 99.7% of bases reading as C. Thus, the true positive detection rate for each of the three modified bases was 98.9% or higher. This accurate detection of the pUC19 and T4-hmC spike-in controls specifically provides strong evidence that copy-strands containing all 5mCs are not being measurably amplified in DM-Seq.

Despite compensating for the 2.8-fold lower library yields with BS-Seq, DM-Seq still provided more information content than BS-Seq in the normalized sequencing libraries (Extended Data Fig. 10). DM-Seq captured 5.9-fold more 1-kb nonoverlapping, unique bins with at least 20 CpGs (444,490 versus 75,022) as compared with BS-Seq (Fig. 4b). Given that the majority of BS-Seq signal comes from 5mC and not 5hmC in GBM, we next explored correlations between the two datasets. Focusing on the





**Fig. 4 | DM-Seq directly detects 5mCpGs in human GBM.** **a**, Experimental design with spike-in controls showing accuracy of C, 5mC and 5hmC detection. **b**, Binned CpG analysis using nonoverlapping 1-kb bins with at least 20 CpGs covered. Left: Venn diagram showing bins covered by BS-Seq and DM-Seq. Right: correlation between DM-Seq and BS-Seq in the 510,977 shared bins. **c**, Percentage cytosine modification at various genomic features. The box shows the lower quartile, median and upper quartile. Minimum and maximum values are shown by the whiskers. Circles are the mean values displayed above each boxplot. **d**, Heatmap representation of all annotated genes for H3K4me3 ChIP-Seq, BS-Seq and DM-Seq. Genes are ranked by their average H3K4me3 signal. TES,

transcription end site. **e**, Observed DM-Seq and BS-Seq signal at 3,876 previously defined 'high 5hmCpG sites' (yellow square, DM-Seq: 61.4%; BS-Seq: 75.6%). The violin plot shows data from the shared BS-Seq and DM-Seq CpGs randomly downsampled 10,000 times to the same coverage as BS-Seq and DM-Seq at these sites. Data represent mean  $\pm$  s.d. (BS-Seq =  $76.0 \pm 1.1\%$ ; DM-Seq =  $75.4 \pm 1.1\%$ ). The dotted line shows the number of standard deviations (13.3) between the downsampled (violin) and observed (yellow box) data at these prognostically significant CpGs. Extended data from these downsamplings are shown in Supplementary Table 4 and Extended Data Fig. 10.

510,977 shared bins showed a strong inverse correlation between signals for BS-Seq (5mC + 5hmC) and DM-Seq (5mC alone) (Fig. 4b; Pearson =  $-0.88$ ). Given the predominance of 5mC in GBM, we found that profiles generated by the two methods track with one another across various genomic elements (Fig. 4c). Notably, DM-Seq signals were distinct for active and inactive promoters ( $22.7 \pm 19.1\%$  versus  $70.9 \pm 20.7\%$ ), as defined by H3K4me3 chromatin immunoprecipitation followed by sequencing (ChIP-Seq), with patterns that are mirrored relative to BS-Seq<sup>35</sup>, a relationship that extends when rank-ordering the enrichment level of H3K4me3 signal (Fig. 4d).

Having established that DM-Seq and BS-Seq are strongly correlated, we next sought to investigate the utility of direct 5mC mapping with DM-Seq by focusing on a limited subset of CpGs that have been shown to harbor DNA modifications of prognostic value. In previous work using oxBS-Seq microarrays, a candidate list of 3,876 CpGs were identified as the top 1% 5hmC sites across 30 GBMs, with an average 5hmC level of 10.1% (ref. 9). These 'high 5hmC sites' were disproportionately enriched in certain genomic elements, correlated with gene expression and could be used to predict a 3.3-fold difference in patient survival. In our GBM tumor, 2,538 and 2,132 of these sites were sequenced by DM-Seq and BS-Seq, respectively, with 1,485

CpGs sequenced across both datasets (Extended Data Fig. 10). At these common sites, DM-Seq reported 61.4% 5mC modification, 14.3% lower than the 75.6% modification level observed by BS-Seq, quantifying the 'blind spot' that BS-Seq harbors to 5hmC at functionally important CpGs within this cancer (Fig. 4e). To determine if the 'low 5mC' level at these CpG sites could serve as a distinct signal, akin to the 'high 5hmC' detected by oxBS-Seq, we performed multiple downsamplings of 1,485 random CpG sites from either BS-Seq or DM-Seq (Fig. 4e). While the measured value for the 'high 5hmC sites' fell within the expected range with BS-Seq ( $76.0 \pm 1.1\%$ ), these CpGs were major outliers in DM-Seq ( $75.4 \pm 1.1\%$ ). These results highlight how direct detection of 5mC from DM-Seq, rather than a pooled 5mC/5hmC signal from BS-Seq, could advance efforts to sequence prognostically significant CpGs.

## Discussion

Here, we describe DM-Seq—a nondestructive and enzyme-only workflow for directly detecting 5mCpGs at single-base resolution. To create DM-Seq, we originally envisioned employing an engineered MTase<sup>\*</sup>:SAM-analog enzyme:substrate pair which could create an unnatural DNA base capable of resisting enzymatic deamination by A3A (Fig. 1a). We identified the CxMTase M.Mpel N374K and CxSAM

pair as a favorable pair but unexpectedly uncovered opposite-strand biases to DNA carboxymethylation that led us to employ several additional innovations. Specifically, we found that ligation of DNA deaminase-resistant adapters containing unnatural SpyC, followed by copying of gDNA with 5mC in lieu of unmodified C, created an ideal substrate for CxMTase activity. The application of our CxMTase offers an important precedent, as engineering or evolution principles can be applied to invent enzymes that expand our sequencing toolbox beyond native activities. Opportunities abound to further improve the enzyme:substrate pair in DM-Seq as there are many SAM analogs besides CxSAM<sup>36</sup>, and new DNA CxMTases could even be, in principle, evolved in vivo given the natural availability of CxSAM as a secondary metabolite in *E. coli*<sup>37</sup>.

Our work highlights how structure–activity studies on DNA-modifying enzymes can be harnessed to devise new sequencing pipelines. In existing DNA deaminase-based pipelines, the ability of A3A to discriminate against the natural DNA modifications, glucosylated 5hmC and 5caC, has been exploited. We now reveal two unnatural cytosine analogs that can be effectively used in sequencing applications: 5cxmC and SpyC (Extended Data Fig. 4). Our finding that 5cxmC is protected from enzymatic deamination is corroborated by an independent study<sup>38</sup>, although their assay design masks issues with opposite-strand modification that would have prevented whole-genome sequencing applications, as we have demonstrated here. We also anticipate added utility for SpyC, as our SpyC adapters can more generally be exploited to improve other nondestructive, DNA deaminase-dependent sequencing workflows.

While both borane and enzymatic deamination methods have been touted as nondestructive, we provide the first quantitative data, to our knowledge, that have shown on the same matched substrate that their relative effects on total DNA preservation are essentially equivalent and both superior to bisulfite (Extended Data Fig. 8d,e). This quantitative trend is likely explained by their contrasting mechanisms, where only bisulfite-catalyzed deamination requires cytosine to undergo electrophilic activation, creating unstable sulfonated intermediates prone to depyrimidation. We unexpectedly found that TAPS underestimates modified bases, a feature that may be attributed to DHU generation. Regions with a high density of DNA modifications in *cis* may be especially prone to underamplification. Although we anticipate that TAPS will remain a useful technology, as demonstrated by recent applications<sup>39</sup>, this behavior is critical to further understand, as caution should likely be taken when comparing new TAPS results with existing reference datasets based on BS-Seq or when deconvoluting mixtures of DNA. It remains to be seen whether quantitative correction for bias or the use of alternative polymerases, such as those evolved to copy across from sulfonated DHU<sup>40</sup>, could improve TAPS.

We believe that the development of DM-Seq highlights the limitations of current methods for 5mC localization, a topic of critical importance. Our GBM analysis revealed how directly sequencing 5mC with DM-Seq at specific CpGs can provide distinctive prognostic information. Extension of the technology to profiling sparse DNA samples, such as cell-free DNA for early cancer detection, can be readily imagined. Given its nondestructive and DHU-free workflow, DM-Seq could also be coupled to rapidly evolving third-generation sequencing platforms. Ultimately, by directly mapping 5mC, rather than modified cytosines in aggregate, DM-Seq can allow for the biological function of 5mC alone to be better ascertained.

## Online content

Any methods, additional references, Nature Portfolio reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at <https://doi.org/10.1038/s41589-023-01318-1>.

## References

- Schubeler, D. Function and information content of DNA methylation. *Nature* **517**, 321–326 (2015).
- Luo, C., Hajkova, P. & Ecker, J. R. Dynamic DNA methylation: in the right place at the right time. *Science* **361**, 1336–1340 (2018).
- Shen, S. Y. et al. Sensitive tumour detection and classification using plasma cell-free DNA methylomes. *Nature* **563**, 579–583 (2018).
- Hoadley, K. A. et al. Multiplatform analysis of 12 cancer types reveals molecular classification within and across tissues of origin. *Cell* **158**, 929–944 (2014).
- Frommer, M. et al. A genomic sequencing protocol that yields a positive display of 5-methylcytosine residues in individual DNA strands. *Proc. Natl Acad. Sci. USA* **89**, 1827–1831 (1992).
- Tanaka, K. & Okamoto, A. Degradation of DNA by bisulfite treatment. *Bioorg. Med. Chem. Lett.* **17**, 1912–1915 (2007).
- Huang, Y. et al. The behaviour of 5-hydroxymethylcytosine in bisulfite sequencing. *PLoS ONE* **5**, e8888 (2010).
- Kriaucionis, S. & Heintz, N. The nuclear DNA base 5-hydroxymethylcytosine is present in Purkinje neurons and the brain. *Science* **324**, 929–930 (2009).
- Johnson, K. C. et al. 5-Hydroxymethylcytosine localizes to enhancer elements and is associated with survival in glioblastoma patients. *Nat. Commun.* **7**, 13177 (2016).
- Wang, T., Loo, C. E. & Kohli, R. M. Enzymatic approaches for profiling cytosine methylation and hydroxymethylation. *Mol. Metab.* **57**, 101314 (2021).
- Booth, M. J. et al. Quantitative sequencing of 5-methylcytosine and 5-hydroxymethylcytosine at single-base resolution. *Science* **336**, 934–937 (2012).
- Yu, M. et al. Base-resolution analysis of 5-hydroxymethylcytosine in the mammalian genome. *Cell* **149**, 1368–1380 (2012).
- Liu, Y. et al. Bisulfite-free direct detection of 5-methylcytosine and 5-hydroxymethylcytosine at base resolution. *Nat. Biotechnol.* **37**, 424–429 (2019).
- Liu, Y. et al. Subtraction-free and bisulfite-free specific sequencing of 5-methylcytosine and its oxidized derivatives at base resolution. *Nat. Commun.* **12**, 618 (2021).
- Schutsky, E. K. et al. Nondestructive, base-resolution sequencing of 5-hydroxymethylcytosine using a DNA deaminase. *Nat. Biotech.* **36**, 1083–1090 (2018).
- Vaisvila, R. et al. Enzymatic methyl sequencing detects DNA methylation at single-base resolution from picograms of DNA. *Genome Res.* **31**, 1280–1289 (2021).
- Wu, H., Wu, X., Shen, L. & Zhang, Y. Single-base resolution analysis of active DNA demethylation using methylase-assisted bisulfite sequencing. *Nat. Biotechnol.* **32**, 1231–1240 (2014).
- Stasevskij, Z., Gibas, P., Gordevicius, J., Kriukiene, E. & Klimasauskas, S. Tethered oligonucleotide-primed sequencing, TOP-Seq: a high-resolution economical approach for DNA epigenome profiling. *Mol. Cell* **65**, 554–564.e6 (2017).
- Kriukiene, E. et al. DNA unmethylome profiling by covalent capture of CpG sites. *Nat. Commun.* **4**, 2190 (2013).
- Wang, T. & Kohli, R. M. Discovery of an unnatural DNA modification derived from a natural secondary metabolite. *Cell. Chem. Biol.* **28**, 97–104.e4 (2021).
- Nabel, C. S. et al. AID/APOBEC deaminases disfavor modified cytosines implicated in DNA demethylation. *Nat. Chem. Biol.* **8**, 751–758 (2012).
- Schutsky, E. K., Nabel, C. S., Davis, A. K. F., DeNizio, J. E. & Kohli, R. M. APOBEC3A efficiently deaminates methylated, but not TET-oxidized, cytosine bases in DNA. *Nucleic Acids Res.* **45**, 7655–7665 (2017).



23. Seiler, C. L. et al. Maintenance DNA methyltransferase activity in the presence of oxidized forms of 5-methylcytosine: structural basis for ten eleven translocation-mediated DNA demethylation. *Biochemistry* **57**, 6061–6069 (2018).
24. Shi, K. et al. Structural basis for targeted DNA cytosine deamination and mutagenesis by APOBEC3A and APOBEC3B. *Nat. Struct. Mol. Biol.* **24**, 131 (2017).
25. Ghanty, U., DeNizio, J. E., Liu, M. Y. & Kohli, R. M. Exploiting substrate promiscuity to develop activity-based probes for TET family enzymes. *J. Am. Chem. Soc.* **140**, 17329–17332 (2018).
26. Chinchilla, R. & Najera, C. The Sonogashira reaction: a booming methodology in synthetic organic chemistry. *Chem. Rev.* **107**, 874–922 (2007).
27. Kelly, T. K. et al. Genome-wide mapping of nucleosome positioning and DNA methylation within individual DNA molecules. *Genome Res.* **22**, 2497–2506 (2012).
28. Liu, Y. et al. Accurate targeted long-read DNA methylation and hydroxymethylation sequencing with TAPS. *Genome Biol.* **21**, 54–56 (2020).
29. Sipa, K. et al. Effect of base modifications on structure, thermodynamic stability, and gene silencing activity of short interfering RNA. *RNA* **13**, 1301–1316 (2007).
30. Dalluge, J. J., Hashizume, T., Sopchik, A. E., McCloskey, J. A. & Davis, D. R. Conformational flexibility in RNA: the role of dihydrouridine. *Nucleic Acids Res.* **24**, 1073–1079 (1996).
31. Onodera, A. et al. Roles of TET and TDG in DNA demethylation in proliferating and non-proliferating immune cells. *Genome Biol.* **22**, 186 (2021).
32. Suvà, M. L. et al. Reconstructing and reprogramming the tumor-propagating potential of glioblastoma stem-like cells. *Cell* **157**, 580–594 (2014).
33. Klughammer, J. et al. The DNA methylation landscape of glioblastoma disease progression shows extensive heterogeneity in time and space. *Nat. Med.* **24**, 1611–1624 (2018).
34. Raiber, E. A. et al. Base resolution maps reveal the importance of 5-hydroxymethylcytosine in a human glioblastoma. *npj Genom. Med.* **2**, 6 (2017).
35. Xie, Q. et al. N<sup>6</sup>-methyladenine DNA modification in glioblastoma. *Cell* **175**, 1228–1243.e20 (2018).
36. Zhang, J. & Zheng, Y. G. SAM/SAH analogs as versatile tools for SAM-dependent methyltransferases. *ACS Chem. Biol.* **11**, 583–597 (2016).
37. Kim, J. et al. Structure-guided discovery of the metabolite carboxy-SAM that modulates tRNA function. *Nature* **498**, 123–126 (2013).
38. Xiong, J. et al. Bisulfite-free and single-base resolution detection of epigenetic DNA modification of 5-methylcytosine by methyltransferase-directed labeling with APOBEC3A deamination sequencing. *Anal. Chem.* **94**, 15489–15498 (2022).
39. Siejka-Zielińska, P. et al. Cell-free DNA TAPS provides multimodal information for early cancer detection. *Sci. Adv.* **7**, eabh0534 (2021).
40. Millar, D., Christova, Y. & Holliger, P. A polymerase engineered for bisulfite sequencing. *Nucleic Acids Res.* **43**, e155 (2015).

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.

© The Author(s), under exclusive licence to Springer Nature America, Inc. 2023

## Methods

### Protein expression and purification

The *E. coli* strain ER1821 (NEB) was used for all cloning and expression to overcome methylation-associated toxicity. The CxMTase M.MpeI N374K was expressed with an N-terminal fusion of maltose binding protein (MBP). Cloning of pMG81-MBP-M.MpeI-N374K-His was performed by Golden Gate Assembly<sup>41</sup>. Purifications of MBP-A3A-His, M.MpeI-WT-His and M.MpeI-N374K-His were performed as previously described<sup>20,42</sup>. MBP-M.MpeI-N374K-His and M.MpeI Q136A/N374A-His (eM.MpeI) were purified using the same single Cobalt column and high-salt wash strategy as previously published<sup>20</sup>.

### Synthesis of SAM and cytosine analogs

The synthesis and characterization of both bSAM and CxSAM have been previously described<sup>20,43</sup>. The synthesis and characterization of 5eyC and 5vC triphosphates have been previously reported<sup>25</sup>. The triphosphates of cytosine (Promega), 5mC (NEB) and 5pyC (TriLink) were purchased.

### Oligonucleotide assay for A3A deamination

A fluorescein (FAM)-labeled 27-bp top-strand oligonucleotide with a single unmethylated TCGA and unlabeled complementary bottom-strand oligonucleotide with a single methylated TCGA were used (Supplementary Table 1). First, 1  $\mu$ M of the duplexed oligonucleotide was reacted in a final volume of 10  $\mu$ l with no enzyme and no SAM, 1  $\mu$ M M.MpeI WT with 40  $\mu$ M SAM, 1  $\mu$ M M.MpeI Q136A/N374A and 40  $\mu$ M bSAM, or 1  $\mu$ M M.MpeI N374K and 40  $\mu$ M CxSAM at 37 °C for 1 h, before heat inactivation at 95 °C for 5 min. The sample was treated with 1  $\mu$ l of Proteinase K (NEB) and incubated at 37 °C for 15 min before purification with an oligonucleotide spin column (Zymo) and elution in 10  $\mu$ l of 0.1 $\times$  low EDTA TE. ESI-MS was obtained (Novatia) to validate the efficiency of 5bC and 5xmC generation. Next, 1  $\mu$ l of the resulting DNA was snap cooled and then incubated with 6  $\mu$ M MBP-A3A-His under ramping conditions for 2 h in a final volume of 50  $\mu$ l, as previously described<sup>42</sup>. DNA was purified using an oligonucleotide spin column (Zymo) and DNA input was normalized. DNA was annealed to 10  $\mu$ M (excess) unmethylated and mismatched opposite-strand in 1 $\times$  CutSmart Buffer (NEB) in a total volume of 9.5  $\mu$ l. The mismatched opposite-strand was a safeguard ensuring that a deaminated TUGA top-strand would not be cut. DNA was then digested with Taq<sup>q</sup>I (NEB) following recommended conditions in a total of 10  $\mu$ l for 1 h at 65 °C. The samples were diluted twofold in 95% formamide and subjected to 20% denaturing PAGE in Tris-Borate-EDTA (TBE) buffer, followed by imaging of the FAM signal using a Typhoon imager.

### Lambda gDNA

Dam<sup>-</sup>/Dcm<sup>-</sup> lambda phage gDNA was obtained (Thermo Fisher). The CpG- and GpC-modified lambda gDNAs were modified with two rounds of DNA methylation as previously described for CpG-modified DNA<sup>42</sup>. The CpG-methylated DNA was modified using M.SssI (NEB) while GpC-methylated DNA was modified using M.CviPI (NEB) in recommended buffers.

### General gDNA processing

Purified gDNA was quantified by Qubit (Thermo Fisher). All gDNA was sheared to ~350 bp using a Covaris sonicator before solid-phase reversible immobilization (SPRI) purification (Beckman, 1.2 $\times$ ). DNA was then end-repaired with NEBNext Ultra End Prep Kit. DNA libraries were separately ligated with IDT xGen Y-shaped adapters containing either all 5mC (used in BS-Seq) or all 5pyC (used in A3A workflows including DM-Seq, custom synthesis from IDT) modifications using an NEBNext Ultra II Prep Kit, purified by SPRI beads (Beckman, 1.2 $\times$ ) and then re-quantified by Qubit. DNA was stored in the -20 °C freezer after this step. All libraries, regardless of deamination method, were quantified (Qubit) and characterized by BioAnalyzer (High Sensitivity Kit,

Agilent) before sequencing. Sequencing was performed using a MiSeq Reagent Kit v2 Nano (Illumina) except for human GBM tumor libraries.

### Initial DM-Seq assay without copy-strand synthesis

Dam<sup>-</sup>/Dcm<sup>-</sup> lambda phage gDNA was obtained (Thermo Fisher) and quantified by Qubit (Thermo Fisher). All gDNA was sheared to ~350 bp using a Covaris sonicator before SPRI purification (Beckman, 1.2 $\times$ ). DNA was then end-repaired with NEBNext Ultra End Prep Kit and ligated to 5mC-containing xGen Y-shaped adapters (IDT). Then, 10 ng of 5mC adapter-ligated gDNA was reacted with untagged 0.5  $\mu$ M M.MpeI N374K and 160  $\mu$ M CxSAM in M.MpeI reaction buffer (50 mM NaCl, 10 mM Tris-HCl pH 7.9, 10 mM EDTA) and incubated for 1 h at 37 °C followed by denaturation for 5 min at 95 °C. Next, 1  $\mu$ l of Proteinase K was added (NEB) and incubated at 37 °C for 15 min. The samples were SPRI purified (1.2 $\times$ ) and subjected to standard BS-Seq (Diagenode; see below). The library was amplified using indexing primers (IDT) and HiFi HotStart Uracil+ Ready Mix (KAPA) before purification over SPRI beads (0.8 $\times$ ) to yield final libraries.

### PCR-based analysis for identifying cytosine analogs resistant to deamination

A template DNA (Supplementary Table 1) was synthesized (IDT). Modified dCTPs were used along with dATP, dTTP and dGTP in a PCR reaction using the C-depleted primer OTF12 and G-depleted primer OTR12 (Supplementary Table 1) and Taq Polymerase (NEB) in a final volume of 50  $\mu$ l (Supplementary Table 2 for all PCR methods). These 254-bp PCR products were purified over an oligonucleotide spin column (Qiagen) and quantified by Qubit (Thermo Fisher). In a volume of 6  $\mu$ l, 2 ng of DNA was snap cooled and then incubated with 6  $\mu$ M MBP-A3A-His under ramping conditions for 2 h in a final volume of 50  $\mu$ l, as previously described<sup>42</sup>. Reaction products were purified using oligonucleotide spin columns (Zymo) and eluted into 10  $\mu$ l. Then, 1  $\mu$ l of reaction product was PCR-amplified using Taq Polymerase and using OTF2 and OTR2 primers containing Illumina TruSeq partial adapters in a total volume of 50  $\mu$ l. For initial assessment of deamination, 5  $\mu$ l of the crude PCR product was digested with Taq<sup>q</sup>I (NEB) following recommended conditions and visualized on a 2% TBE agarose gel pre-stained with SYBR Safe. For next-generation sequencing analysis, the amplicons were indexed using another round of PCR in a total volume of 10  $\mu$ l (IDT TruSeq primers, KAPA HiFi polymerase) before SPRI purification (0.8 $\times$ ) to yield final libraries.

### Oligonucleotide assay for opposite-strand impacts on CxMTase

A fluorescein (FAM)-labeled 27-bp top-strand oligonucleotide with a single unmethylated CCGG and unlabeled complementary bottom-strand oligonucleotides with either an unmethylated or methylated CCGG were used (Supplementary Table 1), as previously described<sup>20</sup>. First, 200 nM of the duplexed oligonucleotide was reacted in a final volume of 5  $\mu$ l with 1  $\mu$ M M.MpeI N374K and either 40  $\mu$ M SAM or CxSAM substrate at 37 °C for 30 min, before heat inactivation at 95 °C for 5 min. A 25 $\times$  excess of unmethylated bottom-strand was added, and the duplex was reannealed before restriction digestion with HpaII (NEB) in a final volume of 50  $\mu$ l. The samples were analyzed by PAGE as described for the oligonucleotide assay above.

### Synthesis of modified adapters and evaluation of ligation efficiency

SpyC adapters were synthesized by standard phosphoramidite chemistry (IDT). The purified oligonucleotides were further characterized by mass spectrometry (Extended Data Fig. 5a). For comparison with standard 5mC-containing adapters (IDT xGen Y-shaped adapters), a 254-bp DNA product with unmodified cytosine (described above; Supplementary Table 1) was ligated to adapters according to the manufacturer instructions (NEB Ultra II). DNA was then purified by SPRI beads

(Beckman) before amplification was attempted with either internal primers OTF2/R2 and Taq Polymerase (NEB) as above or Illumina TruSeq primers (IDT) and HiFi HotStart Polymerase (KAPA). Samples were visualized on a 2% Tris-acetate-EDTA agarose gel.

### Initial unoptimized copy-strand workflow

First, 10 ng of gDNA ligated to 5mC- or 5pyC-containing adapters was used as input. A methylated copy-strand was created. Then, 1  $\mu$ M of copy primer was annealed (Supplementary Table 1, v1) in a total volume of 10  $\mu$ l in CutSmart Buffer and 1 mM final concentration individually of dATP/dGTP/dTTP (Promega) and dmCTP (NEB). Five units of Klenow (exo-) polymerase (NEB) was then added and incubated for 30 min at 37 °C. After purification (Zymo Oligo Clean & Concentrator), libraries were mixed with 1  $\mu$ M untagged M.MpeI N374K and 160  $\mu$ M CxSAM in carboxymethylation buffer (50 mM NaCl, 10 mM Tris-HCl pH 7.9, 10 mM EDTA) and incubated for 1 h at 37 °C followed by denaturation for 5 min at 95 °C. Then, 1  $\mu$ l of Proteinase K was added (NEB) and samples were incubated at 37 °C for 15 min. The samples were purified using SPRI beads (1.2 $\times$ ) and eluted in 1 mM Tris-Cl, pH 8.0. DNA was then subjected to BS-Seq (Diagenode; see below) or to snap-cooling and A3A deamination in a final volume of 50  $\mu$ l as previously described<sup>42</sup>. Purified DNA was then amplified using indexing primers (IDT) and HiFi HotStart Uracil+ Ready Mix (KAPA) before purification over SPRI beads (0.8 $\times$ ) to yield final libraries. The nonoptimized workflow was used in Fig. 2.

### Tumor DNA

Patient GBM tissue was collected at the Hospital of the University of Pennsylvania after informed patient consent under a protocol approved by the University of Pennsylvania's Institutional Review Board. The patient had a surgically resected left temporal tumor. Fresh surgically resected GBM tissue was placed in sterile PBS and taken immediately to the University of Pennsylvania Department of Pathology to confirm a preliminary diagnosis of grade IV glioma by the attending neuropathologist (M.N.). Tumor gDNA was extracted using the Agencourt FormAPure Kit with some protocol alterations as follows. Tissue was lysed at 70 °C for 1 h, proteinase K digestion was performed at 55 °C for 1 h, and tubes were briefly spun and incubated at 80 °C for an additional hour. Tissue was lysed in tissue lysis buffer (Qiagen) at 56 °C overnight, and nucleic acid was extracted by using the FormAPure protocol beginning with the bind 1 step.

### Spike-in controls

The spike-in control contains a 1:1:1 (m:m:m) mixture of unmethylated pUC19 plasmid DNA (NEB), CpG-methylated lambda phage gDNA (see above) and T4-hmC phage gDNA extracted as previously described<sup>15,42</sup>. Unsheared tumor gDNA was pooled 1:100 (m:m) with the spike-in mixture before gDNA processing (see above). To calculate true positive detection rates, it was assumed that pUC19 and T4-hmC were 100% pure of CpGs and 5hmCpGs, respectively, so DM-Seq detection was determined by the percentage of bases sequencing as cytosine. For 5mCpGs, incomplete modification by M.SssI was taken into account and true positives = % T by DM-Seq/% C by BS-Seq.

### Optimized DM-Seq workflow

See the extended Supplementary Note for detailed discussion of specific steps in the optimized and final DM-Seq workflow. Briefly, 10 ng of gDNA ligated to 5pyC-containing adapters was used as input for DM-Seq. A methylated copy-strand was created. Then, 1  $\mu$ M fully methylated copy primer was annealed (Supplementary Table 1, v2) in a total volume of 9  $\mu$ l in CutSmart Buffer and 1 mM final concentration (individually) of dATP/dGTP/dTTP (Promega) and dmCTP (NEB). Next, 1  $\mu$ l or 8 units of Bst polymerase, large fragment (NEB), was added and incubated for 30 min at 65 °C. The 5hmCs were then glucosylated with 40  $\mu$ M UDP-glucose and 1  $\mu$ l or 10 units of T4 Phage  $\beta$ GT (NEB) for 1 h at 37 °C in a final volume of 20  $\mu$ l. Incompletely copied or uncopied

fragments were degraded with 1  $\mu$ l or 10 units of Mung Bean Nuclease (NEB) for 30 min at 30 °C. After SPRI bead purification (1.2 $\times$ ), libraries were mixed with 0.5  $\mu$ M MBP-M.MpeI N374K and 160  $\mu$ M CxSAM in carboxymethylation buffer (50 mM NaCl, 10 mM Tris-HCl pH 7.9, 10 mM EDTA) and incubated overnight at 37 °C followed by denaturation for 5 min at 95 °C. Then, 1  $\mu$ l or 0.8 units of Proteinase K (NEB) was subsequently added and incubated at 37 °C for 15 min. The samples were purified using SPRI beads (1.2 $\times$ ) and eluted in 1 mM Tris-Cl, pH 8.0. DNA was then subjected to snap-cooling and A3A deamination in a final volume of 50  $\mu$ l as previously described<sup>42</sup> before SPRI beads purification (1.2 $\times$ ). DM-Seq libraries were amplified using indexing primers (IDT) and HiFi HotStart Uracil+ Ready Mix (KAPA Biosystems) before purification over SPRI beads (0.8 $\times$ ) to yield final libraries. This optimized workflow was used in Figs. 3 and 4. Alternative deamination conditions using commercially available APOBEC (NEB) and formamide denaturation were used without other alterations for the data generated in Extended Data Figs. 7 and 8.

### BS-Seq

BS-Seq was performed on 10 ng of gDNA ligated to 5mC-containing adapters (xGen, IDT), with no added copy or DM-Seq-specific steps, using manufacturer instructions (Diagenode). Purified BS-Seq libraries were amplified using indexing primers (IDT) and HiFi HotStart Uracil+ Ready Mix (KAPA Biosystems) before purification over SPRI beads (0.8 $\times$ ) to yield final libraries.

### TAPS and TAPS- $\beta$

TAPS and TAPS- $\beta$  were performed as previously described except for the source of TET enzyme (NEB, EM-Seq Conversion Module). First, 1 ng of sheared DNA input (consisting of fully unmodified C pUC19 DNA, 5mCpG-modified lambda phage gDNA and fully 5hmC-modified T4 phage gDNA) was ligated to C-containing Y-shaped adaptors using the same protocol as described above (IDT). For TAPS- $\beta$ , the ligated DNA was added to a 20- $\mu$ l reaction containing 1 $\times$  NEB CutSmart buffer (50 mM potassium acetate, 20 mM Tris-acetate, 10 mM magnesium acetate and 100  $\mu$ g ml<sup>-1</sup> BSA), 0.04 nM UDP-glucose and 10 U of T4- $\beta$ GT. The glucosylated DNA was then purified with SPRI beads (1.2 $\times$ ). This glucosylation step was omitted for standard TAPS. The purified DNA was then incubated in a 50- $\mu$ l reaction containing 1 $\times$  NEB EM-Seq TET buffer (50 mM Tris pH 8.0, 1 mM DTT, 5 mM sodium-L-ascorbate, 20 mM  $\alpha$ KG, 2 mM ATP and 50 mM ammonium iron (II) sulfate hexahydrate) and 16  $\mu$ g of TET2 (NEB). The reaction was then incubated at 37 °C for 80 min. Following oxidation, 0.8 U of Proteinase K (NEB) was added, and the mixture was incubated for 30 min at 37 °C. The oxidized DNA was then purified with SPRI beads (1.2 $\times$ ) and subjected to a second round of TET oxidation, proteinase K treatment and SPRI purification. The oxidized DNA was added to a 50- $\mu$ l reaction containing 600 mM sodium acetate (pH 4.3) and 1 M pyridine borane (Alfa Aesar). The reaction was incubated at 37 °C and 850 r.p.m. in a ThermoMixer (Eppendorf) placed in a chemical fume hood for 16 h and purified by Zymo-IC column (Zymo Research) with Oligo Binding Buffer (Zymo Research). The libraries were amplified using indexing primers (NEB) and HiFi Hotstart Uracil+ Ready Mix (KAPA Biosystems) before purification with SPRI beads (1.2 $\times$ ) to yield final libraries.

### Comparison of sequencing technologies

For each condition, 1 ng of sheared DNA input (consisting of fully unmodified C pUC19 DNA, 5mCpG-modified lambda control gDNA and fully 5hmC-modified T4 phage gDNA) was used. The 5mCpG-modified lambda control gDNA was one of four possibilities: fully unmethylated, fully CpG-methylated, fully GpC-methylated or ~50% CpG-methylated. The ~50% CpG-methylated DNA was made by mixing equal amounts of unmethylated and fully CpG-methylated lambda DNA. DNA was end-prepped and ligated with the same protocol as above except for the adapters used, which were different for each method (no deamination,



TAPS, TAPS- $\beta$ : C adapters; BS-Seq: 5mC adapters; DM-Seq: 5pyC adapters). Each method was then performed according to the protocols described above before sequencing. Bioinformatics were performed as below using a standard Bismark-based alignment for all pipelines. Bioinformatic statistics including number of reads and conversion efficiencies are provided in the Supplementary Data 1 file.

### Comparison of deamination methods

For each condition, 1 ng of sheared DNA input (consisting of fully unmodified CpUC19 DNA, fully 5mCpG-modified lambda control gDNA and fully 5hmC-modified T4 phage gDNA) was used. DNA was added to a 20- $\mu$ l reaction containing 1 $\times$  NEB CutSmart buffer (50 mM potassium acetate, 20 mM Tris-acetate, 10 mM magnesium acetate and 100  $\mu$ g ml<sup>-1</sup> BSA), 0.04 nM UDP-glucose and 10 U T4- $\beta$ GT. The glucosylated DNA was then purified with SPRI beads (1.2 $\times$ ). The purified DNA was then placed in a 50- $\mu$ l reaction containing 1 $\times$  NEB EM-Seq TET buffer (50 mM Tris pH 8.0, 1 mM DTT, 5 mM sodium-L-ascorbate, 20 mM  $\alpha$ KG, 2 mM ATP and 50 mM ammonium iron (II) sulfate hexahydrate) and 16  $\mu$ g of TET2 (NEB). The reaction was then incubated at 37 °C for 80 min. Following oxidation, 0.8 U of Proteinase K (NEB) was added, and the mixture was incubated for 30 min at 37 °C. The oxidized DNA was then purified with SPRI beads. The DNA was then subjected to a second round of TET oxidation, proteinase K treatment and SPRI purification. The no TET control was carried through mock oxidation reactions without TET enzyme. The purified oxidized sample was then end-prepped and ligated with the same protocol as above except for the adapters used, which were different for each method (no deamination, TAPS, TAPS- $\beta$ : C adapters; BS-Seq: 5mC adapters; DM-Seq: 5pyC adapters). Bisulfite (BS) deamination, A3A deamination, pyridine borane deamination or a no deamination control were then performed as above. Bioinformatics were performed as below using a standard Bismark alignment for all pipelines. Bioinformatic statistics including number of reads and conversion efficiencies are provided in the Supplementary Data 1 file. The deamination methods were also compared by qPCR and BioAnalyzer using this same workflow, by using unamplified libraries as input to determine C<sub>q</sub> values (KAPA SYBR FAST ROX, Applied Biosystems).

### General bioinformatics

Reads were quality and length trimmed with Trim Galore! Reads were aligned with Bismark and deduplicated with Picard<sup>44</sup>. All data were analyzed single-end. Reads were filtered if three consecutive CpGs were nonconverted using Bismark's existing filter\_non\_conversion command. Locus-specific amplicons (cytosine analog experiment, see above) were not deduplicated or filtered. Filtering served two purposes (in different experiments). For BS-Seq with copy-strand synthesis, the consecutive CpH conversion eliminated reads from copy-strand amplification which contained all mCpGs, unlike the lambda gDNA template. BS-Seq without copy-strand synthesis was not filtered. For DM-Seq, the copy-strand does not amplify because the copy primer 5mCs are deaminated to Ts by A3A. DM-Seq filtering additionally eliminates double-stranded DNA (dsDNA) hairpins which can cause A3A nondeamination, similar to previously described enzymatic deamination protocols<sup>15,42</sup>. Only reads with mapping quality  $\geq$  30 were analyzed. GraphPad Prism 7 was used to report percentage CpG modification unless otherwise specified. See Supplementary Table 3 for library statistics. A custom R script was employed to visualize target- and opposite-strand modifications as well as GpC-specific analysis. Pearson correlations at the individual CpG level for BS-Seq versus DM-Seq were obtained using the *cor* function in R.

### TAPS mammalian bioinformatics

Data were obtained as mm9 bedfiles from GSE112520 where TAPS and BS-Seq were performed on mouse ESCs. For correlation analyses, the raw TAPS and BS-seq signals were calculated for 1-kb nonoverlapping genomic bins. Only bins containing at least 20 CpGs were analyzed.

Pearson correlation was obtained using the *cor* function in R. ICR locations were previously provided described as mm10 but required liftover (<https://liftover.broadinstitute.org/>) to convert to mm9. Heatmaps were produced using deepTools (v.3.5.1). In the heatmaps, genes were scaled to 3 kb and binned at 300 bases.

### GBM bioinformatics

BS-Seq and DM-Seq GBM libraries were sequenced 75-bp single-end on an Illumina NextSeq using a High Output kit v2.5. Reads were processed as above. For correlation analyses, the raw DM-Seq and BS-seq signals were calculated for 1-kb nonoverlapping genomic bins. Only bins containing at least 20 CpGs were analyzed. Venn diagrams were generated using the VennDiagram package in R. Pearson correlation was obtained using the *cor* function in R, using the same bins as described for generating the Venn diagram. Heatmaps were produced using deepTools (v.3.5.1). In the heatmaps, genes were scaled to 10 kb and binned at 500 bases. Only genes with all bins visualized between DM-Seq and BS-Seq are shown. Genomic elements were defined as follows: Exons and introns: UCSC RefGene for genome build hg19. Promoters:  $\pm$ 1 kb from the transcription start site (TSS). Active promoters: H3K4me3 ChIP-Seq performed on GBM stem cells<sup>35</sup>. Enhancers: nonpromoter regions with a H3K27Ac ChIP-Seq signal<sup>32</sup>. Signals in genomic elements were determined using bedtools with each genomic element as defined as an individual bin (v.2.25.0). The 3,876 'high 5hmC' CpG sites were previously defined as identified by oxBS-Seq and used to calculate percentage modification by both BS-Seq and DM-Seq<sup>9</sup>. Downsampling was performed to obtain 1,485 CpGs from the 16,438,445 total CpGs covered by both BS-Seq and DM-Seq in R. Random downsampling was performed to randomly extract 10,000 CpGs from the obtained BS-Seq and DM-Seq data and extended results are reported in Extended Data Fig. 10 and Supplementary Table 4.

### Research use only statement

IDT products are for research use only and not intended for use in diagnostic procedures. Unless otherwise agreed to in writing, IDT does not intend these products to be used in clinical applications and does not warrant their fitness or suitability for any clinical diagnostic use. The purchaser or user is solely responsible for all decisions regarding the use of these products and any associated regulatory or legal obligations.

### Reporting summary

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

### Data availability

Sequencing data supporting the findings of this study are available in the NCBI Gene Expression Omnibus (GEO, GSE225975). The plasmid encoding MBP-t-M.Mpel-N374K-His has been made available from Addgene (197985). Relevant DNA sequences are provided in Supplementary Information. Source data are provided with this paper.

### Code availability

Software utilized for each analysis is detailed in the relevant Methods section. Scripts have been deposited on GitHub (<https://github.com/twang518/DM-Seq>).

### References

- Engler, C., Kandzia, R. & Marillonnet, S. A one pot, one step, precision cloning method with high throughput capability. *PLoS ONE* **3**, e3647 (2008).
- Wang, T. et al. Bisulfite-free sequencing of 5-hydroxymethylcytosine with APOBEC-Coupled Epigenetic Sequencing (ACE-Seq). *Methods Mol. Biol.* **2198**, 349–367 (2021).
- Arora, S., Horne, W. S. & Islam, K. Engineering methyllysine writers and readers for allele-specific regulation of protein-protein interactions. *J. Am. Chem. Soc.* **141**, 15466–15470 (2019).

44. Krueger, F. & Andrews, S. R. Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications. *Bioinformatics* **27**, 1571–1572 (2011).

## Acknowledgements

We thank Y. Lan, W. Zhou, T. Christopher and the Penn Center for Personalized Diagnostics for useful discussions and reagents. We also thank K. Islam for providing but-2-ynyl-SAM. This work was supported by the National Institutes of Health through grant no. R01-HG010646 (to R.M.K. and H.W.). E.K.S., K.N.B. and J.E.D. were NSF Graduate Research Fellows.

## Author contributions

T.W., E.K.S. and R.M.K. conceived of the approach with input from W.S.G. and H.W. T.W. conducted experiments with assistance from J.M.F., L.L., C.E.L., M.L., E.K.S., K.N.B., J.E.D., S.M. and B.Y.P. A.D. and N.D. supervised synthesis of 5pyC adapters and M.N. contributed GBM gDNA. T.W., J.M.F. and H.W. performed computational analysis and analyzed the results. T.W. and R.M.K. wrote the paper, with contributions from all authors.

## Competing interests

The University of Pennsylvania has patents pending for CxMTase enzymes, DNA deaminase-resistant adapters and the DM-Seq

pipeline. R.M.K. has served as a scientific advisory board member for Cambridge Epigenetix (CEGX). W.S.G. is an employee of CEGX and T.W. was supported by a fellowship from CEGX. A.D. and N.D. are employees of Integrated DNA Technologies, Inc. The other authors declare no competing interests.

## Additional information

**Extended data** is available for this paper at <https://doi.org/10.1038/s41589-023-01318-1>.

**Supplementary information** The online version contains supplementary material available at <https://doi.org/10.1038/s41589-023-01318-1>.

**Correspondence and requests for materials** should be addressed to Rahul M. Kohli.

**Peer review information** *Nature Chemical Biology* thanks Abdulkadir Abakir and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

**Reprints and permissions information** is available at [www.nature.com/reprints](http://www.nature.com/reprints).

a

CpG

5mCpG

5hmCpG

BS-Seq

TAB-Seq

oxBS-Seq

TAPS

TAPS-β

ACE-Seq

EM-Seq

DM-Seq

Sequence as:

	<u>protection/ modification step</u>	<u>deamination step</u>	<u>bases confounded with 5mC</u>	<u>bases directly detected by C-&gt;T</u>
T	N/A	chemical	5hmC	C
T	enzymatic	chemical	C	C, 5mC
T	chemical	chemical	none	C, 5hmC
C	enzymatic	chemical	5hmC	5mC, 5hmC
C	enzymatic	chemical	none	5mC only
T	enzymatic	enzymatic	C	C, 5mC
T	enzymatic	enzymatic	5hmC	C
C	enzymatic	enzymatic	none	5mC only

b

CpG

mCpG

hmCpG

DM-Seq

BS-Seq

TAPS

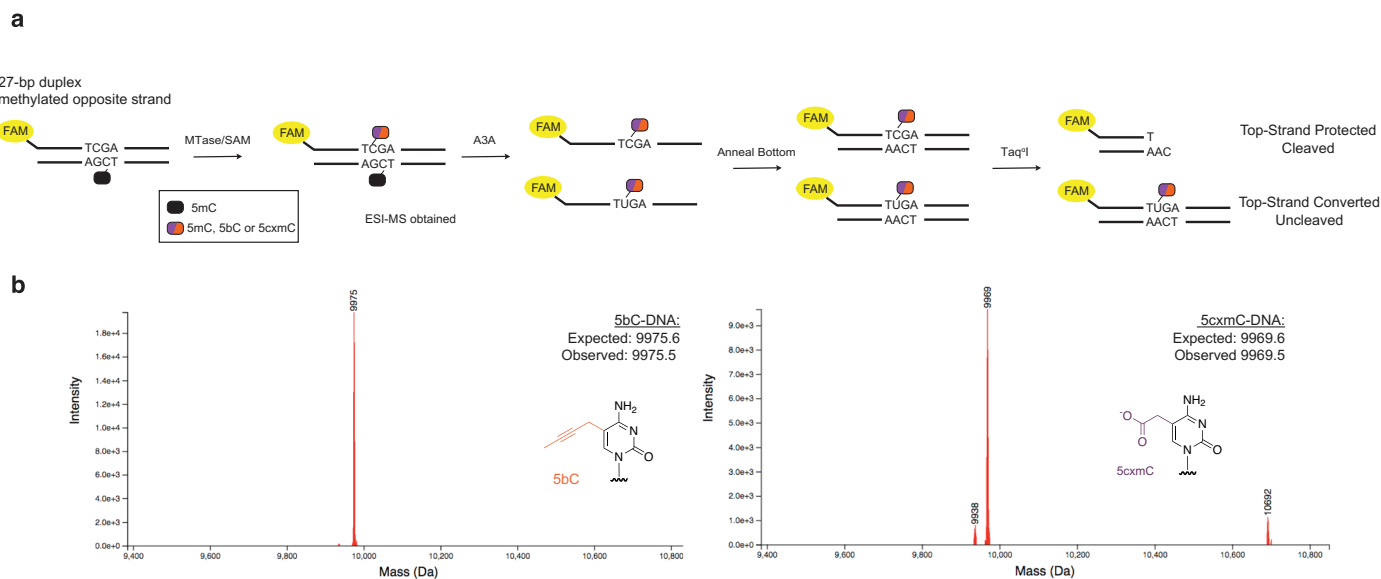
Sequence as:

	CpH	mCpH	hmCpH
C	T	T	C
T	T	C	C
C	C	T	T

**Extended Data Fig. 1 | Chemical and enzymatic sequencing methods for resolving DNA modifications. a)** Methods differ in their use of protection or modification steps to alter C, 5mC or 5hmC. They differ in deamination steps with chemical or enzymatic reagents. In each method, C, 5mC, or 5hmC are detected

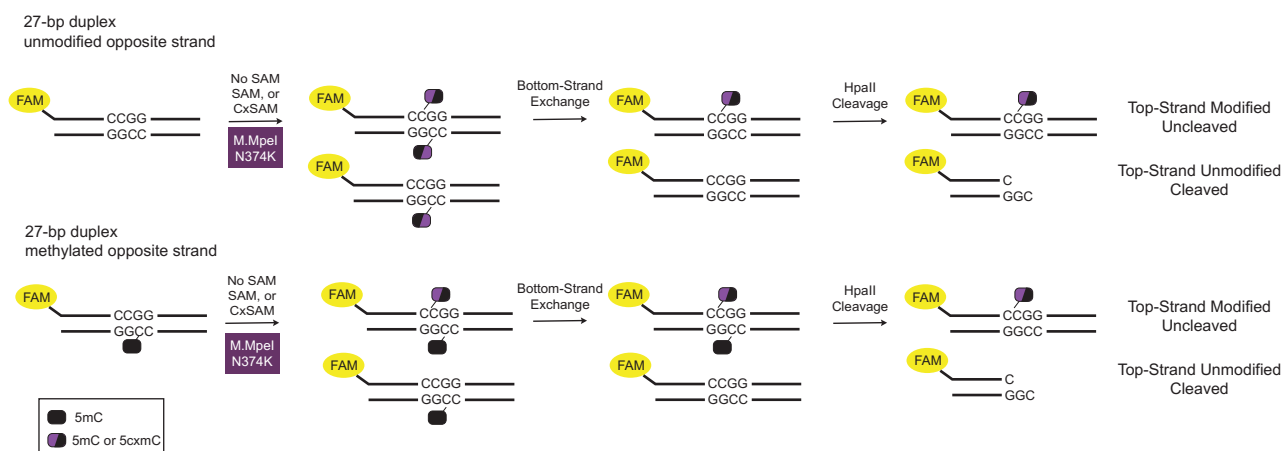
based on the pattern of C-to-T changes in sequencing, resulting in different possible bases that can be confounded with 5mC. **b)** Shown are the anticipated sequencing results for C, 5mC and 5hmC in CpG versus CpH contexts.





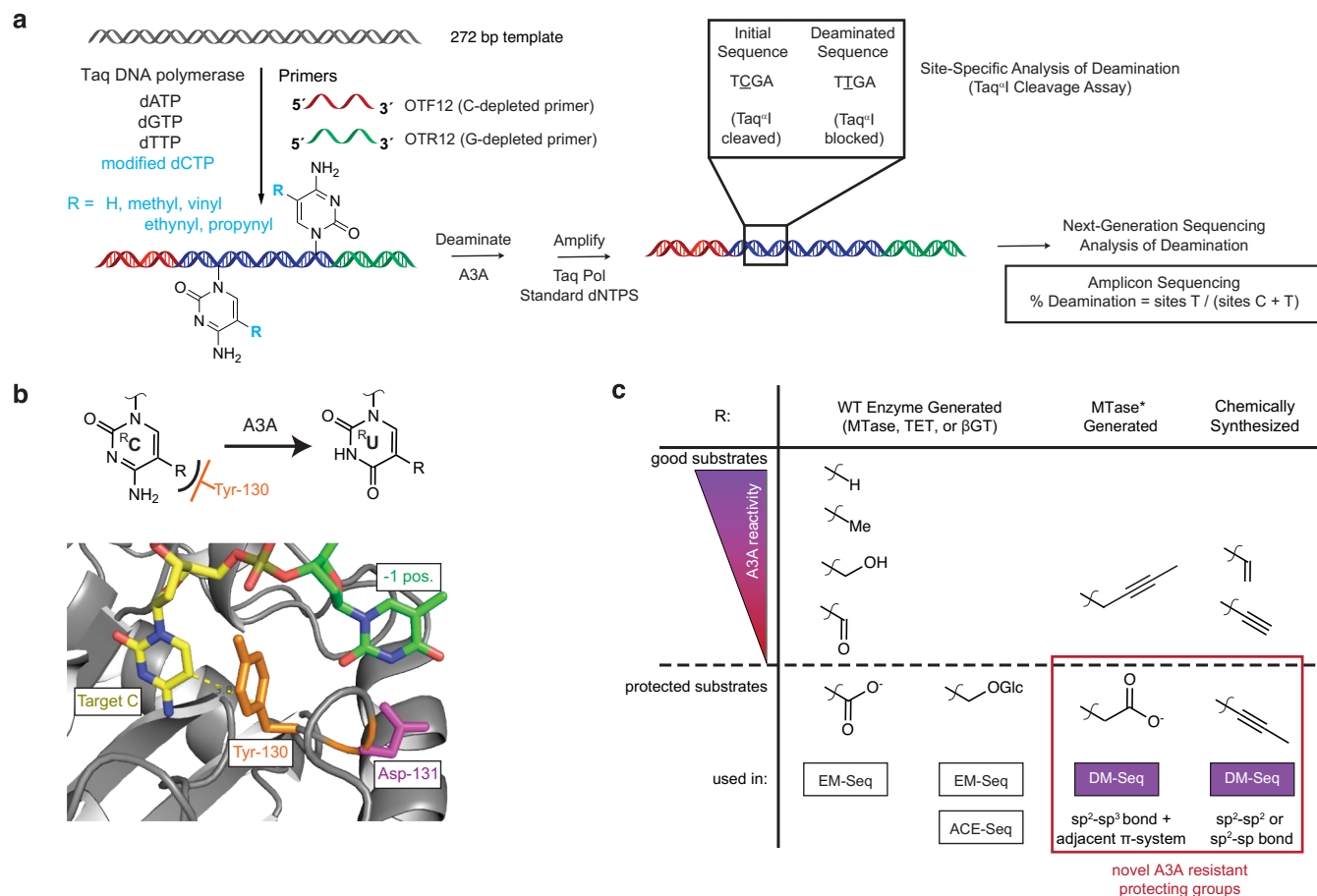
**Extended Data Fig. 2 | TaqI assay for assessment of modified cytosine deamination by A3A. a)** A fluorophore-labelled top-strand is duplexed to a complementary bottom-strand containing a methylated cytosine. The methylated cytosine is represented with a black oval. The substrate is reacted with either WT M.MpeI + SAM, eM.MpeI + bSAM, or M.MpeI N374K + CxSAM. The half-purple/half-orange oval represents a modified cytosine that can either be 5mC, 5bC, or

5cxmC after the action of the MTase variant and SAM analog. The substrate is then deaminated with A3A before duplexing a complement strand. The restriction enzyme TaqI only cleaves DNA if C is protected from A3A deamination. **b)** ESI-MS validating generation of 5bC and 5cxmC substrates before A3A reaction. No unmodified C substrate was detected.



**Extended Data Fig. 3 | HpaII assay for assessment of opposite strand effects in carboxymethylation.** A fluorophore-labelled top-strand is duplexed to a complementary strand containing either an unmodified or methylated cytosine (represented with a black oval). The duplex is incubated with M.MpeI N374K and either no SAM, SAM, or CxSAM. The half-black/half-purple oval

represents the modified cytosine on the labelled top-strand resulting after the action of the M.MpeI N374K and the SAM analog. Excess of unmodified bottom strand exchanges away the modification on the bottom strand. HpaII cleavage interrogates the modification status of the top strand.

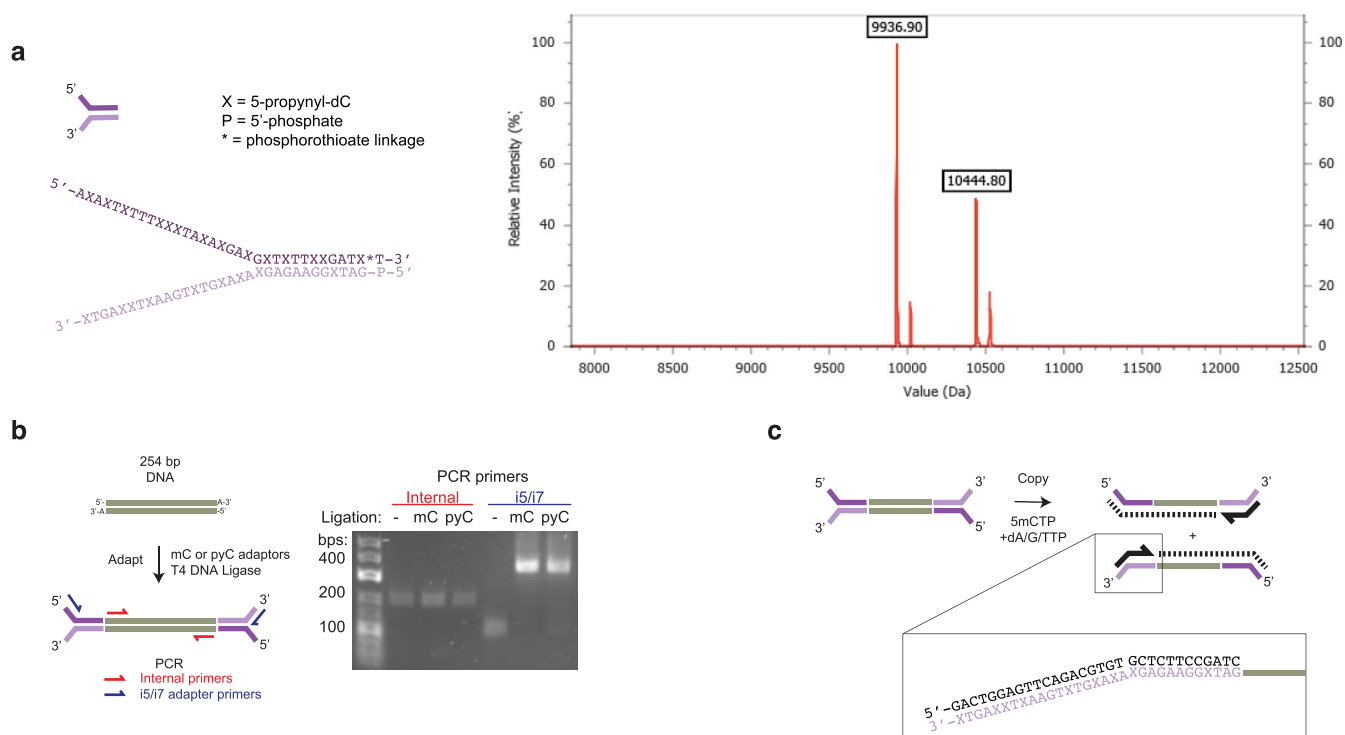


**Extended Data Fig. 4 | Structurally-informed identification of both 5cxmC and 5pyC as new protected cytosines useful for A3A dependent sequencing.**

**a**) Generation of homogenously-modified PCR substrates containing unnatural cytosines. A DNA template is amplified with a C-depleted forward primer (red) and G-depleted reverse primer (green) as well as dA/G/TTP and a modified dCTP (blue). DNA is then A3A deaminated before amplification. Amplicons are interrogated with the Taq<sup>I</sup> restriction enzyme or by Next-Generation Sequencing quantifying all C sites. **b**) Active site of human A3A (PDB: 5SWW) showing gating tyrosine (orange) which abuts the C5-C6 face of the target cytosine (yellow) and is anticipated to limit the size of the 5-position substituent

(dashed yellow line). A cartoon representation is also shown above. **c**) Summary of cytosine analogs and deamination by A3A. Left: WT MTases, TETs, and βGT make naturally-occurring modified cytosines which have different reactivities towards A3A. ScaC and 5ghmC are used in the existing methods, EM-Seq and ACE-Seq, to protect from A3A deamination. Right: 5cxmC and 5pyC are identified as novel, protected A3A substrates, both employed in DM-Seq. Despite their shared utility, 5cxmC and 5pyC also contrast in their bond types at the 5-position of cytosine, which are determined by their contrasting modes of biochemical and chemical synthesis, respectively.

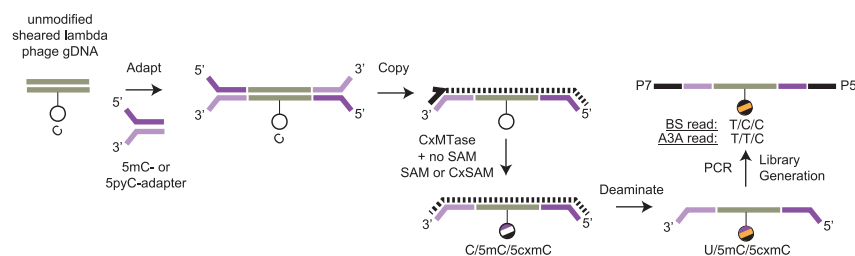




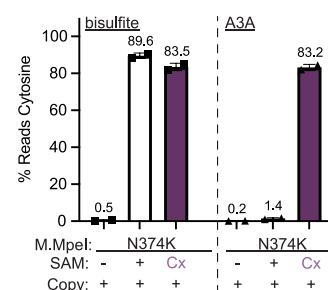
**Extended Data Fig. 5 | SpyC adapters improve DNA carboxymethylation efficiency through the synthesis of a 5mC copy strand. a)** Structure of 5pyC adapters. ESI-MS characterizing 5pyC adapters. Expected m/z of the two strands: 10,444.2 and 9,936.6. The phosphorothioate linkage (\*) substitutes a sulfur in place of a phosphate in the backbone of the oligonucleotide to minimize nuclease degradation. **b)** 5pyC adapter ligation experiment. Template DNA was

ligated to 5mC- or 5pyC-containing Y-shaped adapters. The template DNA was detected by amplification with internal primers (red) or successful ligation was detected by amplification with Illumina indexing primers (blue). Experiment was performed once. **c)** Schematic of copy strand synthesis. A copy strand is made by incubation of a copy primer, polymerase, and dA/G/TTPs with 5mdCTP.

a

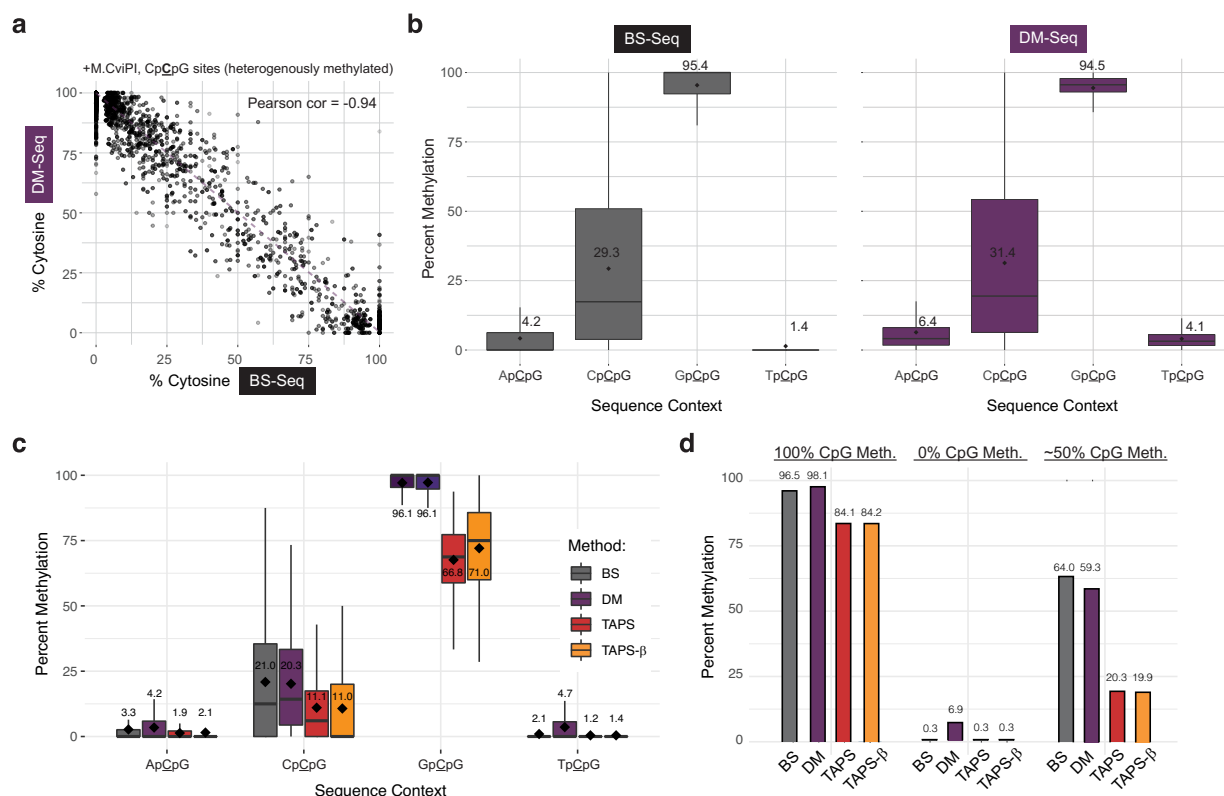


b



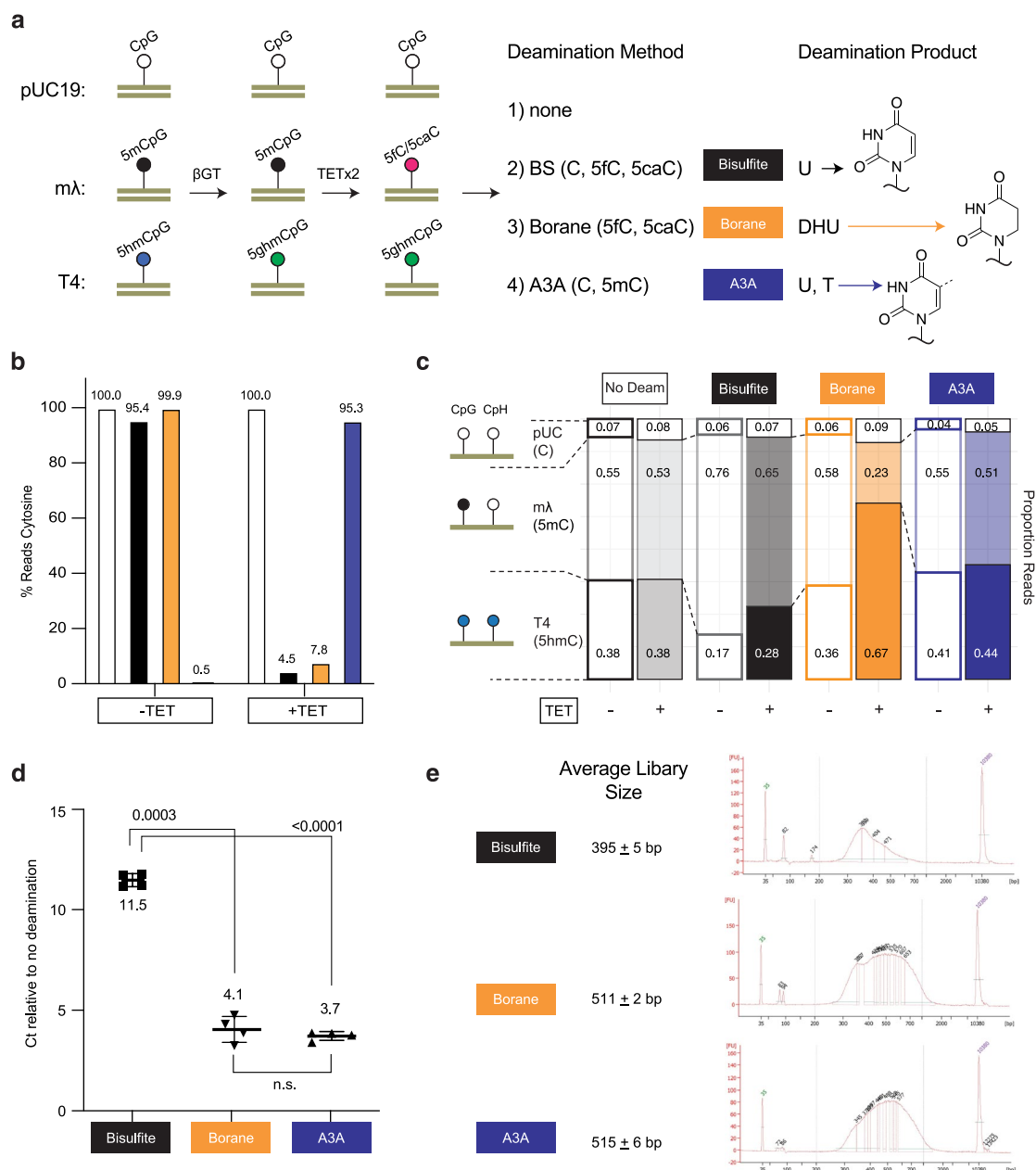
**Extended Data Fig. 6 | Copy strand synthesis improves DNA carboxymethylation efficiency. a)** Experimental scheme. Sheared lambda gDNA is ligated to 5mC- or 5pyC-containing adapters. A copy strand with 5mCs is synthesized before reaction with the CxMTase and either no SAM, SAM, or CxSAM, with the product of this reaction represented by the oval with

mixed colors. Subsequent BS or A3A deamination shows efficiency of DNA modification. Data are presented as mean values  $\pm$  SD ( $n = 2$  independent experiments). **b)** Next-Generation Sequencing quantifying efficiency of CxMTase with methylation or carboxymethylation after copy strand synthesis.



**Extended Data Fig. 7 | Comparison of different sequencing methods on M.CviPI-modified gDNA.** The M.CviPI-methylated lambda phage gDNA shows near complete modification at GpCpG sequencing contexts given the known GpC preference for M.CviPI. The enzyme has known off-target and heterogeneous activity at CpCpG sites. The dashed line shows the readout if BS-Seq signal inversely correlates with DM-Seq as anticipated. **a**) Correlation of BS-Seq to DM-Seq in an M.CviPI-modified substrate. **b**) Comparison of BS-Seq and DM-Seq modification status by 5' sequence context. The box shows the

lower quartile, median, and upper quartile. Minimum and maximum values are shown by the whiskers. Data in **a-b**) corresponds to experiment shown in Fig. 3a-b. **c**) Comparison of BS-Seq, TAPS, TAPS-β, and DM-Seq modification status by 5' sequence context. The box shows the lower quartile, median, and upper quartile. Minimum and maximum values are shown by the whiskers. **d**) Percent modification in CpG contexts of BS-Seq, TAPS, TAPS-β, and DM-Seq of 3 different methylated lambda phages. Data in **c-d**) corresponds to the experiment shown in Fig. 3c-d.

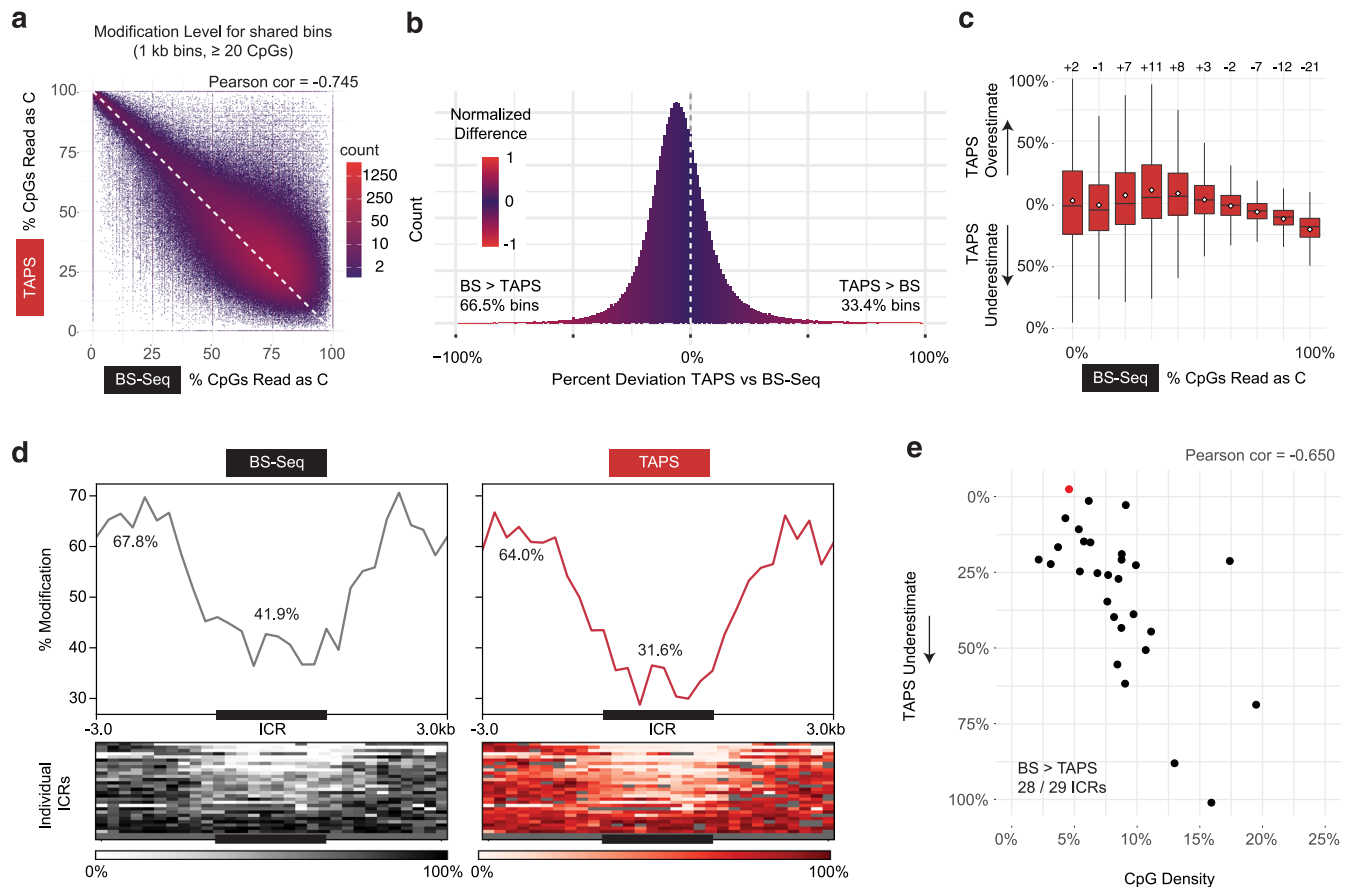


**Extended Data Fig. 8 | Comparison of deamination methods show that TAPS bias is dependent on both TET and borane-mediated deamination.**

**a)** Workflow for comparing deamination methods. A mixture of unmodified pUC19 DNA, 100% CpG methylated lambda phage, and T4-hmC phage (where all C bases are 5hmC) was glucosylated. Samples were then subjected to either two rounds of TET treatment or no TET treatment. DNA was ligated to the appropriate adapter and subjected to one of four conditions: no deamination, BS, pyridine borane, or A3A. The pyridine borane workflow is equivalent to TAPS-β. The bases deaminated by each method (detected as T by sequencing) are noted, with structures of the deamination products at right, including the non-aromatic

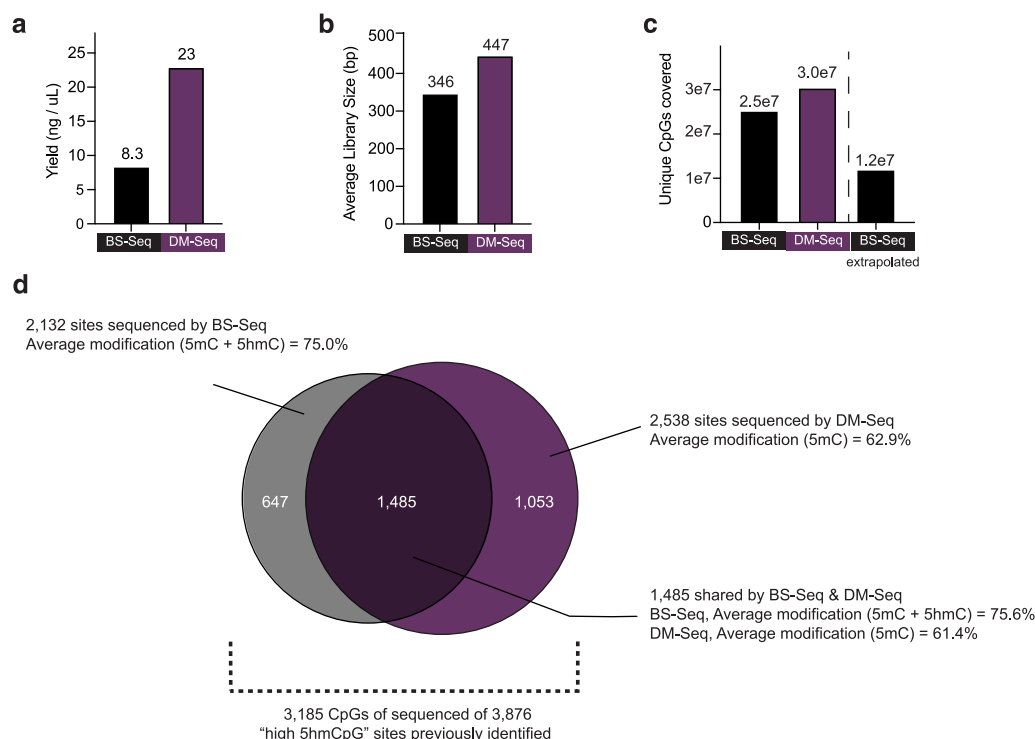
DHU. **b)** Percent reads C as determined by the methylated lambda phage spike-in. The sample with TET and bisulfite indicates efficient conversion of 5mC to 5fC/5caC by TET. **c)** Proportion of reads mapping to each spike-in are shown. Borane deamination shows decreased reads mapping to the methylated lambda phage, with depletion dependent upon TET oxidation. **d)** qPCR detection of amplifiable DNA library after each deamination method. Shown are the p-values from paired two-tailed t-test ( $n = 4$  for each deamination condition,  $3e-4$  between BS and borane,  $2e-5$  between BS and A3A). Data are presented as mean values  $\pm$  SD. **e)** Mean library size  $\pm$  standard deviation for each deamination method. A representative BioAnalyzer trace is shown for each deamination method.





**Extended Data Fig. 9 | Existing mammalian TAPS vs BS-Seq data suggest bias.** **a**) Binned CpG analysis using non-overlapping 1 kb bins with at least 20 CpGs covered. Correlation between TAPS and BS-Seq in mESCs in existing datasets ([GSE112520](#)). **b**) Histogram showing ~2-fold as many 1 kb bins with greater modification detected by BS-Seq than TAPS. Percent Deviation =  $(TAPS \% T - BS-Seq \% C) / (BS-Seq \% C)$ . **c**) Percent deviation of TAPS vs BS-Seq as a function of % modification of CpGs in a given 1 kb bin. The box shows the lower quartile, median, and upper quartile. Minimum and maximum values are

shown by the whiskers. **d**) ICRs show underdetection of 5mC by TAPS relative to BS-Seq. At bottom is the heatmap representation of individual ICRs. The percent modification outside of ICRs (64.2% vs 60.1%) represents the genome-wide average for each method using 1 kb bins vs just at the ICR (41.9% vs 31.6%). **e**) Plot of the CpG density in individual ICRs versus the percent TAPS underestimates the level of 5mC relative to BS-Seq. 28 of 29 ICRs show lower modification density by TAPS than by BS-Seq. The one exception is shown in red. The associated correlation coefficient tracks the % underestimate as a function of CpG density.



**Extended Data Fig. 10 | Mammalian genome DM-Seq metrics.** DM-seq and BS-seq data from gDNA derived from a patient glioblastoma. **a)** Final library yield. **b)** Average size of library fragments (adapters included) determined using a Bioanalyzer. **c)** Unique CpGs covered by BS-Seq and DM-Seq. The extrapolated BS-Seq bar takes into account if the sequencer was loaded with the same volume of each library rather than by normalizing the amount of DNA loaded. **d)** High

5hmCpG sites, previously identified by oxBS-Seq of 30 tumors. The Venn diagram shows the portion of these CpG sites that were covered by BS-seq or DM-Seq with this glioblastoma sample. The metrics for the sites sequenced by either BS-Seq or DM-Seq alone are similar to those at the sites that were sequenced by both methods. The analysis in Fig. 4e focuses on the 1,485 shared CpG sites sequenced by both methods.

## Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our [Editorial Policies](#) and the [Editorial Policy Checklist](#).

### Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a Confirmed

- |                                     |                                     |  |
|-------------------------------------|-------------------------------------|--|
| <input type="checkbox"/>            | <input checked="" type="checkbox"/> | The exact sample size ( $n$ ) for each experimental group/condition, given as a discrete number and unit of measurement  |
| <input type="checkbox"/>            | <input checked="" type="checkbox"/> | A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly  |
| <input type="checkbox"/>            | <input checked="" type="checkbox"/> | The statistical test(s) used AND whether they are one- or two-sided<br><i>Only common tests should be described solely by name; describe more complex techniques in the Methods section.</i>   |
| <input checked="" type="checkbox"/> | <input type="checkbox"/>            | A description of all covariates tested   |
| <input checked="" type="checkbox"/> | <input type="checkbox"/>            | A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons  |
| <input type="checkbox"/>            | <input checked="" type="checkbox"/> | A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals) |
| <input type="checkbox"/>            | <input checked="" type="checkbox"/> | For null hypothesis testing, the test statistic (e.g. $F$ , $t$ , $r$ ) with confidence intervals, effect sizes, degrees of freedom and $P$ value noted<br><i>Give <math>P</math> values as exact values whenever suitable.</i>                            |
| <input checked="" type="checkbox"/> | <input type="checkbox"/>            | For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings   |
| <input checked="" type="checkbox"/> | <input type="checkbox"/>            | For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes   |
| <input type="checkbox"/>            | <input checked="" type="checkbox"/> | Estimates of effect sizes (e.g. Cohen's $d$ , Pearson's $r$ ), indicating how they were calculated   |

Our web collection on [statistics for biologists](#) contains articles on many of the points above.

### Software and code

Policy information about [availability of computer code](#)

Data collection	Sequencing data was collected and demultiplexed by an Illumina MiSeq or NextSeq instrument as specified in the methods.
Data analysis	General data analysis was completed in GraphPad Prism 7. For sequencing data analysis, Trim Galore! (v0.6.5), Bismark (v0.22.3), Picard (v2.21.7), bedtools (v2.25.0), deepTools (v3.5.1), and R (3.6.1) were employed as described in the Methods section in detail. Custom analysis has been provided at <a href="https://github.com/twang518/DM-Seq">https://github.com/twang518/DM-Seq</a> .

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio [guidelines for submitting code & software](#) for further information.

### Data

Policy information about [availability of data](#)

All manuscripts must include a [data availability statement](#). This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our [policy](#)

Sequencing data supporting the findings of this study are available in the NCBI Gene Expression Omnibus (GEO, GSE225975). The plasmid encoding MBP-t-M.Mpel-

N374K-His has been made available from Addgene (197985). Relevant DNA sequences are provided in Supplementary Information.

## Human research participants

Policy information about [studies involving human research participants and Sex and Gender in Research](#).

Reporting on sex and gender	This data has not been collected as it is not relevant to the study.
Population characteristics	The patient had a surgically resected, pathologist diagnosed grade IV glioma resected from the L temporal lobe. Covariates were not considered as we sought to compare DM-Seq to BS-Seq in a single tumor type (human glioblastoma) where 5-hydroxymethylcytosine has previously been shown to be important.
Recruitment	The patient was previously recruited to be in a Penn glioblastoma biobank.
Ethics oversight	University of Pennsylvania's Institutional Review Board

Note that full information on the approval of the study protocol must also be provided in the manuscript.

## Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see [nature.com/documents/nr-reporting-summary-flat.pdf](https://nature.com/documents/nr-reporting-summary-flat.pdf)

## Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	No statistical methods were used to predetermine sample size. A single human tumor was sequenced to show proof of principle for DM-Seq, which was also validated using rigorous in vitro controls.
Data exclusions	No data were excluded.
Replication	Experiments were replicated as described in the figure legends.
Randomization	No randomization was used because only a single human tumor was sequenced.
Blinding	Investigators were not blinded during data collection and analysis because blinding is not relevant in this study with objective spike-in control metrics reported by next generation sequencing.

## Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

### Materials & experimental systems

n/a	Involved in the study
<input checked="" type="checkbox"/>	<input type="checkbox"/> Antibodies
<input checked="" type="checkbox"/>	<input type="checkbox"/> Eukaryotic cell lines
<input checked="" type="checkbox"/>	<input type="checkbox"/> Palaeontology and archaeology
<input checked="" type="checkbox"/>	<input type="checkbox"/> Animals and other organisms
<input checked="" type="checkbox"/>	<input type="checkbox"/> Clinical data
<input checked="" type="checkbox"/>	<input type="checkbox"/> Dual use research of concern

### Methods

n/a	Involved in the study
<input checked="" type="checkbox"/>	<input type="checkbox"/> ChIP-seq
<input checked="" type="checkbox"/>	<input type="checkbox"/> Flow cytometry
<input checked="" type="checkbox"/>	<input type="checkbox"/> MRI-based neuroimaging



# Direct enzymatic sequencing of 5-methylcytosine at single-base resolution

---

In the format provided by the  
authors and unedited

## Supplementary Tables

**Supplementary Table 1. Oligonucleotides.**

DNA	Sequence	Purpose
254mer GeneBlock	gtcactcagATGTATAGAATGATGAGTTAGGTAGTGTTGAT ATGGGTTATGAATGAAGTAGTCGATCTTTCATCATATT CTAGATCCCTCTGAAAAAATCTCCGAGTTTGCTAGGC AGTGATACATAACTCTTTTCCAATAATTGGGGAAGTCA TTCAAATCTATAATAGGTTTCAGATTTAATTCTGACTGT AGCTGCTGAAACGTTGCGGAGTGTTAAGGTATATGAG TAGATGATTGATTGGGTATGTTGATAAGTGTAgtcactcag	Generate DNA substrate with homogenously modified cytosines
OTF12	ATGTATAGAATGATGAGTTAGGTAGTTGATATGGGT TATGAATGAAGTA	Generate DNA substrate with homogenously modified cytosines
OTR12	TACACTTATCAACATACCCAATCAATCATCTACTCATAT ACCTTAACACT	Generate DNA substrate with homogenously modified cytosines
OTF2_TruSeq	ACACTCTTTCCCTACACGACGCTCTTCCGATCTTTGAT ATGGGTTATGAATGAAGTA	Primers for installing Illumina overhangs
OTR2_TruSeq	GACTGGAGTTCAGACGTGTGCTCTTCCGATCTAGTGT TAAGGTATATGAGTAGATGA	Primers for installing Illumina overhangs
Taq <sup>q</sup> I-Top-C-FAM	FAM- AGTATGTAGTTGAATTCGATTGATAGGAGG	A3A deamination assay
Taq <sup>q</sup> I-Top-T-FAM	FAM- AGTATGTAGTTGAATTTGATTGATAGGAGG	A3A deamination assay
Taq <sup>q</sup> I-Bottom-5mC	CCTCCTATCAAT/5mC/GAATTCAACTACATA CT	A3A deamination assay
Taq <sup>q</sup> I-Bottom-C	CCTCCTATCAATCAAATTCAACTACATACT	A3A deamination assay
HpaII-Top-C-FAM	GTATCTAGTTCAATCCGGTTCATAGCA-FAM	CxMTase opposite strand assay
HpaII-Bottom-5mC	TGCTATGAAC/5mC/GGATTGAACTAGATAC	CxMTase opposite strand assay
HpaII-Bottom-C	TGCTATGAACCGGATTGAACTAGATAC	CxMTase opposite strand assay
Unmethylated Copy Primer (v1)	GACTGGAGTTCAGACGTGT	copy strand synthesis
Methylated Copy Primer (v2)	GA/5mC/TGGAGTT/5mC/AGA/5mC/GTGTG/5mC/T/5mC/ TT/5mC/GAT/5mC/	copy strand synthesis

**Supplementary Table 2. PCR Methods.**

Purpose	Method
Indexing PCRs	98°C for 30s, 10x [98°C for 10s, 65°C for 30s, 68°C for 60s], 4°C hold
254mer generation PCR (OTF12/R12)	95°C for 5 min, 35x [95°C for 30s, 62°C for 30s, 72°C for 30s], 72°C for 5 min, 4°C hold
254mer post deamination PCR	95°C 5min, 25x [95°C 15s, 50°C 30s, 60°C 15s], 60°C 5min, 4°C hold.

**Supplementary Table 3. Final Library Statistics**

Method	Condition	Indexing PCR cycles	ng / $\mu$ L final library	total reads	reads after trimming	reads after alignment	reads after deduplication	reads after filtering
BS-Seq	-MTase	17	0.308	16716	16507	16050	14067	
BS-Seq	+M.SssI	17	0.222	4161	3150	2965	2826	
BS-Seq	+M.CviPI	17	0.664	15809	15636	13831	12429	
DM-Seq	-MTase	17	36.0	67400	67333	63973	44799	44377
DM-Seq	+M.SssI	17	30.6	24830	24598	23803	20572	20180
DM-Seq	+M.CviPI	17	23.8	55542	55304	53533	39379	37865
BS-Seq	GBM	14	8.30	223430253	222555471	172961545	89542258	
DM-Seq	GBM	13	23.0	223862027	223764985	165058698	107711545	105968995

**Supplementary Table 4. Downsampling Statistics.**

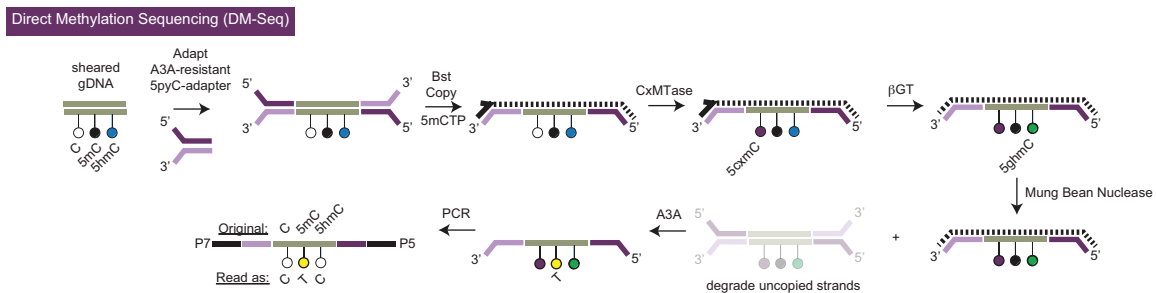
	BS-Seq	DM-Seq
unique CpGs sequenced	25,138,620	30,369,633
% modification of all CpGs sequenced	74.4	75.9
shared CpGs sequenced	16,438,445	
% modification of shared CpGs	76.0	75.4
High 5hmCpGs covered by each data set	2,132	2,538
% modification at high 5hmCpGs	75.0	62.9
High 5hmCpGs covered by both datasets	1,485	
% modification of shared high 5hmCpGs	75.6	61.4
Number of downsamplings	10,000	
Number of CpGs in each simulation	1,485	
Mean modification for all downsamplings	76.0	75.4
Standard deviation for all downsamplings	1.14	1.06

## Supplementary Note: Full DM-Seq Protocol

### Introduction

Direct Methylation Sequencing (DM-Seq) detects 5-methylcytosine (5mC) at single-base resolution at CpGs using low genomic DNA input. In this method, 5-propynylcytosine (5pyC) containing adapters, which are resistant to deamination by the enzyme APOBEC3A (A3A), are first ligated to sheared genomic DNA. These adapters also serve as a scaffold to prime the creation of a methylated copy strand which is a favorable substrate for DNA carboxymethylation. A DNA carboxymethyltransferase (CxMTase) protects all unmodified CpGs by creating 5-carboxymethylcytosines (5cxmC) and glucosylation with  $\beta$ GT protects all 5-hydroxymethylcytosines (5hmC). The A3A deaminase then only acts on the unprotected 5mC, generating T, and subsequent PCR amplification identifies 5mCs as converted Ts in sequencing. Key advantages of this workflow are that it is 1) direct for sequencing 5mC and 2) non-destructive as compared to chemical methods such as bisulfite sequencing.

A detailed protocol for DM-Seq is provided below.





## Materials

Note: while in-house purified A3A enzyme was used for the majority of experiments in this manuscript, we have validated both ACE-Seq and DM-Seq (**Supplementary Figs. 7 and 8**) using commercially-available A3A enzyme (E7125, NEB) and have written this protocol using this more readily available commercial source. We further anticipate that critical reagents such as the DNA CxMTase, CxSAM, and 5pyC-containing adapters will also be available from accessible sources in the future.

- Samples of interest
- Sonicator for shearing DNA (e.g. Covaris M220 Focused Ultrasonicator) with corresponding shearing tubes (e.g. 50  $\mu$ L microTUBEs, 520166)
- Spike-in controls
  - Unmethylated pUC19 (as a control for DNA carboxymethylation and protection of 5cxmC from deamination)
    - *Unmethylated pUC19 plasmid DNA (dam+/dcm+/cpg-), from standard plasmid preparation.*
  - Methylated M.SssI-treated lambda phage gDNA (as a control for complete deamination of 5mC)
    - *Unmethylated lambda phage DNA (D1521, Promega) was enzymatically methylated at all CpG sites using M.SssI methyltransferase (EM0821, Thermo).*
  - Hydroxymethylated T4-hmC DNA (as a control for  $\beta$ GT protection of 5hmC from deamination)
    - *Genomic DNA from a mutant T4 phage in which inactivating mutations exist in both the alpha- and beta-glucosyltransferase genes, but the machinery to synthesize dhmCTP remains intact. Therefore, every "C" in this genome is 5hmC.*
- NEBNext Ultra II End Prep Mix and Buffer (E7546S)
- NEBNext Ultra II Ligation Master Mix and Ligation Enhancer (E7595S)
- IDT xGEN Duplexed Y-shaped Adapters (15  $\mu$ M)
  - 5pyC (custom synthesis, IDT, analogous to 10005974)
- Methylated Copy Primer (v2, 10  $\mu$ M)
- Bst large fragment polymerase (8000 U/mL, NEB – M0275S)
- 10x NEB CutSmart (B7004S)
- Nucleotides
  - dATP, dGTP, dTTP (10 mM, NEB - U120B, U121B, U123C)
  - dmCTP (10 mM, NEB - N0356S)
- Mung Bean Nuclease (NEB - M0250S)
- Proteinase K (NEB - P8107S)
- CxMTase reagents
  - 5x CxMTase Buffer (50 mM Tris-HCl, 250 mM NaCl, 50 mM EDTA, 5 mM DTT, pH 7.9 at 25°C)
  - MBP-M.Mpel-N374K (2.5  $\mu$ M) – prepped in-house
  - CxSAM (diluted to 1.6 mM) – prepped in-house
- APOBEC3A deamination
  - A3A enzyme (E7125, NEB)
  - Two thermocyclers
  - PCR sample cooler rack stored at -20°C
  - 100% Formamide

- 10x A3A reaction buffer (E7125, NEB)
- NEBNext Multiplex Oligos for Illumina - Dual Indices (E7600S)
- PCR Master Mixes
  - KAPA qPCR Master Mix (Roche – KK4621)
  - KAPA 2x Hot Start Uracil+ Ready Mix (Roche – KK2802)
- qPCR 384 well plate (4309849) and covers (AB0558)
- Agilent 2100 Bioanalyzer Instrument and High Sensitivity DNA Kit (5067-4626)
- Illumina reagents (e.g. MiSeq Reagent Nano Kit V2, MS-103-1001)
- 0.2M NaOH, diluted from 1M
- Nuclease Free H<sub>2</sub>O (AM9937)
- Low EDTA TE (0.22 µm filtered, 10 mM Tris Cl, pH 8.0 at 25°C, 0.1 mM EDTA)
- SPRIselect Beads (Beckman Coulter – B23318) and magnetic rack (S1515S)
- Qubit 3.0 Fluorimeter
  - Qubit dsDNA BR reagent (Q32850, 100 pg/µL - 1000 ng/µL)
  - Qubit dsDNA HS reagent (Q32851, 10 pg/µL - 100 ng/µL)

## Procedure

### Shear DNA

1. Dilute DNA into a DNA LoBind tube with low EDTA TE to a final volume of 50  $\mu$ L.
2. Add whole-genome spike-ins at a total ratio of 1:100 spike-ins : gDNA (m:m, 1:300 each). Note that more dilute concentrations can be used if not utilizing shallow (e.g. MiSeq) sequencing as a control.
  - a. Unmethylated pUC19
  - b. CpG-Methylated M.SssI-treated lambda gDNA
  - c. Fully hydroxymethylated T4-hmC phage DNA
3. Use Covaris Sonicator as in typical sonication protocol for appropriate size (50  $\mu$ L volume)
  - a. ~550 bp: 45s duration, 500 cycles per burst, 25 watts, 10% duty factor
4. Perform a left-sided 1.2x SPRI purification and elute in 40  $\mu$ L low EDTA TE
5. Quantify samples with Qubit HS reagent.

### End-Repair DNA

1. Assemble the reactions in a total volume of 60 $\mu$ L using the following chart on a per reaction basis.

	$\mu$ L
NEBNext Ultra II End Prep Enzyme Mix:	3
NEBNext Ultra II End Prep Reaction Buffer:	6
Fragmented DNA (500 pg - 1 $\mu$ g total):	51

2. Incubate samples with the following thermocycler method: 20°C for 30 minutes, 65°C for 30 minutes, 4°C hold. Set the heated lid to 75°C
3. Fragmented DNA input should be within the range of 500 pg to 1  $\mu$ g, and diluted in low EDTA TE per NEB kit instructions, although lower input is possible.

### Ligate Illumina Y-Shaped Adapters

1. Anneal the appropriate volume of Y-shaped adapters following the thermocycler method: 95°C 5min, 70°C to 40°C in steps of 5°C per min, 37°C 5min, 4°C hold.
2. Assemble the reactions in a total volume of 93.5  $\mu$ L using the following chart on a per reaction basis.

	Vol ( $\mu$ L)
End Prep Reaction Mixture (unpurified)	60
Annealed IDT xGEN Y-shaped Adapter (15 $\mu$ M, in low EDTA TE)	2.5
NEBNext Ultra II Ligation Master Mix	30
NEBNext Ligation Enhancer	1
Total Volume	93.5

3. Incubate the samples at the thermocycler method: 20°C for 15 minutes, 4°C hold. Turn the heated lid off.
4. Perform a left-sided 1.2x SPRI purification and elute in > 6  $\mu$ L Nuclease-Free H<sub>2</sub>O. Do not vortex samples.

- Quantify samples using Qubit HS reagent.

*Note: Following adapter ligation, the samples may be stored at -20 °C until ready to proceed.*

### Anneal copy primer and extend 5mC copy strand

*Note 1: We chose the Bst large fragment polymerase since it lacks 5'-3' exonuclease activity but has excellent strand displacement activity. We also found that a methylated copy primer increases copy strand efficiency, possibly due to an increase in annealing temperature.*

*Note 2: We find that the copy strand does not readily sequence based on the CpG and 5hmCpG spike-in phage controls (Figure 4a). If the copy strand were to be amplified after enzymatic deamination, we would anticipate reads where the associated CpGs from the copy stand would read as T rather than C. The accuracy of our detection of unmodified CpG (98.9%) and 5hmCpG (99.7%) with the spike-in control phage DNA indicate that this is not the case.*

- Decide on the amount of ligated input DNA to use. ~1-10 ng was used in our study.
- Assemble the reactions in a total volume of 9 µL using the following chart on a per reaction basis.

5pyC ligated sample	Up to 9 µL in Nuclease-Free H <sub>2</sub> O
10x NEB CutSmart	1 µL
Methylated Copy Primer (v2, 10 µM)	1 µL
10 mM dATP, dGTP, dTTP	1 µL
10 mM dmCTP	1 µL

- Incubate samples at the thermocycler method: 95°C 5min, 70°C to 40°C in steps of 5°C per min, 37°C 5min, 4°C hold.
- Add 1µL Bst large fragment polymerase to each sample.
- Incubate samples at the thermocycler method: 65°C 30min, 4°C hold.

### Protect 5hmC with βGT

- Assemble the reactions in a total volume of 20 µL using the following chart on a per reaction basis.

Unpurified and copied sample	10 µL
UDP-Glucose	0.4 µL
T4-βGT	1 µL
10x NEB CutSmart buffer	2 µL
nfH <sub>2</sub> O	6.6 µL

- Incubate the samples at the thermocycler method: 37°C 1 hour, 4°C hold

### Degrade uncopied strands

- Adding 1µL of Mung Bean Nuclease to each sample and incubate samples at the thermocycler method: 30°C 30 min, 4°C hold
- Perform a left-sided 1.2x SPRI purification and elute in 6 µL low EDTA TE.



*Note: Make sure sample cools to 4°C before adding Mung Bean Nuclease. This step is an added precaution and is optional.*

### **DNA carboxymethylation**

*Note: This is the critical enzymatic reaction will convert C to 5cxmC, which is resistant to A3A deamination. 16 hrs was selected for a convenient overnight protocol. Omission of carboxymethylation and copy strand steps sequences 5hmC at base resolution, analogous to ACE-Seq.*

1. Assemble the reactions in a total volume of 10 µL using the following chart on a per reaction basis.

Purified DNA	5 µL
5x CxMTase buffer	2 µL
MBP-M.MpeI N374K-His enzyme	2 µL
CxSAM (1.6mM)	1 µL

2. Incubate the samples at the thermocycler method: 37°C 16 hours, 95°C 5 min, 4°C hold

### **Denaturation and A3A Treatment**

1. Treat the remaining post-carboxymethylation DNA with 0.5 µL Proteinase K and incubate at 37°C for 15 minutes.
2. Perform a left-sided 1.2x SPRI purification and elute in 16 µL nuclease-free H<sub>2</sub>O. Do not vortex samples.
3. Preheat one thermocycler to 85°C (with heated lid). On the other thermocycler, preheat with the “ramp method,” which follows the method:

4°C	10 min
4°C - 50°C	2:15 per degree of ramp (~2 hours total)
50°C	10 min
4°C	Hold

*Note 1: The rationale for the ramping conditions is based on balancing the kinetic tendency for reannealing with A3A deamination efficiency. We compared this to incubation at 37°C alone and this ramping method is more favorable for complete deamination. When the ramp starts at a low temperature, the strands should not favor reannealing and A3A deamination will start. As the temperature slowly increases, A3A deamination efficiency will increase, leading to more deamination events; even though the thermal fluctuations may increase and the tendency of the DNA to reanneal will be higher, the fact that some deamination will already have occurred can disrupt base pairing and disfavor reannealing. The protocol thus aims to increase single-strandedness and promote more efficient deamination.*

*Note 2: We have seen slightly more consistent performance using formamide denaturation / heat denaturation relative to DMSO / heat denaturation.*

- Assemble the reactions in a total volume of 20  $\mu\text{L}$  using the following chart on a per reaction basis by pipetting up and down multiple times.

Purified post-CxMTase DNA	16 $\mu\text{L}$
Formamide	4 $\mu\text{L}$

- Incubate the samples at 85°C for 10 minutes.
- Assemble the below master mix in a total volume of 80  $\mu\text{L}$  using the following chart on a per reaction basis.

nfH <sub>2</sub> O	68 $\mu\text{L}$
NEB APOBEC Reaction Buffer	10 $\mu\text{L}$
NEB BSA	1 $\mu\text{L}$
NEB APOBEC	1 $\mu\text{L}$

- After DNA is fully denatured for 10 minutes in formamide, swiftly transfer samples to PCR sample cooling rack pre-chilled to -20°C.
- Add the master mix containing the APOBEC enzyme (80  $\mu\text{L}$  / reaction) to each denatured sample (20  $\mu\text{L}$ ) while still on the 20°C block.
- Spin down on tabletop microfuge.
- Incubate the samples using the “ramp method.”
- Following deamination, perform a left-sided 1.2x SPRI purification and elute in 10  $\mu\text{L}$  nuclease-free H<sub>2</sub>O. Do not vortex samples.

## qPCR

*Note: The purpose of the qPCR is to determine an approximate number of indexing cycles for each sample, as over-indexing leads to bias in the final sequencing libraries.*

- Assemble the reactions in a total volume of 10  $\mu\text{L}$  using the following chart on a per reaction basis.

nfH <sub>2</sub> O	3 $\mu\text{L}$
2x KAPA qPCR Master Mix (KK4621)	5 $\mu\text{L}$
Purified, deaminated DNA	1 $\mu\text{L}$
2 $\mu\text{M}$ indexing primers	1 $\mu\text{L}$

- Carefully pipette each sample into the 384-well plate and start qPCR method (Applied Biosystems Real-Time PCR instrument): 95°C 3min, 25x (95°C 15s, 63°C 30s, 72°C 30s), 4°C hold.
  - This method will take 52 minutes to run.
- For each sample, record the number of cycles required to reach the half-maximum value. The maximum value can be found by approximating where each sample curve plateaus.
  - Use these values to assign cycles for the indexing PCR.

## Indexing PCR

- Assign unique indices to each sample.

2. Assemble the reactions in a total volume of 50  $\mu\text{L}$  using the following chart on a per reaction basis.
  - a. Note: the 50  $\mu\text{L}$  volume should be used for mammalian DNA samples.

nfH <sub>2</sub> O	5 $\mu\text{L}$
2x KAPA Uracil+ PCR Master Mix (KK2802)	25 $\mu\text{L}$
Purified, deaminated DNA	19 $\mu\text{L}$
Combined i5/i7 indexing primer	1 $\mu\text{L}$

3. Incubate samples with the following thermocycler method: 98°C 30s, variable x (98°C 10s, 65°C 30s, 68°C 1 min), 4°C hold.
4. Perform a left-sided 0.8x SPRI purification and elute in 20  $\mu\text{L}$  low EDTA TE.

### **BioAnalyzer Quality Control**

1. Quantify all samples with Qubit HS dsDNA reagent
2. Ensuring 100pg/ $\mu\text{L}$  – 10ng/ $\mu\text{L}$  concentrations before loading on BioAnalyzer.
3. Use BioAnalyzer to estimate average library size, as well as assessing need for an additional SPRI purification due to adapter dimers.

### **Illumina Sequencer**

1. Load sequencer as in manufacturer instructions utilizing library size and DNA concentration.