

BISULFITE-FREE, BASE-RESOLUTION IDENTIFICATION OF CYTOSINE MODIFICATIONS

CROSS-REFERENCE TO RELATED APPLICATIONS

[0001] This application claims the benefit of U.S. Provisional Application Nos. 62/614,798 filed January 8, 2018, 62/660,523 filed April 20, 2018, and 62/771,409 filed November 26, 2018, each of which is incorporated herein by reference in its entirety.

FIELD OF THE INVENTION

[0002] This disclosure provides methods for identifying in a nucleic acid sequence the locations of 5-methylcytosine, 5-hydroxymethylcytosine, 5-carboxylcytosine and/or 5-formylcytosine.

BACKGROUND

[0003] 5-Methylcytosine (5mC) and 5-hydroxymethylcytosine (5hmC) are the two major epigenetic marks found in the mammalian genome. 5hmC is generated from 5mC by the ten-eleven translocation (TET) family dioxygenases. Tet can further oxidize 5hmC to 5-formylcytosine (5fC) and 5-carboxylcytosine (5caC), which exists in much lower abundance in the mammalian genome compared to 5mC and 5hmC (10-fold to 100-fold lower than that of 5hmC). Together, 5mC and 5hmC play crucial roles in a broad range of biological processes from gene regulation to normal development. Aberrant DNA methylation and hydroxymethylation have been associated with various diseases and are well-accepted hallmarks of cancer. Therefore, the determination of 5mC and 5hmC in DNA sequence is not only important for basic research, but also is valuable for clinical applications, including diagnosis and therapy.

[0004] 5fC and 5caC are the two final oxidized derivatives of 5mC and can be converted to unmodified cytosine by Thymine DNA glycosylase (TDG) in base excision repair pathway. Therefore, 5fC and 5caC are two important key intermediates in the active demethylation process, which plays important role in embryonic development. 5fC and 5caC are found in these contexts and may serve as indicator of nearly complete 5mC demethylation. 5fC and 5caC may also play additional functions such as bind specific proteins and affect the rate and specificity of RNA polymerase II.

[0005] 5mC is also a post-transcriptional RNA modification that has been identified in both stable and highly abundant tRNAs and rRNAs, and in mRNAs. In addition, 5mC has been detected in snRNA (small nuclear RNA), miRNA (microRNA), lncRNA (long noncoding RNA) and eRNA (enhancer RNA). However, there appears to be differences in the occurrence of 5mC in specific RNA types in different organisms. For example, 5mC appears not to be present in tRNA and mRNA from bacteria, while it has been found in tRNA and mRNA in eukaryotes and archaea.

[0006] 5hmC has also been detected in RNA. For example, mRNA from *Drosophila* and mouse has been found to contain 5hmC. The same family of enzymes that oxidize 5mC in DNA was reported to catalyze the formation of 5hmC in mammalian total RNA. In flies, a transcriptome wide study using methylation RNA immunoprecipitation sequencing (MeRIP-seq) with 5hmC antibodies, detected the presence of 5hmC in many mRNA coding sequences, with particularly high levels in the brain. It was also reported that active translation is associated with high 5hmC levels in RNA, and flies lacking the TET enzyme responsible for 5hmC deposition in RNA have impaired brain development.

[0007] The current gold standard and most widely used method for DNA methylation and hydroxymethylation analysis is bisulfite sequencing (BS), and its derived methods such as Tet-assisted bisulfite sequencing (TAB-Seq) and oxidative bisulfite sequencing (oxBS). All of these methods employ bisulfite treatment to convert unmethylated cytosine to uracil while leaving 5mC and/or 5hmC intact. Through PCR amplification of the bisulfite-treated DNA, which reads uracil as thymine, the modification information of each cytosine can be inferred at a single base resolution (where the transition of C to T provides the location of the unmethylated cytosine). There are, however, at least two main drawbacks to bisulfite sequencing. First, bisulfite treatment is a harsh chemical reaction, which degrades more than 90% of the DNA due to depurination under the required acidic and thermal conditions. This degradation severely limits its application to low-input samples, such as clinical samples including circulating cell-free DNA and single-cell sequencing. Second, bisulfite sequencing relies on the complete conversion of unmodified cytosine to thymine. Unmodified cytosine accounts for approximately 95% of the total cytosine in the human genome. Converting all these positions to thymine severely reduces sequence complexity, leading to poor sequencing quality, low mapping rates, uneven genome coverage and increased sequencing cost. Bisulfite sequencing methods are also susceptible to false detection of 5mC and 5hmC due to incomplete conversion of unmodified cytosine to thymine.

[0008] Bisulfite sequencing has also been used to detect cytosine methylation in RNA. Unlike other methods for detecting 5mC in RNA such as methylated-RNA-immunoprecipitation, RNA-bisulfite-sequencing (RNA-BS-seq) has the advantage of being able to determine of the extent of methylation of a specific C position in RNA. RNA-BS-seq, however, suffers from the same drawbacks described above for bisulfite sequencing of DNA. In particular, the reaction conditions can cause substantial degradation of RNA.

[0009] There is a need for a method for DNA methylation and hydroxymethylation analysis that is a mild reaction that can detect the modified cytosine (5mC and 5hmC) at base-resolution quantitatively without affecting the unmodified cytosine. Likewise, there is a need for a method for RNA methylation and hydroxymethylation analysis that employs mild reaction conditions and can detect the modified cytosine quantitatively at base resolution without affecting the unmodified cytosine.

SUMMARY OF THE INVENTION

[0010] The present invention provides methods for identifying the location of one or more of 5-methylcytosine, 5- hydroxymethylcytosine, 5-carboxylcytosine and/or 5-formylcytosine in a nucleic acid. The methods described herein provide for DNA or RNA methylation and hydroxymethylation analysis involving mild reactions that detect the modified cytosine quantitatively with base-resolution without affecting the unmodified cytosine. Provided herein is a new method for identifying 5mC and 5hmC by combining TET oxidation and reduction by borane derivatives (e.g., pyridine borane and 2-picoline borane (pic-BH₃)), referred to herein as TAPS (TET Assisted Pyridine borane Sequencing) (Table 1). TAPS detects modifications directly with high sensitivity and specificity, without affecting unmodified cytosines, and can be adopted to detect other cytosine modifications. It is non-destructive, preserving RNA and DNA up to 10 kbs long. Compared with bisulfite sequencing, TAPS results in higher mapping rates, more even coverage and lower sequencing costs, enabling higher quality, more comprehensive and cheaper methylome analyses. Variations of this method that do not employ the oxidation step are used to identify 5fC and/or 5caC as described herein.

[0011] In one aspect, the present invention provides a method for identifying 5-methylcytosine (5mC) in a target nucleic acid comprising the steps of:

- a. providing a nucleic acid sample comprising the target nucleic acid;
- b. modifying the nucleic acid comprising the steps of:

- i. adding a blocking group to the 5-hydroxymethylcytosine (5hmC) in the nucleic acid sample;
- ii. converting the 5mC in the nucleic acid sample to 5-carboxylcytosine (5caC) and/or 5-formylcytosine (5fC); and
- iii. converting the 5caC and/or 5fC to dihydrouracil (DHU) to provide a modified nucleic acid sample comprising a modified target nucleic acid; and
- c. detecting the sequence of the modified target nucleic acid; wherein a cytosine (C) to thymine (T) transition in the sequence of the modified target nucleic acid compared to the target nucleic acid provides the location of a 5mC in the target nucleic acid.

[0012] In embodiments of the method for identifying 5mC in a target nucleic acid, the percentages of the T at each transition location provide a quantitative level of 5mC at each location in the target nucleic acid. In embodiments, the nucleic acid is DNA. In other embodiments, nucleic acid is RNA.

[0013] In another aspect, the present invention provides a method for identifying 5mC or 5hmC in a target nucleic acid comprising the steps of:

- a. providing a nucleic acid sample comprising the target nucleic acid;
- b. modifying the nucleic acid comprising the steps of:
 - i. converting the 5mC and 5hmC in the nucleic acid sample to 5-carboxylcytosine (5caC) and/or 5fC; and
 - ii. converting the 5caC and/or 5fC to DHU to provide a modified nucleic acid sample comprising a modified target nucleic acid; and
- c. detecting the sequence of the modified target nucleic acid; wherein a cytosine (C) to thymine (T) transition in the sequence of the modified target nucleic acid compared to the target nucleic acid provides the location of either a 5mC or 5hmC in the target nucleic acid.

[0014] In embodiments of the method for identifying 5mC or 5hmC, the percentages of the T at each transition location provide a quantitative level of 5mC or 5hmC at each location in the target nucleic acid. In embodiments, the nucleic acid is DNA. In other embodiments, nucleic acid is RNA.

[0015] In another aspect, the invention provides a method for identifying 5mC and identifying 5hmC in a target nucleic acid comprising:

- a. identifying 5mC in the target nucleic acid comprising the steps of:
 - i. providing a first nucleic acid sample comprising the target nucleic acid;

- ii. modifying the nucleic acid in the first sample comprising the steps of:
 - 1. adding a blocking group to the 5-hydroxymethylcytosine (5hmC) in the first nucleic acid sample;
 - 2. converting the 5mC in the first nucleic acid sample to 5caC and/or 5fC; and
 - 3. converting the 5caC and/or 5fC to DHU to provide a modified first DNA sample comprising a modified target nucleic acid;
 - iii. optionally amplifying the copy number of the modified target nucleic acid; and
 - iv. detecting the sequence of the modified target nucleic acid; wherein a cytosine (C) to thymine (T) transition in the sequence of the modified target nucleic acid compared to the target nucleic acid provides the location of a 5mC in the target nucleic acid.
- b. identifying 5mC or 5hmC in the target nucleic acid comprising the steps of:
- i. providing a second nucleic acid sample comprising the target nucleic acid;
 - ii. modifying the nucleic acid in the second sample comprising the steps of:
 - 1. converting the 5mC and 5hmC in the second nucleic acid sample to 5caC and/or 5fC; and
 - 2. converting the 5caC and/or 5fC to DHU to provide a modified second nucleic acid sample comprising a modified target nucleic acid;
 - iii. optionally amplifying the copy number of the modified target nucleic acid;
 - iv. detecting the sequence of the modified target nucleic acid from the second sample; wherein a cytosine (C) to thymine (T) transition in the sequence of the modified target nucleic acid compared to the target nucleic acid provides the location of either a 5mC or 5hmC in the target nucleic acid; and
- c. comparing the results of steps (a) and (b), wherein a C to T transitions present in step (b) but not in step (a) provides the location of 5hmC in the target nucleic acid.

[0016] In embodiments for identifying 5mC and identifying 5hmC in a target nucleic acid, in step (a) the percentages of the T at each transition location provide a quantitative level of 5mC in the target nucleic acid; in step (b), the percentages of the T at each transition location provide a quantitative level of 5mC or 5hmC in the target nucleic acid; and in step (c) the differences in percentages for a C to T transition identified in step (b), but not in step (a) provides the quantitative level of a 5hmC at each location in the target nucleic acid. In embodiments, the nucleic acid is DNA. In other embodiments, nucleic acid is RNA.

[0017] In embodiments of the invention, the blocking group added to 5hmC in the nucleic acid sample is a sugar. In embodiments, the sugar is a naturally-occurring sugar or a modified sugar, for example glucose or a modified glucose. In embodiments of the invention, the blocking group is added to 5hmC by contacting the nucleic acid sample with UDP linked to a sugar, for example UDP-glucose or UDP linked to a modified glucose in the presence of a glucosyltransferase enzyme, for example, T4 bacteriophage β -glucosyltransferase (β GT) and T4 bacteriophage α -glucosyltransferase (α GT) and derivatives and analogs thereof.

[0018] In embodiments of the invention, the step of converting the 5mC in the nucleic acid sample to 5caC and/or 5fC and the step of converting the 5mC and 5hmC in the nucleic acid sample to 5caC and/or 5fC each comprises contacting the nucleic acid sample with a ten eleven translocation (TET) enzyme. In further embodiments, the TET enzyme is one or more of human TET1, TET2, and TET3; murine Tet1, Tet2, and Tet3; *Naegleria* TET (NgTET); *Coprinopsis cinerea* (CcTET) and derivatives or analogues thereof. In embodiments, the TET enzyme is NgTET.

[0019] In another aspect, the invention provides a method for identifying 5caC or 5fC in a target nucleic acid comprising the steps of:

- a. providing a nucleic acid sample comprising the target nucleic acid;
- b. converting the 5caC and 5fC to DHU to provide a modified nucleic acid sample comprising a modified target nucleic acid;
- c. optionally amplifying the copy number of the modified target nucleic acid; and
- d. detecting the sequence of the modified target nucleic acid; wherein a cytosine (C) to thymine (T) transition in the sequence of the modified target nucleic acid compared to the target nucleic acid provides the location of either a 5caC or 5fC in the target nucleic acid.

[0020] In embodiments of the method for identifying 5caC or 5fC in a target nucleic acid, the percentages of the T at each transition location provide a quantitative level for 5caC or 5fC at each location in the target nucleic acid.

[0021] In another aspect, the invention provides a method for identifying 5caC in a target nucleic acid comprising the steps of:

- a. providing a nucleic acid sample comprising the target nucleic acid;
- b. adding a blocking group to the 5fC in the nucleic acid sample;
- c. converting the 5caC to DHU to provide a modified nucleic acid sample comprising a modified target nucleic acid;

- d. optionally amplifying the copy number of the modified target nucleic acid; and
- e. determining the sequence of the modified target nucleic acid; wherein a cytosine (C) to thymine (T) transition in the sequence of the modified target nucleic acid compared to the target nucleic acid provides the location of a 5caC in the target nucleic acid.

[0022] In embodiments of the method for identifying 5caC in a target nucleic acid, the percentages of the T at each transition location provide a quantitative level for 5caC at each location in the target nucleic acid. In embodiments, the nucleic acid is DNA. In other embodiments, nucleic acid is RNA.

[0023] In embodiments of the invention, adding a blocking group to the 5fC in the nucleic acid sample comprises contacting the nucleic acid with an aldehyde reactive compound including, for example, hydroxylamine derivatives (such as O-ethylhydroxylamine), hydrazine derivatives, and hydrazide derivatives.

[0024] In another aspect, the invention provides a method for identifying 5fC in a target nucleic acid comprising the steps of:

- a. providing a nucleic acid sample comprising the target nucleic acid;
- b. adding a blocking group to the 5caC in the nucleic acid sample
- c. converting the 5fC to DHU to provide a modified nucleic acid sample comprising a modified target nucleic acid;
- d. optionally amplifying the copy number of the modified target nucleic acid;
- e. detecting the sequence of the modified target nucleic acid; wherein a cytosine (C) to thymine (T) transition in the sequence of the modified target nucleic acid compared to the target nucleic acid provides the location of a 5fC in the target nucleic acid.

[0025] In embodiments of the method for identifying 5fC in a target nucleic acid, the percentages of the T at each transition location provide a quantitative level for 5fC at each location in the target nucleic acid. In embodiments, the nucleic acid (sample and target) is DNA. In other embodiments, nucleic acid (sample and target) is RNA.

[0026] In embodiments, the step of adding a blocking group to the 5caC in the nucleic acid sample comprises contacting the nucleic acid sample with a carboxylic acid derivatization reagent, including, for example, 1-ethyl-3-(3-dimethylaminopropyl)carbodiimide (EDC) and (ii) an amine (such as ethylamine), hydrazine, or hydroxylamine compound.

[0027] In embodiments of the invention, the methods above further comprise the step of amplifying the copy number of the modified target nucleic acid. In embodiments, this amplification step is performed prior to the step of detecting the sequence of the modified target nucleic acid. The step of amplifying the copy number when the modified target nucleic acid is DNA may be accomplished by performing the polymerase chain reaction (PCR), primer extension, and/or cloning. When the modified target nucleic acid is RNA, the step of amplifying the copy number may be accomplished by RT-PCR using oligo(dT) primer (for mRNA), random primers, and/or gene specific primers.

[0028] In embodiments of the invention, the DNA sample comprises picogram quantities of DNA. In embodiments of the invention, the DNA sample comprises about 1 pg to about 900 pg DNA, about 1 pg to about 500 pg DNA, about 1 pg to about 100 pg DNA, about 1 pg to about 50 pg DNA, about 1 to about 10 pg, DNA, less than about 200 pg, less than about 100 pg DNA, less than about 50 pg DNA, less than about 20 pg DNA, and less than about 5 pg DNA. In other embodiments of the invention, the DNA sample comprises nanogram quantities of DNA. In embodiments of the invention, the DNA sample contains about 1 to about 500 ng of DNA, about 1 to about 200 ng of DNA, about 1 to about 100 ng of DNA, about 1 to about 50 ng of DNA, about 1 ng to about 10 ng of DNA, about 1 ng to about 5 ng of DNA, less than about 100 ng of DNA, less than about 50 ng of DNA less than about 5 ng of DNA, or less than about 2 ng of DNA. In embodiments of the invention, the DNA sample comprises circulating cell-free DNA (cfDNA). In embodiments of the invention the DNA sample comprises microgram quantities of DNA.

[0029] In embodiments of the invention, the step of converting the 5caC and/or 5fC to DHU comprises contacting the nucleic acid sample with a reducing agent including, for example, pyridine borane, 2-picoline borane (pic-BH₃), borane, sodium borohydride, sodium cyanoborohydride, and sodium triacetoxymethylborohydride. In a preferred embodiment, the reducing agent is pic-BH₃ and/or pyridine borane.

[0030] In embodiments of the invention, the step of determining the sequence of the modified target nucleic acid comprises chain termination sequencing, microarray, high-throughput sequencing, and restriction enzyme analysis.

BRIEF DESCRIPTION OF THE FIGURES

[0031] **Fig. 1. Borane-containing compounds screening.** Borane-containing compounds were screened for conversion of 5caC to DHU in an 11mer oligonucleotide (“oligo”), with conversion rate estimated by MALDI. 2-picoline borane (pic-borane), borane

pyridine, and tert-butylamine borane could completely convert 5caC to DHU while ethylenediamine borane and dimethylamine borane gave around 30% conversion rate. No detectable products measured (n.d.) with dicyclohexylamine borane, morpholine borane, 4-methylmorpholine borane, and trimethylamine borane. Other reducing agents such as sodium borohydride and sodium tri(acetoxy)borohydride decomposed rapidly in acidic media and lead to incomplete conversion. Sodium cyanoborohydride was not used due to potential for hydrogen cyanide formation under acidic condition. Pic-borane and pyridine borane were chosen because of complete conversion, low toxicity and high stability.

[0032] Fig. 2A-B. Pic-borane reaction on DNA oligos. (A) MALDI characterization of 5caC-containing 11mer model DNA treated with pic-borane. Calculated mass (m/z) shown above each graph, observed mass shown to the left of the peak. (B) The conversion rates of dC and various cytosine derivatives were quantified by HPLC-MS/MS. Data shown as mean $\pm$ SD of three replicates.

[0033] Fig. 3A-B. Single nucleoside pic-borane reaction. ^1H and ^{13}C NMR results were in accordance with previous report on 2'-deoxy-5,6-dihydrouridine (I. Aparici-Espert et al., *J. Org. Chem.* **81**, 4031–4038 (2016)). (A) ^1H NMR (MeOH- d_4 , 400 MHz) chart of the single nucleoside pic-borane reaction product. δ ppm: 6.28 (t, 1H, $J = 7$ Hz), 4.30 (m, 1H), 3.81 (m, 1H), 3.63 (m, 2H), 3.46 (m, 2H), 2.65 (t, 2H, $J = 6$ Hz), 2.20 (m, 1H), 2.03 (m, 1H). (B) ^{13}C NMR (MeOH- d_4 , 400 MHz) chart of the single nucleoside pic-borane reaction product. δ ppm: 171.56 (CO), 153.54 (CO), 85.97 (CH), 83.86 (CH), 70.99 (CH), 61.92 (CH_2), 36.04 (CH_2), 35.46 (CH_2), 30.49 (CH_2).

[0034] Fig. 4A-B. A diagram showing (A) borane conversion of 5caC to DHU and a proposed mechanism for borane reaction of 5caC to DHU; and (B) borane conversion of 5fC to DHU and a proposed mechanism for borane reaction of 5fC to DHU.

[0035] Fig. 5A-B. (A) Diagram showing that the TAPS method converts both 5mC and 5hmC to DHU, which upon replication acts as thymine. (B) Overview of the TAPS, TAPS β , and CAPS methods.

[0036] Fig. 6. MALDI characterization of 5fC and 5caC containing model DNA oligos treated by pic-borane with or without blocking 5fC and 5caC. 5fC and 5caC are converted to dihydrouracil (DHU) with pic-BH₃. 5fC was blocked by hydroxylamine derivatives such as O-ethylhydroxylamine (EtONH₂) which would become oxime and resist pic-borane conversion. 5caC was blocked by ethylamine via EDC conjugation and converted to amide which blocks conversion by pic-borane. Calculated MS (m/z) shown above each graph, observed MS shown to the left of the peak.

[0037] Fig. 7. MALDI characterization of 5mC and 5hmC containing model DNA oligos treated by K₂Cr₂O₇ and pic-borane with or without blocking of 5hmC. 5hmC could be blocked by β GT with glucose and converted to 5gmC. 5mC, 5hmC and 5gmC could not be converted by pic-borane. 5hmC could be oxidized by K₂Cr₂O₇ to 5fC, and then converted to DHU by pic-borane. Calculated MS (m/z) shown above each graph, observed MS shown to the left of the peak.

[0038] Fig. 8A-B. Restriction enzyme digestion showed TAPS could effectively convert 5mC to T. (A) Illustration of restriction enzyme digestion assay to confirm sequence change caused by TAPS. (B) Taq α I-digestion tests to confirm the C-to-T transition caused by TAPS. TAPS was performed on a 222 bp model DNA having a Taq α I restriction site and containing 5 fully methylated CpG sites (5mC) and its unmethylated control (C). PCR-amplified 222 bp model DNA can be cleaved with Taq α I to ~160 bp and ~60 bp fragments as shown in the 5mC, C and C TAPS. After TAPS on the methylated DNA, the T(mC)GA sequence is converted to TTGA and is no longer cleaved by Taq α I digestion as shown in the 5mC-TAPS lanes.

[0039] Fig. 9A-B. TAPS on a 222 bp model DNA and mESC gDNA. (A) Sanger sequencing results for the 222 bp model DNA containing 5 fully methylated CpG sites and its unmethylated control before (5mC, C) and after TAPS (5mC TAPS, C TAPS). Only 5mC is converted to T by the TAPS method. (B) HPLC-MS/MS quantification of relative modification levels in the mESCs gDNA control, after NgTET1 oxidation and after pic-borane reduction. Data shown as mean $\pm$ SD of three replicates.

[0040] Fig. 10A-D. TAPS caused no significant DNA degradation compared to bisulfite. Agarose gel images of 222 bp unmethylated DNA, 222 bp methylated DNA, and mESC gDNA (A) before and (B) after chilling in an ice bath. No detectable DNA degradation was observed after TAPS and DNA remained double-stranded and could be visualized without chilling. Bisulfite conversion created degradation and DNA became single-stranded and could be visualized only after chilled on ice. (C) Agarose gel images of mESC gDNA of various fragment lengths treated with TAPS and bisulfite before (left panel) and after (right panel) cooling down on ice. DNA after TAPS remained double-stranded and could be directly visualized on the gel. Bisulfite treatment caused more damage and fragmentation to the samples and DNA became single-stranded and could be visualized only after chilled on ice. TAPS conversion was complete for all gDNA regardless of fragment length as shown in Fig. 15. (D) Agarose gel imaging of a 222 bp model DNA before and after TAPS (three independent repeats) showed no detectable degradation after the reaction.

[0041] Fig. 11. Comparison of amplification curves and melting curves between model DNAs before and after TAPS. qPCR assay showed minor difference on model DNAs before and after TAPS in amplification curves. Melting curve of methylated DNA (5mC) shifted to lower temperature after TAPS indicated possible T_m-decreasing C-to-T transition while there was no shift for unmethylated DNA (C).

[0042] Fig. 12. Complete C-to-T transition induced after TAPS, TAPS β and CAPS as demonstrated by Sanger sequencing. Model DNA containing single methylated and single hydroxymethylated CpG sites was prepared as described herein. TAPS conversion was done following NgTET1 oxidation and pyridine borane reduction protocol as described herein. TAPS β conversion was done following 5hmC blocking, NgTET1 Oxidation and Pyridine borane reduction protocol. CAPS conversion was done following 5hmC oxidation and Pyridine borane reduction protocol. After conversion, 1 ng of converted DNA sample was PCR amplified by Taq DNA Polymerase and processed for Sanger sequencing. TAPS converted both 5mC and 5hmC to T. TAPS β selectively converted 5mC whereas CAPS selectively converted 5hmC. None of the three methods caused conversion on unmodified cytosine and other bases.

[0043] Fig. 13A-B. (A) TAPS is compatible with various DNA and RNA polymerase and induces complete C-to-T transition as shown by Sanger sequencing. Model DNA containing methylated CpG sites for the polymerase test and primer sequences are described herein. After TAPS treatment, 5mC was converted to DHU. KAPA HiFi Uracil plus polymerase, Taq polymerase, and Vent exo-polymerase would read DHU as T and therefore induce complete C-to-T transition after PCR. Alternatively, primer extension was done with biotin-labelled primer and isothermal polymerases including Klenow fragment, Bst DNA polymerase, and phi29 DNA polymerase. The newly synthesized DNA strand was separated by Dynabeads MyOne Streptavidin C1 and then amplified by PCR with Taq polymerase and processed for Sanger sequencing. T7 RNA polymerase could efficiently bypass DHU and insert adenine opposite to DHU site, which is proved by RT-PCR and Sanger sequencing. **(B)** Certain other commercialized polymerases did not amplify DHU containing DNA efficiently. After TAPS treatment, 5mC was converted to DHU. KAPA HiFi Uracil plus polymerase and Taq polymerase would read DHU as T and therefore induce complete C-to-T transition. Low or no C-to-T transition was observed with certain other commercialized polymerases including KAPA HiFi polymerase, Pfu polymerase, Phusion polymerase and NEB Q5 polymerase (not shown).

[0044] Fig. 14. DHU does not show PCR bias compared to T and C. Model DNA containing one DHU/U/T/C modification was synthesized with the corresponding DNA oligos as described in herein. Standard curves for each model DNA with DHU/U/T/C modification were plotted based on qPCR reactions with 1:10 serial dilutions of the model DNA input (from 0.1 pg to 1 ng, every qPCR experiment was run in triplicates). The slope of the regression between the log concentration (ng) values and the average Ct values was calculated by SLOPE function in Excel. PCR efficiency was calculated using the following equation: $\text{Efficiency\%} = (10^{(-1/\text{Slope})} - 1) * 100\%$. Amplification factor was calculated using the following equation: $\text{Amplification factor} = 10^{(-1/\text{Slope})}$. PCR efficiency for model DNAs with DHU or T or C modification were almost the same, which demonstrated that DHU could be read through as a regular base and would not cause PCR bias.

[0045] Fig. 15A-B. TAPS completely converted 5mC to T regardless of DNA fragment length. (A) Agarose gel images of TaqαI-digestion assay confirmed complete 5mC to T conversion in all samples regardless of DNA fragment lengths. 194 bp model sequence from lambda genome was PCR amplified after TAPS and digested with TaqαI enzyme. PCR product amplified from unconverted sample could be cleaved, whereas products amplified on TAPS treated samples stayed intact, suggesting loss of restriction site and hence complete 5mC-to-T transition. (B) The C-to-T conversion percentage was estimated by gel band quantification and shown 100% for all DNA fragment lengths tested.

[0046] Fig. 16. The conversion and false positive for different TAPS conditions. The combination of mTet1 and pyridine borane achieved the highest conversion rate of methylated C (96.5%, calculated with fully CpG methylated Lambda DNA) and the lowest conversion rate of unmodified C (0.23%, calculated with 2 kb unmodified spike-in), compared to other conditions with NgTET1 or pic-borane. Showing above bars the conversion rate +/- SE of all tested cytosine sites.

[0047] Fig. 17A-B. Conversion rate on short spike-ins. (A) 120mer-1 and (B) 120mer-2 containing 5mC and 5hmC. Near complete conversion was archived on 5mC and 5hmC sites from both strands. Actual sequences with modification status shown on top and bottom.

[0048] Fig. 18A-E. Improved sequencing quality of TAPS over Whole Genome Bisulfite Sequencing (WGBS). (A) Conversion rate of 5mC and 5hmC in TAPS-treated DNA. Left: Synthetic spike-ins (CpN) methylated or hydroxymethylated at known positions. (B) False positive rate of TAPS from unmodified 2 kb spike-in. (C) Total run time of TAPS and WGBS when processing 1 million simulated reads on one core of one Intel Xeon CPU. (D) Fraction of all sequenced read pairs (after trimming) mapped to the genome. (E)

Sequencing quality scores per base for the first and second reads in all sequenced read pairs, as reported by Illumina BaseSpace. Top: TAPS. Bottom: WGBS.

[0049] Fig. 19A-B. TAPS resulted in more even coverage and fewer uncovered positions than WGBS. Comparison of depth of coverage across (A) all bases and (B) CpG sites between WGBS and TAPS, computed on both strands. For “TAPS (down-sampled)”, random reads out of all mapped TAPS reads were selected so that the median coverage matched the median coverage of WGBS. Positions with coverage above 50× are shown in the last bin.

[0050] Fig. 20. Distribution of modification levels across all chromosomes. Average modification levels in 100 kb windows along mouse chromosomes, weighted by the coverage of CpG, and smoothed using a Gaussian weighted moving average filter with window size 10.

[0051] Fig. 21A-E. Comparison of genome-wide methylome measurements by TAPS and WGBS. (A) Average sequencing coverage depth in all mouse CpG islands (binned into 20 windows) and 4 kbp flanking regions (binned into 50 equally sized windows). To account for differences in sequencing depth, all mapped TAPS reads were down-sampled to match the median number of mapped WGBS reads across the genome. (B) CpG sites covered by at least three reads by TAPS alone, both TAPS and WGBS, or WGBS alone. (C) Number of CpG sites covered by at least three reads and modification level > 0.1 detected by TAPS alone, TAPS and WGBS, or WGBS alone. (D) Example of the chromosomal distribution of modification levels (in %) for TAPS and WGBS. Average fraction of modified CpGs per 100 kb windows along mouse chromosome 4, smoothed using a Gaussian-weighted moving average filter with window size 10. (E) Heatmap representing the number of CpG sites covered by at least three reads in both TAPS and WGBS, broken down by modification levels as measured by each method. To improve contrast, the first bin, containing CpGs unmodified in both methods, was excluded from the color scale and is denoted by a star.

[0052] Fig. 22. Modification levels around CpG Islands. Average modification levels in CpG islands (binned into 20 windows) and 4 kbp flanking regions (binned into 50 equally sized windows). Bins with coverage below 3 reads were ignored.

[0053] Fig. 23A-B. TAPS exhibits smaller coverage-modification bias than WGBS. All CpG sites were binned according to their coverage and the mean (circles) and the median (triangles) modification value is shown in each bin for WGBS (A) and TAPS (B). The CpG sites covered by more than 100 reads are shown in the last bin. The lines represent a linear fit through the data points.

[0054] Fig. 24A-C. Low-input gDNA and cell-free DNA TAPS prepared with dsDNA and ssDNA library preparation kits. (A) Sequencing libraries were successfully constructed with down to 1 ng of murine embryonic stem cell (mESC) genomic DNA (gDNA) with dsDNA library kits NEBNext Ultra II or KAPA Hyperplus kits. ssDNA library kit Accel-NGS Methyl-Seq kit was used to further lower the input DNA amount down to (B) 0.01 ng of mESC gDNA or (C) 1 ng of cell-free DNA.

[0055] Fig. 25A-B. Low-input gDNA and cell-free DNA TAPS libraries prepared with dsDNA KAPA Hyperplus library preparation kit. Sequencing libraries were successfully constructed with as little as 1 ng of (A) mESC gDNA and (B) cell-free DNA with KAPA Hyperplus kit. Cell-free DNA has a sharp length distribution around 160 bp (nucleosome size) due to plasma nuclease digestion. After library construction, it becomes ~300 bp, which is the sharp band in (B).

[0056] Fig. 26A-D. High-quality cell-free DNA TAPS. (A) Conversion rate of 5mC in TAPS-treated cfDNA. (B) False positive rate in TAPS-treated cfDNA. (C) Fraction of all sequenced read pairs that were uniquely mapped to the genome. (D) Fraction of all sequenced read pairs that were uniquely mapped to the genome and after removal of PCR duplication reads. CHG and CHH are non-CpG contexts.

[0057] Fig. 27. TAPS can detect genetic variants. Methylation (MOD, top row) and C-to-T SNPs (bottom row) showed distinct base distribution patterns in original top strand (OT)/original bottom strand (OB), left column, and in strands complementary to OT (CTOT) and OB (CTOB), right column.

DETAILED DESCRIPTION OF THE INVENTION

[0058] The present invention provides a bisulfite-free, base-resolution method for detecting 5mC and 5hmC in a sequence, herein named TAPS. TAPS consists of mild enzymatic and chemical reactions to detect 5mC and 5hmC directly and quantitatively at base-resolution without affecting unmodified cytosine. The present invention also provides methods to detect 5fC and 5caC at base resolution without affecting unmodified cytosine. Thus, the methods provided herein provide mapping of 5mC, 5hmC, 5fC and 5caC and overcome the disadvantages of previous methods such as bisulfite sequencing.

Table 1. Comparison of BS and related methods versus TAPS, TAPS β and CAPS for 5mC and 5hmC sequencing.

Base	BS	TAB-Seq	oxBS	TAPS	TAPS β	CAPS
C	T	T	T	C	C	C
5mC	C	T	C	T	T	C

5hmC		C	C	T	T	C	T
-------------	--	---	---	---	---	---	---

[0059] Methods for Identifying 5mC

[0060] In one aspect, the present invention provides a method for identifying 5-methylcytosine (5mC) in a target DNA comprising the steps of:

- a. providing a DNA sample comprising the target DNA;
- b. modifying the DNA comprising the steps of:
 - i. adding a blocking group to the 5-hydroxymethylcytosine (5hmC) in the DNA sample;
 - ii. converting the 5mC in the DNA sample to 5-carboxylcytosine (5caC) and/or 5-formylcytosine (5fC); and
 - iii. converting the 5caC and/or 5fC to DHU to provide a modified DNA sample comprising a modified target DNA;
- c. detecting the sequence of the modified target DNA; wherein a cytosine (C) to thymine (T) transition in the sequence of the modified target DNA compared.

[0061] In embodiments of the method for identifying 5mC in the target DNA, the method provides a quantitative measure for the frequency of 5mC modification at each location where the modification was identified in the target DNA. In embodiments, the percentages of the T at each transition location provide a quantitative level of 5mC at each location in the target DNA.

[0062] In order to identify 5mC in a target DNA without including 5hmC, the 5hmC in the sample is blocked so that it is not subject to conversion to 5caC and/or 5fC. In the methods of the present invention, 5hmC in the sample DNA are rendered non-reactive to the subsequent steps by adding a blocking group to the 5hmC. In one embodiment, the blocking group is a sugar, including a modified sugar, for example glucose or 6-azide-glucose (6-azido-6-deoxy-D-glucose). The sugar blocking group is added to the hydroxymethyl group of 5hmC by contacting the DNA sample with uridine diphosphate (UDP)-sugar in the presence of one or more glucosyltransferase enzymes.

[0063] In embodiments, the glucosyltransferase is T4 bacteriophage β -glucosyltransferase (β GT), T4 bacteriophage α -glucosyltransferase (α GT), and derivatives and analogs thereof. β GT is an enzyme that catalyzes a chemical reaction in which a beta-D-glucosyl (glucose) residue is transferred from UDP-glucose to a 5-hydroxymethylcytosine residue in a nucleic acid.

[0064] By stating that the blocking group is, for example, glucose, this refers to a glucose moiety (e.g., a beta-D-glucosyl residue) being added to 5hmC to yield glucosyl 5-hydroxymethyl cytosine. The sugar blocking group can be any sugar or modified sugar that is a substrate of the glucosyltransferase enzyme and blocks the subsequent conversion of the 5hmC to 5caC and/or 5fC. The step of converting the 5mC in the DNA sample to 5caC and/or 5fC is then accomplished by the methods provided herein, such as by oxidation using a TET enzyme. And converting the 5caC and/or 5fC to DHU is accomplished by the methods provided herein, such by borane oxidation.

[0065] The method for identifying 5-methylcytosine (5mC) can be performed on an RNA sample to identify the location of, and provide a quantitative measure of, 5mC in a target RNA.

[0066] Methods for Identifying 5mC or 5hmC (together)

[0067] In another aspect, the present invention provides a method for identifying 5mC or 5hmC in a target DNA comprising the steps of:

- a. providing a DNA sample comprising the target DNA;
- b. modifying the DNA comprising the steps of:
 - i. converting the 5mC and 5hmC in the DNA sample to 5-carboxylcytosine (5caC) and/or 5fC; and
 - ii. converting the 5caC and/or 5fC to DHU to provide a modified DNA sample comprising a modified target DNA;
- c. detecting the sequence of the modified target DNA; wherein a cytosine (C) to thymine (T) transition in the sequence of the modified target DNA compared to the target DNA provides the location of either a 5mC or 5hmC in the target DNA.

[0068] In embodiments of the method for identifying 5mC or 5hmC in the target DNA, the method provides a quantitative measure for the frequency the of 5mC or 5hmC modifications at each location where the modifications were identified in the target DNA. In embodiments, the percentages of the T at each transition location provide a quantitative level of 5mC or 5hmC at each location in the target DNA.

[0069] This method for identifying 5mC or 5hmC provides the location of 5mC and 5hmC, but does not distinguish between the two cytosine modifications. Rather, both 5mC and 5hmC are converted to DHU. The presence of DHU can be detected directly, or the modified DNA can be replicated by known methods where the DHU is converted to T.

[0070] The method for identifying 5mC or 5hmC can be performed on an RNA sample to identify the location of, and provide a quantitative measure of, 5mC or 5hmC in a target RNA.

[0071] Methods for Identifying 5mC and Identifying 5hmC

[0072] The present invention provides a method for identifying 5mC and identifying 5hmC in a target DNA by (i) performing the method for identifying 5mC on a first DNA sample described herein, and (ii) performing the method for identifying 5mC or 5hmC on a second DNA sample described herein. The location of 5mC is provided by (i). By comparing the results of (i) and (ii), wherein a C to T transitions detected in (ii) but not in (i) provides the location of 5hmC in the target DNA. In embodiments, the first and second DNA samples are derived from the same DNA sample. For example, the first and second samples may be separate aliquots taken from a sample comprising DNA to be analyzed.

[0073] Because the 5mC and 5hmC (that is not blocked) are converted to 5fC and 5caC before conversion to DHU, any existing 5fC and 5caC in the DNA sample will be detected as 5mC and/or 5hmC. However, given the extremely low levels of 5fC and 5caC in genomic DNA under normal conditions, this will often be acceptable when analyzing methylation and hydroxymethylation in a DNA sample. The 5fC and 5caC signals can be eliminated by protecting the 5fC and 5caC from conversion to DHU by, for example, hydroxylamine conjugation and EDC coupling, respectively.

[0074] The above method identifies the locations and percentages of 5hmC in the target DNA through the comparison of 5mC locations and percentages with the locations and percentages of 5mC or 5hmC (together). Alternatively, the location and frequency of 5hmC modifications in a target DNA can be measured directly. Thus, in one aspect the invention provides a method for identifying 5hmC in a target DNA comprising the steps of:

- a. providing a DNA sample comprising the target DNA;
- b. modifying the DNA comprising the steps of:
 - i. converting the 5hmC in the DNA sample to 5fC; and
 - ii. converting the 5fC to DHU to provide a modified DNA sample comprising a modified target DNA;
- c. detecting the sequence of the modified target DNA; wherein a cytosine (C) to thymine (T) transition in the sequence of the modified target DNA compared to the target DNA provides the location of a 5hmC in the target DNA.

[0075] In embodiments, the step of converting the 5hmC to 5fC comprises oxidizing the 5hmC to 5fC by contacting the DNA with, for example, potassium perruthenate (KRuO₄) (as

described in Science. 2012, 33, 934-937 and WO2013017853, incorporated herein by reference); or Cu(II)/TEMPO (copper(II) perchlorate and 2,2,6,6-tetramethylpiperidine-1-oxyl (TEMPO)) (as described in Chem. Commun., 2017,53, 5756-5759 and WO2017039002, incorporated herein by reference). The 5fC in the DNA sample is then converted to DHU by the methods disclosed herein, e.g., by the borane reaction.

[0076] The method for identifying 5mC and identifying 5hmC can be performed on an RNA sample to identify the location of, and provide a quantitative measure of, 5mC and 5hmC in a target RNA.

[0077] Methods for Identifying 5caC or 5fC

[0078] In one aspect, the invention provides a method for identifying 5caC or 5fC in a target DNA comprising the steps of:

- a. providing a DNA sample comprising the target DNA;
- b. converting the 5caC and/or 5fC to DHU to provide a modified DNA sample comprising a modified target DNA;
- c. optionally amplifying the copy number of the modified target DNA;
- d. detecting the sequence of the modified target DNA; wherein a cytosine (C) to thymine (T) transition in the sequence of the modified target DNA compared to the target DNA provides the location of either a 5caC or 5fC in the target DNA.

[0079] This method for identifying 5fC or 5caC provides the location of 5fC or 5caC, but does not distinguish between these two cytosine modifications. Rather, both 5fC and 5caC are converted to DHU, which is detected by the methods described herein.

[0080] Methods for Identifying 5caC

[0081] In another aspect, the invention provides a method for identifying 5caC in a target DNA comprising the steps of:

- a. providing a DNA sample comprising the target DNA;
- b. adding a blocking group to the 5fC in the DNA sample;
- c. converting the 5caC to DHU to provide a modified DNA sample comprising a modified target DNA;
- d. optionally amplifying the copy number of the modified target DNA; and
- e. determining the sequence of the modified target DNA; wherein a cytosine (C) to thymine (T) transition in the sequence of the modified target DNA compared to the target DNA provides the location of a 5caC in the target DNA.

[0082] In embodiments of the method for identifying 5caC in the target DNA, the method provides a quantitative measure for the frequency the of 5caC modification at each location

where the modification was identified in the target DNA. In embodiments, the percentages of the T at each transition location provide a quantitative level of 5caC at each location in the target DNA.

[0083] In this method, 5fC is blocked (and 5mC and 5hmC are not converted to DHU) allowing identification of 5caC in the target DNA. In embodiments of the invention, adding a blocking group to the 5fC in the DNA sample comprises contacting the DNA with an aldehyde reactive compound including, for example, hydroxylamine derivatives, hydrazine derivatives, and hyrazide derivatives. Hydroxylamine derivatives include ashydroxylamine; hydroxylamine hydrochloride; hydroxylammonium acid sulfate; hydroxylamine phosphate; O-methylhydroxylamine; O-hexylhydroxylamine; O-pentylhydroxylamine; O-benzylhydroxylamine; and particularly, O-ethylhydroxylamine (EtONH₂), O-alkylated or O-arylated hydroxylamine, acid or salts thereof. Hydrazine derivatives include N-alkylhydrazine, N-arylhydrazine, N-benzylhydrazine, N,N-dialkylhydrazine, N,N-diarylhydrazine, N,N-dibenzylhydrazine, N,N-alkylbenzylhydrazine, N,N-arylbenzylhydrazine, and N,N-alkylarylhydrazine. Hydrazide derivatives include -toluenesulfonylhydrazide, N-acylhydrazide, N,N-alkylacylhydrazide, N,N-benzylacylhydrazide, N,N-arylacylhydrazide, N-sulfonylhydrazide, N,N-alkylsulfonylhydrazide, N,N-benzylsulfonylhydrazide, and N,N-arylsulfonylhydrazide.

[0084] The method for identifying 5caC can be performed on an RNA sample to identify the location of, and provide a quantitative measure of, 5caC in a target RNA.

[0085] Methods for Identifying 5fC

[0086] In another aspect, the invention provides a method for identifying 5fC in a target DNA comprising the steps of:

- a. providing a DNA sample comprising the target DNA;
- b. adding a blocking group to the 5caC in the DNA sample;
- c. converting the 5fC to DHU to provide a modified DNA sample comprising a modified target DNA;
- d. optionally amplifying the copy number of the modified target DNA;
- e. detecting the sequence of the modified target DNA; wherein a cytosine (C) to thymine (T) transition in the sequence of the modified target DNA compared to the target DNA provides the location of a 5fC in the target DNA.

[0087] In embodiments of the method for identifying 5fC in the target DNA, the method provides a quantitative measure for the frequency of 5fC modification at each location where the modification was identified in the target DNA. In embodiments, the percentages of

the T at each transition location provide a quantitative level of 5fC at each location in the target DNA.

[0088] Adding a blocking group to the 5caC in the DNA sample can be accomplished by (i) contacting the DNA sample with a coupling agent, for example a carboxylic acid derivatization reagent like carbodiimide derivatives such as 1-ethyl-3-(3-dimethylaminopropyl)carbodiimide (EDC) or N,N'-dicyclohexylcarbodiimide (DCC) and (ii) contacting the DNA sample with an amine, hydrazine or hydroxylamine compound. Thus, for example, 5caC can be blocked by treating the DNA sample with EDC and then benzylamine, ethylamine or other amine to form an amide that blocks 5caC from conversion to DHU by, e.g., pic-BH₃. Methods for EDC-catalyzed 5caC coupling are described in WO2014165770, and are incorporated herein by reference.

[0089] The method for identifying 5fC can be performed on an RNA sample to identify the location of, and provide a quantitative measure of, 5fC in a target RNA.

[0090] Nucleic Acid Sample / Target Nucleic Acid

[0091] The present invention provides methods for identifying the location of one or more of 5-methylcytosine, 5-hydroxymethylcytosine, 5-carboxylcytosine and/or 5-formylcytosine in a target nucleic acid quantitatively with base-resolution without affecting the unmodified cytosine. In embodiments, the target nucleic acid is DNA. In other embodiments, the target nucleic acid is RNA. Likewise the nucleic acid sample that comprises the target nucleic acid may be a DNA sample or an RNA sample.

[0092] The target nucleic acid may be any nucleic acid having cytosine modifications (i.e., 5mC, 5hmC, 5fC, and/or 5caC). The target nucleic acid can be a single nucleic acid molecule in the sample, or may be the entire population of nucleic acid molecules in a sample (or a subset thereof). The target nucleic acid can be the native nucleic acid from the source (e.g., cells, tissue samples, etc.) or can pre-converted into a high-throughput sequencing-ready form, for example by fragmentation, repair and ligation with adaptors for sequencing. Thus, target nucleic acids can comprise a plurality of nucleic acid sequences such that the methods described herein may be used to generate a library of target nucleic acid sequences that can be analyzed individually (e.g., by determining the sequence of individual targets) or in a group (e.g., by high-throughput or next generation sequencing methods).

[0093] A nucleic acid sample can be obtained from an organism from the Monera (bacteria), Protista, Fungi, Plantae, and Animalia Kingdoms. Nucleic acid samples may be obtained from a from a patient or subject, from an environmental sample, or from an

organism of interest. In embodiments, the nucleic acid sample is extracted or derived from a cell or collection of cells, a body fluid, a tissue sample, an organ, and an organelle.

[0094] RNA Sample / Target RNA

[0095] The present invention provides methods for identifying the location of one or more of 5-methylcytosine, 5-hydroxymethylcytosine, 5-carboxylcytosine and/or 5-formylcytosine in a target RNA quantitatively with base-resolution without affecting the unmodified cytosine. In embodiments, the RNA is one or more of mRNA (messenger RNA), tRNA (transfer RNA), rRNA (ribosomal RNA), snRNA (small nuclear RNA), miRNA (microRNA), lncRNA (long noncoding RNA) and eRNA (enhancer RNA). The target RNA can be a single RNA molecule in the sample, or may be the entire population of RNA molecules in a sample (or a subset thereof). Thus, target RNA can comprise a plurality of RNA sequences such that the methods described herein may be used to generate a library of target RNA sequences that can be analyzed individually (e.g., by determining the sequence of individual targets) or in a group (e.g., by high-throughput or next generation sequencing methods).

[0096] DNA Sample / Target DNA

[0097] The methods of the invention utilize mild enzymatic and chemical reactions that avoid the substantial degradation associated with methods like bisulfite sequencing. Thus, the methods of the present invention are useful in analysis of low-input samples, such as circulating cell-free DNA and in single-cell analysis.

[0098] In embodiments of the invention, the DNA sample comprises picogram quantities of DNA. In embodiments of the invention, the DNA sample comprises about 1 pg to about 900 pg DNA, about 1 pg to about 500 pg DNA, about 1 pg to about 100 pg DNA, about 1 pg to about 50 pg DNA, about 1 to about 10 pg, DNA, less than about 200 pg, less than about 100 pg DNA, less than about 50 pg DNA, less than about 20 pg DNA, and less than about 5 pg DNA. In other embodiments of the invention, the DNA sample comprises nanogram quantities of DNA. The sample DNA for use in the methods of the invention can be any quantity including, DNA from a single cell or bulk DNA samples. In embodiments, the methods of the present invention can be performed on a DNA sample comprising about 1 to about 500 ng of DNA, about 1 to about 200 ng of DNA, about 1 to about 100 ng of DNA, about 1 to about 50 ng of DNA, about 1 to about 10 ng of DNA, about 2 to about 5 ng of DNA, less than about 100 ng of DNA, less than about 50 ng of DNA less than 5 ng, and less than 2 ng of DNA. In embodiments of the invention the DNA sample comprises microgram quantities of DNA.

[0099] A DNA sample used in the methods described herein may be from any source including, for example a body fluid, tissue sample, organ, organelle, or single cells. In embodiments, the DNA sample is circulating cell-free DNA (cell-free DNA or cfDNA), which is DNA found in the blood and is not present within a cell. cfDNA can be isolated from blood or plasma using methods known in the art. Commercial kits are available for isolation of cfDNA including, for example, the Circulating Nucleic Acid Kit (Qiagen). The DNA sample may result from an enrichment step, including, but is not limited to antibody immunoprecipitation, chromatin immunoprecipitation, restriction enzyme digestion-based enrichment, hybridization-based enrichment, or chemical labeling-based enrichment.

[0100] The target DNA may be any DNA having cytosine modifications (i.e., 5mC, 5hmC, 5fC, and/or 5caC) including, but not limited to, DNA fragments or genomic DNA purified from tissues, organs, cells and organelles. The target DNA can be a single DNA molecule in the sample, or may be the entire population of DNA molecules in a sample (or a subset thereof). The target DNA can be the native DNA from the source or pre-converted into a high-throughput sequencing-ready form, for example by fragmentation, repair and ligation with adaptors for sequencing. Thus, target DNA can comprise a plurality of DNA sequences such that the methods described herein may be used to generate a library of target DNA sequences that can be analyzed individually (e.g., by determining the sequence of individual targets) or in a group (e.g., by high-throughput or next generation sequencing methods).

[0101] Converting 5mC and 5hmC to 5caC and/or 5fC

[0102] Embodiments of the present invention, such as the TAPS method described herein, include the step of converting the 5mC and 5hmC (or just the 5mC if the 5hmC is blocked) to 5caC and/or 5fC. In embodiments of the invention, this step comprises contacting the DNA or RNA sample with a ten eleven translocation (TET) enzyme. The TET enzymes are a family of enzymes that catalyze the transfer of an oxygen molecule to the N5 methyl group on 5mC resulting in the formation of 5-hydroxymethylcytosine (5hmC). TET further catalyzes the oxidation of 5hmC to 5fC and the oxidation of 5fC to form 5caC (*see* Fig. 5A). TET enzymes useful in the methods of the invention include one or more of human TET1, TET2, and TET3; murine Tet1, Tet2, and Tet3; *Naegleria* TET (NgTET); *Coprinopsis cinerea* (CcTET) and derivatives or analogues thereof. In embodiments, the TET enzyme is NgTET. In other embodiments the TET enzyme is human TET1 (hTET1).

[0103] Converting 5caC and/or 5fC to DHU

[0104] Methods of the present invention include the step of converting the 5caC and/or 5fC in a nucleic acid sample to DHU. In embodiments of the invention, this step comprises contacting the DNA or RNA sample with a reducing agent including, for example, a borane reducing agent such as pyridine borane, 2-picoline borane (pic-BH₃), borane, sodium borohydride, sodium cyanoborohydride, and sodium triacetoxyborohydride. In a preferred embodiment, the reducing agent is pyridine borane and/or pic-BH₃.

[0105] Amplifying the copy number of modified target nucleic acid

[0106] The methods of the invention may optionally include the step of amplifying (increasing) the copy number of the modified target Nucleic acid by methods known in the art. When the modified target nucleic acid is DNA, the copy number can be increased by, for example, PCR, cloning, and primer extension. The copy number of individual target DNAs can be amplified by PCR using primers specific for a particular target DNA sequence. Alternatively, a plurality of different modified target DNA sequences can be amplified by cloning into a DNA vector by standard techniques. In embodiments of the invention, the copy number of a plurality of different modified target DNA sequences is increased by PCR to generate a library for next generation sequencing where, e.g., double-stranded adapter DNA has been previously ligated to the sample DNA (or to the modified sample DNA) and PCR is performed using primers complimentary to the adapter DNA.

[0107] Detecting the Sequence of the Modified Target nucleic acid

[0108] In embodiments of the invention, the method comprises the step of detecting the sequence of the modified target nucleic acid. The modified target DNA or RNA contains DHU at positions where one or more of 5mC, 5hmC, 5fC, and 5caC were present in the unmodified target DNA or RNA. DHU acts as a T in DNA replication and sequencing methods. Thus, the cytosine modifications can be detected by any direct or indirect method that identifies a C to T transition known in the art. Such methods include sequencing methods such as Sanger sequencing, microarray, and next generation sequencing methods. The C to T transition can also be detected by restriction enzyme analysis where the C to T transition abolishes or introduces a restriction endonuclease recognition sequence.

[0109] Kits

[0110] The invention additionally provides kits for identification of 5mC and 5hmC in a target DNA. Such kits comprise reagents for identification of 5mC and 5hmC by the methods described herein. The kits may also contain the reagents for identification of 5caC and for the identification of 5fC by the methods described herein. In embodiments, the kit comprises a TET enzyme, a borane reducing agent and instructions for performing the

method. In further embodiments, the TET enzyme is TET1 and the borane reducing agent is selected from one or more of the group consisting of pyridine borane, 2-picoline borane (pic-BH₃), borane, sodium borohydride, sodium cyanoborohydride, and sodium triacetoxymethylborohydride. In a further embodiment, the TET1 enzyme is NgTet1 or murine Tet1 and the borane reducing agent is pyridine borane and/or pic-BH₃.

[0111] In embodiments, the kit further comprises a 5hmC blocking group and a glucosyltransferase enzyme. In further embodiments, the 5hmC blocking group is uridine diphosphate (UDP)-sugar where the sugar is glucose or a glucose derivative, and the glucosyltransferase enzyme is T4 bacteriophage β -glucosyltransferase (β GT), T4 bacteriophage α -glucosyltransferase (α GT), and derivatives and analogs thereof.

[0112] In embodiments the kit further comprises an oxidizing agent selected from potassium perruthenate (K₂RuO₄) and/or Cu(II)/TEMPO (copper(II) perchlorate and 2,2,6,6-tetramethylpiperidine-1-oxyl (TEMPO)).

[0113] In embodiments, the kit comprises reagents for blocking 5fC in the nucleic acid sample. In embodiments, the kit comprises an aldehyde reactive compound including, for example, hydroxylamine derivatives, hydrazine derivatives, and hydrazone derivatives as described herein. In embodiments, the kit comprises reagents for blocking 5caC as described herein.

[0114] In embodiments, the kit comprises reagents for isolating DNA or RNA. In embodiments the kit comprises reagents for isolating low-input DNA from a sample, for example cfDNA from blood, plasma, or serum.

EXAMPLES

[0115] **Methods**

[0116] ***Preparation of model DNA.***

[0117] **DNA oligos for MALDI and HPLC-MS/MS test.** DNA oligonucleotides ("oligos") with C, 5mC and 5hmC were purchased from Integrated DNA Technologies (IDT). All the sequences and modifications could be found in Figs. 6 and 7. DNA oligo with 5fC was synthesized by the C-tailing method: DNA oligos 5'-GTCGACCGGATC-3' and 5'-TTGGATCCGGTCTGACTT-3' were annealed and then incubated with 5-formyl-2'-dCTP (Trilink Biotech) and Klenow Fragment 3'→5' exo- (New England Biolabs) in NEBuffer 2 for 2 hr at 37°C. The product was purified with Bio-Spin P-6 Gel Columns (Bio-Rad).

[0118] DNA oligo with 5caC was synthesized using Expedite 8900 Nucleic Acid Synthesis System with standard phosphoramidites (Sigma) 5-Carboxy-dC-CE

Phosphoramidite (Glen Research). Subsequent deprotection and purification were carried out with Glen-Pak Cartridges (Glen Research) according to the manufacturer's instructions. Purified oligonucleotides were characterized by Voyager-DE MALDI-TOF (matrix-assisted laser desorption ionization time-of-flight) Biospectrometry Workstation.

[0119] 222 bp Model DNA for conversion test. To generate 222 bp model DNA containing five CpG sites, bacteriophage lambda DNA (Thermo Fisher) was PCR amplified using Taq DNA Polymerase (New England Biolabs) and purified by AMPure XP beads (Beckman Coulter). Primers sequences are as follows: FW-5'-CCTGATGAAACAAGCATGTC-3', RV-5'-CAUTACTCACUTCCCCACUT-3'. The uracil base in the reverse strand of PCR product was removed by USER enzyme (New England Biolabs). 100 ng of purified PCR product was then methylated in 20 µl solution containing 1x NEBuffer 2, 0.64 mM S-adenosylmethionine and 20 U M.SssI CpG Methyltransferase (New England Biolabs) for 2 hr at 37°C, followed by 20 min heat inactivation at 65°C. The methylated 222 bp model DNA was purified by AMPure XP beads.

[0120] Model DNA for TAPS, TAPSβ and CAPS validation with Sanger sequencing. 34 bp DNA oligo containing single 5mC and single 5hmC site was annealed with other DNA oligos in annealing buffer containing 5 mM Tris-Cl (pH 7.5), 5 mM MgCl₂, and 50 mM NaCl, and then ligated in a reaction containing 400 U T4 ligase (NEB) at 25°C for 1 hr and purified by 1.8X AMPure XP beads.

DNA	Sequence (5' to 3')
34 bp mC and hmC	CCCGA ^m CGCATGATCTGTACTTGATCGAC ^{hm} CGTGCAAC
TruSeq Universal Adapter	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGA
TruSeq Adapter (Index 6)	/5Phos/GATCGGAAGAGCACACGTCTGAACTCCAGTCACG CCAATATCTCGTATGCCGTCTTCTGCTTG
Uracil linker	TCTTCCGAUCGTTGCACGGUCGATCAAGUACAGATCAT GCGUCGGGAGAUCGGAAG

[0121] The Uracil linker was removed by USER enzyme after ligation reaction resulting in a final product sequence (5' to 3'):

AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGA TCTCCCGA^mCGCATGATCTGTACTTGATCGAC^{hm}CGTGCAACGATCGGAAGAGCA CACGTCTGAACTCCAGTCACGCCAATATCTCGTATGCCGTCTTCTGCTTG. PCR primers for amplification of the model DNA were: P5: 5'-AATGATACGGCGACCACCGAG-3' and P7: 5'-CAAGCAGAAGACGGCATAACGAG-3'.

[0122] Model DNA for polymerase test and Sanger sequencing. Model DNA for polymerase test and Sanger sequencing was prepared with the same ligation method above except different DNA oligos were used:

DNA	Sequence (5' to 3')
34 bp mC	AGCAGTCT ^m CGATCAGCTG ^m CTACTGTA ^m CGTAGCAT
TruSeq Universal Adapter	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCT
TruSeq Adapter (Index 6)	/5Phos/GATCGGAAGAGCACACGTCTGAACTCCAGTCACGC CAATATCTCGTATGCCGTCTTCTGCTTG
Insert_1_40_bp	/5Phos/AGGTGCGCTAAGTTCTAGATCGCCAACTGGTTGTG GCCTT
Insert_2_60_bp	/5Phos/CTATAGCCGGCTTGCTCTCTCTGCCTCTAGCAGCTG CTCCCTATAGTGAGTCGTATTAAC
40_bp-Linker-1	ATCTAGAACTTAGCGCACCTAGATCGGAAGAGCGTCGTG T
80_bp-Linker:	AGAGAGCAAGCCGGCTATAGATGCTACGTACAGTAGCAG CTGATCAAGACTGCTAAGGCCACAACCAAGTTGGCG
42_bp-Linker-2:	AGACGTGTGCTCTTCCGATCGTTAATACGACTCACTATAG GG

[0123] The final product sequence (5' to 3') was:

AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGA TCTAGGTGCGCTAAGTTCTAGATCGCCAACTGGTTGTGGCCTTAGCAGTCT^mCGA TCAGCTG^mCTACTGTA^mCGTAGCATCTATAGCCGGCTTGCTCTCTCTGCCTCTAGC AGCTGCTCCCTATAGTGAGTCGTATTAACGATCGGAAGAGCACACGTCTGAACTC CAGTCACGCCAATATCTCGTATGCCGTCTTCTGCTTG. PCR primers to amplify the model DNA are the P5 and P7 primers provided above. Biotin-labelled primer sequence for primer extension is biotin linked to the 5' end of the P7 primer. PCR primers for RT-PCR after T7 RNA polymerase transcription were the P5 primer and RT: 5'- TGCTAGAGGCAGAGAGAGCAAG-3'.

[0124] Model DNA for PCR bias test. Model DNA for PCR bias test was prepared with the same ligation method above except different DNA oligos were used:

DNA	Sequence (5' to 3')
17 bp X	AGCAGTCTXGATCAGCT (X= DHU or U or T or C)
17 bp No Modification	GCTACTGTACGTAGCAT
TruSeq Universal Adapter	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCT
TruSeq Adapter (Index 6)	/5Phos/GATCGGAAGAGCACACGTCTGAACTCCAGTCACGC CAATATCTCGTATGCCGTCTTCTGCTTG

Insert_1_40_bp	/5Phos/AGGTGCGCTAAGTTCTAGATCGCCAACTGGTTGTG GCCTT
Insert_2_60_bp	/5Phos/CTATAGCCGGCTTGCTCTCTCTGCCTCTAGCAGCTG CTCCCTATAGTGAGTCGTATTAAC
40_bp-Linker-1	ATCTAGAACTTAGCGCACCTAGATCGGAAGAGCGTCGTG T
80_bp-Linker	AGAGAGCAAGCCGGCTATAGATGCTACGTACAGTAGCAG CTGATCAAGACTGCTAAGGCCACAACCAGTTGGCG
42_bp-Linker-2	AGACGTGTGCTCTTCCGATCGTTAATACGACTCACTATAG GG

[0125] Final product sequence (5' to 3'):

AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGA
TCTAGGTGCGCTAAGTTCTAGATCGCCAACTGGTTGTGGCCTTAGCAGTCTXGAT
CAGCTGCTACTGTACGTAGCATCTATAGCCGGCTTGCTCTCTCTGCCTCTAGCAGC
TGCTCCCTATAGTGAGTCGTATTAACGATCGGAAGAGCACACGTCTGAACTCCAG
TCACGCCAATATCTCGTATGCCGTCTTCTGCTTG, where X= DHU or U or T or C.

PCR primer to amplify the model DNA are the P5 and P7 primers provided above.

[0126] *Preparation of methylated bacteriophage lambda genomic DNA*

[0127] 1 µg of unmethylated bacteriophage lambda DNA (Promega) was methylated in 50 µL reaction containing 0.64 mM SAM and 0.8 U/µl M.SssI enzyme in Mg²⁺-free buffer (10 mM Tris-Cl pH 8.0, 50 mM NaCl, and 10 mM EDTA) for 2 hours at 37°C. Then, 0.5 µL of M.SssI enzyme and 1 µL of SAM were added and the reaction was incubated for additional 2 hours at 37°C. Methylated DNA was subsequently purified on 1X Ampure XP beads. To assure complete methylation, the whole procedure was repeated in NEB buffer 2. DNA methylation was then validated with HpaII digestion assay. 50 ng of methylated and unmethylated DNA were digested in 10 µL reaction with 2 U of HpaII enzyme (NEB) in CutSmart buffer (NEB) for 1 h at 37°C. Digestion products were run on 1% agarose gel together with undigested lambda DNA control. Unmethylated lambda DNA was digested after the assay whereas methylated lambda DNA remained intact confirming complete and successful CpG methylation. Sequence of lambda DNA can be found in GenBank - EMBL Accession Number: [J02459](#).

[0128] *Preparation of 2 kb unmodified spike-in controls*

[0129] 2 kb spike-in controls (2kb-1, 2, 3) were PCR amplified from pNIC28-Bsa4 plasmid (Addgene, cat. no. 26103) in the reaction containing 1 ng DNA template, 0.5 µM primers, 1 U Phusion High-Fidelity DNA Polymerase (Thermo Fisher). PCR primer sequences are listed in Table 2.

Table 2. Sequences of PCR primers for spike-ins.

Primer name	Sequence (5' to 3')
2kb-3_Foward	CACAGATGTCTGCCTGTTCA
2kb-3_Reverse	AGGGTGGTGAATGTGAAACC

[0130] PCR product was purified on Zymo-Spin column. 2 kb unmodified control sequence (5' to 3'):

CACAGATGTCTGCCTGTTTCATCCGCGTCCAGCTCGTTGAGTTTCTCCAGAAGCGTT
AATGTCTGGCTTCTGATAAAGCGGGCCATGTAAAGGGCGGTTTTTCTGTTTGGT
CACTGATGCCTCCGTGTAAGGGGGATTTCTGTTTCATGGGGGTAATGATACCGATG
AAACGAGAGAGGATGCTCACGATACGGGTACTGATGATGAACATGCCCGGTTA
CTGGAACGTTGTGAGGGTAAACAACCTGGCGGTATGGATGCGGCGGGACCAGAGA
AAAATCACTCAGGGTCAATGCCAGCGCTTCGTTAATACAGATGTAGGTGTTCCAC
AGGGTAGCCAGCAGCATCCTGCGATGCAGATCCGGAACATAATGGTGCAGGGCG
CTGACTTCCGCGTTTCCAGACTTTACGAAACACGGAAACCGAAGACCATTTCATGT
TGTTGCTCAGGTCGCAGACGTTTTGCAGCAGCAGTCGCTTCACGTTGCTCGCGT
ATCGGTGATTCAATTCTGCTAACCAGTAAGGCAACCCCGCCAGCCTAGCCGGGTCC
TCAACGACAGGAGCACGATCATGCGCACCCGTGGGGCCGCCATGCCGGCGATAA
TGGCCTGCTTCTCGCCGAAACGTTTGGTGGCGGGACCAGTGACGAAGGCTTGAGC
GAGGGCGTGCAAGATTCCGAATACCGCAAGCGACAGGCCGATCATCGTCGCGCT
CCAGCGAAAGCGGTCCTCGCCGAAAATGACCCAGAGCGCTGCCGGCACCTGTCC
TACGAGTTGCATGATAAAGAAGACAGTCATAAGTGCGGCGACGATAGTCATGCC
CCGCGCCCAACCGGAAGGAGCTGACTGGGTTGAAGGCTCTCAAGGGCATCGGTGCG
AGATCCCGGTGCCTAATGAGTGAGCTAACTTACATTAATTGCGTTGCGCTCACTG
CCCCTTTCCAGTCGGGAAACCTGTCGTGCCAGCTGCATTAATGAATCGGCCAAC
GCGCGGGGAGAGGCGGTTTGCGTATTGGGCGCCAGGGTGGTTTTTCTTTTACCA
GTGAGACGGGCAACAGCTGATTGCCCTTACCCGCTGGCCCTGAGAGAGTTGCA
GCAAGCGGTCCACGCTGGTTTGCCCCAGCAGGCGAAAATCCTGTTTGATGGTGGT
TAACGGCGGGATATAACATGAGCTGTCTTCGGTATCGTCGTATCCCACTACCGAG
ATATCCGCACCAACGCGCAGCCCGGACTCGGTAATGGCGCGCATTGCGCCCAGC
GCCATCTGATCGTTGGCAACCAGCATCGCAGTGGGAACGATGCCCTCATTACAGCA
TTTGATGGTTTTGTTGAAAACCGGACATGGCACTCCAGTCGCTTCCCGTTCCGCT
ATCGGCTGAATTTGATTGCGAGTGAGATATTTATGCCAGCCAGCCAGACGCAGAC
GCGCCGAGACAGAACTTAATGGGCCCCGCTAACAGCGCGATTTGCTGGTGACCCA
ATGCGACCAGATGCTCCACGCCCAGTCGCGTACCGTCTTCATGGGAGAAAATAAT

ACTGTTGATGGGTGTCTGGTCAGAGACATCAAGAAATAACGCCGGAACATTAGT
 GCAGGCAGCTTCCACAGCAATGGCATCCTGGTCATCCAGCGGATAGTTAATGATC
 AGCCCACTGACGCGTTGCGCGAGAAGATTGTGCACCGCCGCTTTACAGGCTTCGA
 CGCCGCTTCGTTCTACCATCGACACCACACGCTGGCACCCAGTTGATCGGCGCG
 AGATTTAATCGCCGCGACAATTTGCGACGGCGCGTGCAGGGCCAGACTGGAGGT
 GGCAACGCCAATCAGCAACGACTGTTTGGCCGCCAGTTGTTGTGCCACGCGGTTG
 GGAATGTAATTCAGCTCCGCCATCGCCGCTTCCACTTTTCCCGCGTTTTTCGCAGA
 AACGTGGCTGGCCTGGTTCACCACGCGGGAAACGGTCTGATAAGAGACACCGGC
 ATACTCTGCGACATCGTATAACGTTACTGGTTTCACATTCACCACCCT

[0131] Preparation of 120mer spike-in controls

[0132] 120mer spike-in controls were produced by primer extension. Oligo sequences and primers are listed in the Table 3.

Table 3. Sequences of DNA oligos and primers used for preparation of 120mer control spike-ins.

Spike-in control	Template sequence (5' to 3')	Primer for extension (5' to 3')
120mer-1	ATACTCATCATTAAGTTCGCCCTTACCTAC CACTTCGTGTATGTAGATAGGTAGTATACA ATTGATATCGAAATGAGTACGTAGATAGTA GAAAGTAAGATGGAGGTGAGAGTGAGAGT GCGGCGTGATACTGGTCCCAG5hmCCTGA AGTTAGGCC5hmCGGGATGACTGA5hmCAG TCTCCGAGACCGACGACACAGGTCTCCCT ATAGTGAGTCGTATTATGGCGAGAGAATGA ATCTCCATC	ATACTCATCATTAAG CTTCGCCCTTACCTAC CACTTCG GATGGAGATTCATTC TCTCGCCA TAATACGACTCACTA TAGG

Briefly, for 120mer-1 spike-in, 3 μ M oligo was annealed with 10 μ M primer in the annealing buffer containing 5 mM Tris-Cl (pH 7.5), 5 mM MgCl₂, and 50 mM NaCl. For 120mer-2 spike-in, 5 μ M oligo was annealed with 7.5 μ M primer. Primer extension was performed in the NEB buffer 2 with 0.4 μ M dNTPs (120mer-1: dATP/dGTP/dTTP/dhmCTP, 120mer-2: dATP/dGTP/dTTP/dCTP) and 5 U of Klenow Polymerase (New England Biolabs) for 1 hour at 37°C. After reaction spike-in controls were purified on Zymo-Spin columns (Zymo Research). The 120mer spike-in controls were then methylated in 50 μ L reaction containing 0.64 mM SAM and 0.8 U/ μ L M.SssI enzyme in NEB buffer 2 for 2 hours at 37°C and purified with Zymo-Spin columns. All spike-in sequences used can be downloaded from <https://figshare.com/s/80c3ab713c261262494b>.

[0133] Generation of Synthetic spike-in with N5mCNN and N5hmCNN

[0134] Synthetic oligo with N5mCNN and N5hmCNN sequences was produced by annealing and extension method. Oligo sequences are listed in Table 4, below.

Table 4.

Oligo	Template sequence (5' to 3')
N5mCNN	GAAGATGCAGAAGACAGGAAGGATGAAACACTCAGGCG CACGCTGGCATN ^m CNNGACAAACCACAAGAACAGGCTAG TGAGAATGAAGGGA
N5hmCNN	CCAAGTCTGAAACCCACCAACGCCAACATCCACCACACA ACCCAAGATN ^{hm} CNNGACCATCTTACAAACATATCCCTTC ATTCTCACTAGCC

[0135] Briefly, 10 μ M N5mCNN and N5hmCNN oligos (IDT) were annealed together in the annealing buffer containing 5 mM Tris-Cl (pH 7.5), 5 mM MgCl₂, and 50 mM NaCl.

Extension was performed in the NEB buffer 2 with 0.4 mM dNTPs

(dATP/dGTP/dTTP/dCTP) and 5 U of Klenow Polymerase (NEB) for 1 hour at 37°C. After reaction, spike-in control was purified on Zymo-Spin column (Zymo Research). Synthetic spike-in with N5mCNN and N5hmCNN (5' to 3'):

GAAGATGCAGAAGACAGGAAGGATGAAACACTCAGGCGCACGCTGGCATNmCN
NGACAAACCACAAGAACAGGCTAGTGAGAATGAAGGGATATGTTTGTAAAGATGG
TCNNGNATCTTGGGTTGTGTGGTGGATGTTGGCGTTGGTGGGTTTCAGAGTTGG.

Complementary strand (5' to 3'):

CCAAGTCTGAAACCCACCAACGCCAACATCCACCACACAACCCAAGATNhmCNN
GACCATCTTACAAACATATCCCTTCATTCTCACTAGCCTGTTCTTGTGGTTTGTCTN
NGNATGCCAGCGTGCGCCTGAGTGTTTCATCCTTCCTGTCTTCTGCATCTTC.

[0136] DNA digestion and HPLC-MS/MS analysis

[0137] DNA samples were digested with 2 U of Nuclease P1 (Sigma-Aldrich) and 10 nM deaminase inhibitor erythro-9-Amino- β -hexyl- α -methyl-9H-purine-9-ethanol hydrochloride (Sigma-Aldrich). After overnight incubation at 37°C, the samples were further treated with 6 U of alkaline phosphatase (Sigma-Aldrich) and 0.5 U of phosphodiesterase I (Sigma-Aldrich) for 3 hours at 37°C. The digested DNA solution was filtered with Amicon Ultra-0.5 mL 10 K centrifugal filters (Merck Millipore) to remove the proteins, and subjected to HPLC-MS/MS analysis.

[0138] The HPLC-MS/MS analysis was carried out with 1290 Infinity LC Systems (Agilent) coupled with a 6495B Triple Quadrupole Mass Spectrometer (Agilent). A ZORBAX Eclipse Plus C18 column (2.1 x 150 mm, 1.8-Micron, Agilent) was used. The column temperature was maintained at 40°C, and the solvent system was water containing 10

mM ammonium acetate (pH 6.0, solvent A) and water-acetonitrile (60/40, v/v, solvent B) with 0.4 mL/min flow rate. The gradient was: 0-5 min; 0 solvent B; 5-8 min; 0-5.63 % solvent B; 8-9 min; 5.63 % solvent B; 9-16 min; 5.63-13.66% solvent B; 16-17 min; 13.66-100% solvent B; 17-21 min; 100% solvent B; 21-24.3 min; 100-0% solvent B; 24.3-25 min; 0% solvent B. The dynamic multiple reaction monitoring mode (dMRM) of the MS was used for quantification. The source-dependent parameters were as follows: gas temperature 230°C, gas flow 14 L/min, nebulizer 40 psi, sheath gas temperature 400°C, sheath gas flow 11 L/min, capillary voltage 1500 V in the positive ion mode, nozzle voltage 0 V, high pressure RF 110 V and low pressure RF 80 V, both in the positive ion mode. The fragmentor voltage was 380 V for all compounds, while other compound-dependent parameters were as summarized in Table 5.

Table 5. Compound-dependent HPLC-MS/MS parameters used for nucleosides quantification. RT: retention time, CE: collision energy; CAE: cell accelerator voltage. All the nucleosides were analyzed in the positive mode.

Compound	Precursor Ion (m/z)	Product Ion (m/z)	RT (min)	Delta RT(min)	CE (V)	CAE (V)
dA+H	252	136	13.78	2	10	4
dT+H	243	127	11.07	2	10	4
dT+Na	265	149	11.07	2	10	4
dG+H	268	152	9.64	2	10	4
dC+H	228	112	3.71	1.5	10	4
dC+Na	250	134	3.71	1.5	10	4
mdC+H	242	126	9.05	1.5	10	4
mdC+Na	264	148	9.05	1.5	10	4
hmdC+H	258	142	4.34	2	12	4
hmdC+Na	280	164	4.34	2	12	4
fdC+H	256	140	10.69	2	8	4
fdC+Na	278	162	10.69	2	8	4
cadC+H	272	156	1.75	3	12	4
cadC+Na	294	178	1.75	3	12	4
DHU+H	231	115	3.45	3	10	4
DHU+Na	253	137	3.45	3	10	4

[0139] Expression and purification of NgTET1

[0140] pRSET-A plasmid encoding His-tagged NgTET1 protein (GG739552.1) was designed and purchased from Invitrogen. Protein was expressed in *E. coli* BL21 (DE3) bacteria and purified as previously described with some modifications (J. E. Pais et al., Biochemical characterization of a Naegleria TET-like oxygenase and its application in single molecule sequencing of 5-methylcytosine. Proc. Natl. Acad. Sci. U.S.A. 112, 4316-4321 (2015), incorporated herein by reference). Briefly, for protein expression bacteria from overnight small-scale culture were grown in LB medium at 37°C and 200 rpm until OD600

was between 0.7-0.8. Then cultures were cooled down to room temperature and target protein expression was induced with 0.2 mM isopropyl- β -D-1-thiogalactopyranoside (IPTG). Cells were maintained for additional 18 hours at 18°C and 180 rpm. Subsequently, cells were harvested and re-suspended in the buffer containing 20 mM HEPES (pH 7.5), 500 mM NaCl, 1 mM DTT, 20 mM imidazole, 1 μ g/mL leupeptin, 1 μ g/mL pepstatin A and 1 mM PMSF. Cells were broken with EmulsiFlex-C5 high-pressure homogenizer, and lysate was clarified by centrifugation for 1 hour at 30,000 x g and 4°C. Collected supernatant was loaded on Ni-NTA resins and NgTET1 protein was eluted with buffer containing 20 mM HEPES (pH 7.5), 500 mM imidazole, 2 M NaCl, 1 mM DTT. Collected fractions were then purified on HiLoad 16/60 Sdx 75 (20 mM HEPES pH 7.5, 2 M NaCl, 1 mM DTT). Fractions containing NgTET1 were then collected, buffer exchanged to the buffer containing 20 mM HEPES (pH 7.0), 10 mM NaCl, 1 mM DTT, and loaded on HiTrap HP SP column. Pure protein was eluted with the salt gradient, collected and buffer-exchanged to the final buffer containing 20 mM Tris-Cl (pH 8.0), 150 mM NaCl and 1 mM DTT. Protein was then concentrated up to 130 μ M, mixed with glycerol (30% v/v) and aliquots were stored at -80°C.

[0141] Expression and purification of mTET1CD

[0142] mTET1CD catalytic domain (NM_001253857.2, 4371-6392) with N-terminal Flag-tag was cloned into pcDNA3-Flag between KpnI and BamHI restriction sites. For protein expression, 1 mg plasmid was transfected into 1 L of Expi293F (Gibco) cell culture at density 1×10^6 cells/mL and cells were grown for 48 h at 37°C, 170 rpm and 5% CO₂. Subsequently, cells were harvested by centrifugation, re-suspended in the lysis buffer containing 50 mM Tris-Cl pH = 7.5, 500 mM NaCl, 1X cOmplete Protease Inhibitor Cocktail (Sigma), 1 mM PMSF, 1% Triton X-100 and incubated on ice for 20 min. Cell lysate was then clarified by centrifugation for 30 min at 30000 x g and 4°C. Collected supernatant was purified on ANTI-FLAG M2 Affinity Gel (Sigma) and pure protein was eluted with buffer containing 20 mM HEPES pH = 8.0, 150 mM NaCl, 0.1 mg/mL 3X Flag peptide (Sigma), 1X cOmplete Protease Inhibitor Cocktail (Sigma), 1 mM PMSF. Collected fractions were concentrated and buffer-exchanged to the final buffer containing 20 mM HEPES pH = 8.0, 150 mM NaCl and 1 mM DTT. Concentrated protein was mixed with glycerol (30% v/v), frozen in liquid nitrogen and aliquots were stored at -80°C. Activity and quality of recombinant mTET1CD was checked by MALDI Mass Spectrometry analysis. Based on this assay, recombinant mTET1CD is fully active and able to catalyze oxidation of 5mC to 5caC. Any significant digestion of tested model oligo was detected by MALDI confirming that protein is free from nucleases.

[0143] TET Oxidation

[0144] NgTET1 Oxidation. For Tet oxidation of the 222 bp model DNA oligos, 100 ng of 222 bp DNA was incubated in 20 μ l solution containing 50 mM MOPS buffer (pH 6.9), 100 mM ammonium iron (II) sulfate, 1 mM α -ketoglutarate, 2 mM ascorbic acid, 1 mM dithiothreitol (DTT), 50 mM NaCl, and 5 μ M NgTET for 1 hr at 37 °C. After that, 0.4 U of Proteinase K (New England Biolabs) was added to the reaction mixture and incubated for 30 min at 37°C. The product was purified by Zymo-Spin column (Zymo Research) following manufacturer's instruction.

[0145] For NgTET1 oxidation of genomic DNA, 500 ng of genomic DNA were incubated in 50 μ l solution containing 50 mM MOPS buffer (pH 6.9), 100 mM ammonium iron (II) sulfate, 1 mM α -ketoglutarate, 2 mM ascorbic acid, 1 mM dithiothreitol, 50 mM NaCl, and 5 μ M NgTET1 for 1 hour at 37°C. After that, 4 U of Proteinase K (New England Biolabs) were added to the reaction mixture and incubated for 30 min at 37°C. The product was cleaned-up on 1.8X Ampure beads following the manufacturer's instruction.

[0146] mTET1 Oxidation. 100 ng of genomic DNA was incubated in 50 μ l reaction containing 50 mM HEPES buffer (pH 8.0), 100 μ M ammonium iron (II) sulfate, 1 mM α -ketoglutarate, 2 mM ascorbic acid, 1 mM dithiothreitol, 100 mM NaCl, 1.2 mM ATP and 4 μ M mTET1CD for 80 min at 37°C. After that, 0.8 U of Proteinase K (New England Biolabs) were added to the reaction mixture and incubated for 1 hour at 50°C. The product was cleaned-up on Bio-Spin P-30 Gel Column (Bio-Rad) and 1.8X Ampure XP beads following the manufacturer's instruction.

[0147] Borane Reduction

[0148] Pic-BH₃ reduction 25 μ L of 5 M α -picoline-borane (pic-BH₃, Sigma-Aldrich) in MeOH and 5 μ L of 3 M sodium acetate solution (pH 5.2, Thermo Fisher) was added into 20 μ L DNA sample and incubated at 60°C for 1 h. The product was purified by Zymo-Spin column (Zymo Research) following manufacturer's instructions for the 222 bp or by Micro Bio-Spin 6 Columns (Bio-Rad) following manufacturer's instruction for the oligos.

[0149] Alternatively, 100 mg of 2-picoline-borane (pic-borane, Sigma-Aldrich) was dissolved in 187 μ L of DMSO to give around 3.26 M solution. For each reaction, 25 μ L of pic-borane solution and 5 μ L of 3 M sodium acetate solution (pH 5.2, Thermo Fisher) were added into 20 μ L of DNA sample and incubated for 3 hours at 70°C. The product was purified by Zymo-Spin column for genomic DNA or by Micro Bio-Spin 6 Columns (Bio-Rad) for DNA oligos following the manufacturer's instructions.

[0150] Pyridine borane reduction. 50-100 ng of oxidised DNA in 35 μ L of water were reduced in 50 μ L reaction containing 600 mM sodium acetate solution (pH = 4.3) and 1 M pyridine borane for 16 hours at 37°C and 850 rpm in Eppendorf ThermoMixer. The product was purified by Zymo-Spin column.

[0151] Single nucleoside pic-borane reaction. 500 μ L of 3.26 M 2-picoline-borane (pic-borane, Sigma-Aldrich) in MeOH and 500 μ L of 3 M sodium acetate solution (pH 5.2, Thermo Fisher) were added into 10 mg of 2'-deoxycytidine-5-carboxylic acid sodium salt (Berry&Associates). The mixture was stirred for 1 hour at 60°C. The product was purified by HPLC to give pure compound as white foam. High resolution MS (Q-TOF) m/z [M + Na]⁺ calculated for C₉H₁₄N₂O₅Na: 253.0800; found: 253.0789.

[0152] 5hmC blocking

[0153] 5hmC blocking was performed in 20 μ L solution containing 50 mM HEPES buffer (pH 8), 25 mM MgCl₂, 200 μ M uridine diphosphoglucose (UDP-Glc, New England Biolabs), and 10 U β GT (Thermo Fisher), and 10 μ M 5hmC DNA oligo for 1 hr at 37 °C. The product was purified by Micro Bio-Spin 6 Columns (Bio-Rad) following manufacturer's instruction.

[0154] 5fC blocking

[0155] 5fC blocking was performed in 100 mM MES buffer (pH 5.0), 10 mM O-ethylhydroxylamine (Sigma- Aldrich), and 10 μ M 5fC DNA oligo for 2 hours at 37 °C. The product was purified by Micro Bio-Spin 6 Columns (Bio-Rad) following manufacturer's instruction.

[0156] 5caC blocking

[0157] 5caC blocking was performed in 75 mM MES buffer (pH 5.0), 20 mM N-hydroxysuccinimide (NHS, Sigma-Aldrich), 20 mM 1-(3-dimethylaminopropyl)-3-ethylcarbodiimide hydrochloride (EDC, Fluorochem), and 10 μ M 5caC DNA oligo at 37 °C for 0.5 h. The buffer was then exchanged to 100 mM sodium phosphate (pH 7.5), 150 mM NaCl using Micro Bio-Spin 6 Columns (Bio-Rad) following manufacturer's instructions. 10 mM ethylamine (Sigma-Aldrich) was added to the oligo and incubated for 1 hour at 37°C. The product was purified by Micro Bio-Spin 6 Columns (Bio-Rad) following manufacturer's instructions.

[0158] 5hmC oxidation

[0159] 46 μ L of 5hmC DNA oligo was denatured with 2.5 μ L of 1 M NaOH for 30 min at 37°C in a shaking incubator, then oxidized with 1.5 μ L of solution containing 50 mM NaOH and 15 mM potassium perruthenate (KRuO₄, Sigma-Aldrich) for 1 hour on ice. The product was purified by Micro Bio-Spin 6 Columns following manufacturer's instructions.

[0160] *Validation of TAPS conversion with TaqαI assay*

[0161] 5mC conversion after TAPS was tested by PCR amplification of a target region which contains TaqαI restriction site (TCGA) and subsequent TaqαI digestion. For example, 5mC conversion in our TAPS libraries can be tested based on 194 bp amplicon containing single TaqαI restriction site that is amplified from CpG methylated lambda DNA spike-in control. PCR product amplified from the 194 bp amplicon is digested with TaqαI restriction enzyme and digestion product is checked on 2% agarose gel. PCR product amplified on unconverted control DNA is digested by TaqαI and shows two bands on the gel. In TAPS-converted sample restriction site is lost due to C-to-T transition, so the 194 bp amplicon would remain intact. Overall conversion level can be assessed based on digested and undigested gel bands quantification and for successful TAPS samples should be higher than 95%.

[0162] Briefly, the converted DNA sample was PCR amplified by Taq DNA Polymerase (New England Biolabs) with corresponding primers. The PCR product was incubated with 4 units of TaqαI restriction enzyme (New England Biolabs) in 1X CutSmart buffer (New England Biolabs) for 30 min at 65°C and checked by 2% agarose gel electrophoresis.

[0163] *Quantitative polymerase chain reaction (qPCR)*

[0164] For comparison of amplification curves and melting curves between model DNAs before and after TAPS (Fig. 11), 1 ng of DNA sample was added into 19 μL of PCR master mix containing 1× LightCycler 480 High Resolution Melting Master Mix (Roche Diagnostics Corporation), 250 nM of primers FW-CCTGATGAAACAAGCATGTC and RV-CATTACTCACTTCCCCACTT and 3 mM of MgSO₄. For PCR amplification, an initial denaturation step was performed for 10 min at 95°C, followed by 40 cycles of 5 sec denaturation at 95°C, 5 sec annealing at customized annealing temperature and 5 sec elongation at 72°C. The final step included 1 min at 95°C, 1 min at 70°C and a melting curve (0.02°C step increments, 5 sec hold before each acquisition) from 65°C to 95°C.

[0165] For other assays, qPCR was performed by adding the required amount of DNA sample into 19 μL of PCR master mix containing 1× Fast SYBR Green Master Mix (Thermo Fisher), 200 nM of forward and reverse primers. For PCR amplification, an initial denaturation step was performed for 20 sec at 95°C, followed by 40 cycles of 3 s denaturation at 95°C, 20 s annealing and elongation at 60°C.

[0166] *Validation of C^m CGG methylation level in mESC gDNA with HpaII-qPCR assay.*

[0167] 1 µg mESC gDNA was incubated with 50 units of HpaII (NEB, 50 units/µL) and 1X CutSmart buffer in 50 µL reaction for 16 hours at 37°C. No HpaII was added for control reaction. 1 µL Proteinase K was added to the reaction and incubated at 40°C for 30 minutes followed by inactivation of Proteinase K for 10 minutes at 95°C. Ct value of HpaII digested sample or control sample was measured by qPCR assay as above with corresponding primer sets for specific CCGG positions (listed in Table 9).

[0168] ***Sanger sequencing***

[0169] The PCR product was purified by Exonuclease I and Shrimp Alkaline Phosphatase (New England Biolabs) or Zymo-Spin column and processed for Sanger sequencing.

[0170] ***DNA damage test on fragments with different length.***

[0171] mESC genomic DNA was spiked-in with 0.5% of CpG methylated lambda DNA and left unfragmented or sonicated with Covaris M220 instrument and size-selected to 500-1 kb or 1 kb–3 kb on Ampure XP beads. 200 ng of DNA were single-oxidised with mTET1CD and reduced with Pyridine borane complex as described above or converted with EpiTect Bisulfite Kit (Qiagen) according to manufacturer's protocol. 10 ng of DNA before and after TAPS and Bisulfite conversion were run on 1% agarose gel. To visualize bisulfite converted gel was cooled down for 10 min samples in ice bath. 5mC conversion in TAPS samples was tested by TaqαI digestion assay as described above.

[0172] ***mESCs culture and isolation of genomic DNA***

[0173] Mouse ESCs (mESCs) E14 were cultured on gelatin-coated plates in Dulbecco's Modified Eagle Medium (DMEM) (Invitrogen) supplemented with 15% FBS (Gibco), 2 mM L-glutamine (Gibco), 1% non-essential amino acids (Gibco), 1% penicillin/streptavidin (Gibco), 0.1 mM β-mercaptoethanol (Sigma), 1000 units/mL LIF (Millipore), 1 µM PD0325901 (Stemgent), and 3 µM CHIR99021 (Stemgent). Cultures were maintained at 37°C and 5% CO₂ and passaged every 2 days.

[0174] For isolation of genomic DNA, cells were harvested by centrifugation for 5 min at 1000 x g and room temperature. DNA was extracted with Quick-DNA Plus kit (Zymo Research) according to manufacturer's protocol.

[0175] ***Preparation of mESC gDNA for TAPS and WGBS.***

[0176] For whole-genome bisulfite sequencing (WGBS), mESC gDNA was spiked-in with 0.5% of unmethylated lambda DNA. For whole-genome TAPS, mESC gDNA was spiked-in with 0.5% of methylated lambda DNA and 0.025% of unmodified 2 kb spike-in control. DNA samples were fragmented by Covaris M220 instrument and size-selected to

200-400 bp on Ampure XP beads. DNA for TAPS was additionally spiked-in with 0.25% of N5mCNN and N5hmCNN control oligo after size-selection with Ampure XP beads.

[0177] *Whole Genome Bisulfite Sequencing*

[0178] For Whole Genome Bisulfite Sequencing (WGBS), 200 ng of fragmented mESC gDNA spiked-in with 0.5% of unmethylated bacteriophage lambda DNA was used. End-repaired and A-tailing reaction and ligation of methylated adapter (NextFlex) were prepared with KAPA HyperPlus kit (Kapa Biosystems) according to manufacturer's protocol. Subsequently, DNA underwent bisulfite conversion with EpiTect Bisulfite Kit (Qiagen) according to Illumina's protocol. Final library was amplified with KAPA Hifi Uracil Plus Polymerase (Kapa Biosystems) for 6 cycles and cleaned-up on 1X Ampure beads. WGBS sequencing library was paired-end 80 bp sequenced on a NextSeq 500 sequencer (Illumina) using a NextSeq High Output kit with 15% PhiX control library spike-in.

[0179] *Whole-genome TAPS*

[0180] For whole genome TAPS, 100 ng of fragmented mESC gDNA spiked-in with 0.5% of methylated lambda DNA and 0.025% of unmodified 2 kb spike-in control were used. End-repair and A-tailing reaction and ligation of Illumina Multiplexing adapters were prepared with KAPA HyperPlus kit according to manufacturer's protocol. Ligated DNA was oxidized with mTET1CD twice and then reduced with pyridine borane according to the protocols described above. Final sequencing library was amplified with KAPA Hifi Uracil Plus Polymerase for 5 cycles and cleaned-up on 1X Ampure beads. Whole-genome TAPS sequencing library was paired-end 80 bp sequenced on a NextSeq 500 sequencer (Illumina) using one NextSeq High Output kit with 1% PhiX control library spike-in.

[0181] *Low-input whole-genome TAPS with dsDNA library preparation kits*

[0182] mESC gDNA prepared as described above for whole-genome TAPS was used for low-input whole-genome TAPS. Briefly, samples containing 100 ng, 10 ng, and 1 ng of mESC gDNA were oxidized with NgTET1 once according to the protocol described above. End-repaired and A-tailing reaction and ligation were performed with NEBNext Ultra II (New England Biolabs) or KAPA HyperPlus kit according to manufacturer's protocol. Subsequently DNA underwent pic-borane reaction as described above. Converted libraries were amplified with KAPA Hifi Uracil Plus Polymerase and cleaned-up on 1X Ampure beads.

[0183] *Low-input whole-genome TAPS with ssDNA library preparation kit*

[0184] mESC gDNA prepared as described above for whole-genome TAPS was used for low-input whole-genome TAPS. Briefly, samples containing 100 ng, 10 ng, 1 ng, 100 pg,

and 10 pg of mESC gDNA were oxidized with NgTET1 once and reduced with pic-borane as described above. Sequencing libraries were prepared with Accel-NGS Methyl-Seq DNA Library Kit (Swift Biosciences) according to manufacturer's protocol. Final libraries were amplified with KAPA Hifi Uracil Plus Polymerase for 6 cycles (100 ng), 9 cycles (10 ng), 13 cycles (1 ng), 16 cycles (100 pg), and 21 cycles (10 pg) and cleaned-up on 0.85X Ampure beads.

[0185] In other experiments, mESC gDNA prepared as described above for whole-genome TAPS were used for low-input whole-genome TAPS. Briefly, samples containing 100 ng, 10 ng, and 1 ng of mESC gDNA were used for End-repaired and A-tailing reaction and ligated to Illumina Multiplexing adaptors with KAPA HyperPlus kit according to manufacturer's protocol. Ligated samples were then oxidized with mTET1CD once and then reduced with pyridine borane according to the protocols described above. Converted libraries were amplified with KAPA Hifi Uracil Plus Polymerase for 5 cycles (100 ng), 8 cycles (10 ng), and 13 cycles (1 ng) and cleaned-up on 1X Ampure XP beads.

[0186] *Cell-free DNA TAPS*

[0187] Cell-free DNA TAPS samples were prepared from 10 ng and 1 ng of cell-free DNA sample. Briefly, samples were oxidized with NgTET1 once and reduced with pic-borane as described above. Sequencing libraries were prepared with Accel-NGS Methyl-Seq DNA Library Kit (Swift Biosciences) according to manufacturer's protocol. Final libraries were amplified with KAPA Hifi Uracil Plus Polymerase for 9 cycles (10 ng) and 13 cycles (1 ng) and cleaned-up on 0.85X Ampure beads.

[0188] In other experiments, cell-free DNA TAPS samples were prepared from 10 ng and 1 ng of cell-free DNA sample as described above for whole-genome TAPS. Briefly, cell-free DNA samples were used for End-repaired and A-tailing reaction and ligated to Illumina Multiplexing adaptors with KAPA HyperPlus kit according to manufacturer's protocol. Ligated samples were then oxidized with mTET1CD once and then reduced with pyridine borane according to the protocols described above. Converted libraries were amplified with KAPA Hifi Uracil Plus Polymerase for 7 cycles (10 ng), and 13 cycles (1 ng) and cleaned-up on 1X Ampure XP beads.

[0189] *WGBS data processing*

[0190] Paired-end reads were download as FASTQ from Illumina BaseSpace and subsequently quality-trimmed with Trim Galore! v0.4.4

(https://www.bioinformatics.babraham.ac.uk/projects/trim_galore/). Read pairs where at least one read was shorter than 35 bp after trimming were removed. Trimmed reads were

mapped to a genome combining the mm9 version of the mouse genome, lambda phage and PhiX (sequence from Illumina iGENOMES) using Bismark v0.19 using --no_overlap option (F. Krueger, S. R. Andrews, Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications. *Bioinformatics* **27**, 1571-1572 (2011), incorporated herein by reference). The 'three-C' filter was used to remove reads with excessive non-conversion rates. PCR duplicates were called using Picard v1.119 (<http://broadinstitute.github.io/picard/>) MarkDuplicates. Regions known to be prone to mapping artefacts were downloaded (<https://sites.google.com/site/anshulkundaje/projects/blacklists>) and excluded from further analysis (E. P. Consortium, An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57-74 (2012), incorporated herein by reference).

[0191] TAPS data pre-processing

[0192] Paired-end reads were downloaded from Illumina BaseSpace and subsequently quality-trimmed with Trim Galore! v0.4.4. Read pairs where at least one read was shorter than 35 bp after trimming were removed. Trimmed reads were mapped to a genome combining spike-in sequences, lambda phage and the mm9 version of the mouse genome using BWA mem v.0.7.15 (H. Li, R. Durbin, Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754-1760 (2009), incorporated herein by reference) with default parameters. Regions known to be prone to mapping artefacts were downloaded (<https://sites.google.com/site/anshulkundaje/projects/blacklists>) and excluded from further analysis (E. P. Consortium, *Nature* **489**, 57-74 (2012)).

[0193] Detection of converted bases in TAPS

[0194] Aligned reads were split into original top (OT) and original bottom (OB) strands using a custom python3 script (MF-filter.py). PCR duplicates were then removed with Picard MarkDuplicates on OT and OB separately. Overlapping segments in read pairs were removed using BamUtil clipOverlap (<https://github.com/statgen/bamUtil>) on the deduplicated, mapped OT and OB reads separately. Modified bases were then detected using samtools mpileup and a custom python3 script (MF-caller_MOD.py).

[0195] Sequencing quality analysis of TAPS and WGBS

[0196] Quality score statistics per nucleotide type were extracted from original FASTQ files as downloaded from Illumina BaseSpace with a python3 script (MF-phredder.py).

[0197] Coverage analysis of TAPS and WGBS

[0198] Per-base genome coverage files were generated with Bedtools v2.25 genomecov (A. R. Quinlan, I. M. Hall, BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841-842 (2010), incorporated herein by reference). To compare

the relative coverage distributions between TAPS and WGBS, TAPS reads were subsampled to the corresponding coverage median in WGBS using the `-s` option of `samtools view`. In the analyses comparing coverage in WGBS and subsampled TAPS, `clipOverlap` was used on both TAPS and WGBS bam files.

[0199] *Analysis of cytosine modifications measured by TAPS and WGBS*

[0200] The fraction of modified reads per base was calculated from Bismark output, and the output of `MF-caller_MOD.py`, respectively. Intersections were performed using `Bedtools intersect`, and statistical analyses and figures were generated in R and Matlab. Genomic regions were visualized using IGV v2.4.6 (J. T. Robinson et al., Integrative genomics viewer. *Nat. Biotechnol.* **29**, 24-26 (2011), incorporated herein by reference). To plot the coverage and modification levels around CGIs, all CGI coordinates for mm9 were downloaded from the UCSC genome browser, binned into 20 windows, and extended by up to 50 windows of size 80 bp on both sides (as long as they did not reach half the distance to the next CGI). Average modification levels (in CpGs) and coverage (in all bases, both strands) in each bin were computed using `Bedtools map`. The values for each bin were again averaged and subsequently plotted in Matlab.

[0201] *Data processing time simulation*

[0202] Synthetic pair-end sequencing reads were simulated using ART42 based on the lambda phage genome (with parameters `-p -ss NS50 --errfree --minQ 15 -k 0 -nf 0 -l 75 -c 1000000 -m 240 -s 0 -ir 0 -ir2 0 -dr 0 -dr2 0 -sam -rs 10`). 50% of all CpG positions were subsequently marked as modified and two libraries were produced, either as TAPS (convert modified bases) or as WGBS (convert unmodified bases), using a custom python3 script. The reads were then processed following the pipeline used for each of the methods in the paper. Processing time was measured with Linux command `time`. All steps of the analysis were performed in single-threaded mode on one Intel Xeon CPU with 250GB of memory.

[0203] Results and Discussion

[0204] It was discovered that pic-BH_3 can readily convert 5fC and 5caC to DHU by a previously unknown reductive decarboxylation/deamination reaction (Fig. 4). The reaction was shown to be quantitative both in single nucleoside and in oligonucleotides using MALDI (Figs. 2-3, and 6-7).

[0205] An 11mer 5caC-containing DNA oligo was used as a model to screen chemicals that could react with 5caC, as monitored by matrix-assisted laser desorption/ionization mass spectroscopy (MALDI). Certain borane-containing compounds were found to efficiently react with the 5caC oligo, resulting in a molecular weight reduction of 41 Da (Figs. 1 and 2).

Pyridine borane and its derivative 2-picoline borane (pic-borane) were selected for further study as they are commercially available and environmentally benign reducing agents.

[0206] The reaction on a single 5caC nucleoside was repeated and confirmed that pyridine borane and pic-borane convert 5caC to dihydrouracil (DHU) (Figs. 3, 4B). Interestingly, pyridine borane and pic-borane was found to also convert 5fC to DHU through an apparent reductive decarboxylation/ deamination mechanism (Figs. 4C and 6). The detailed mechanism of both reactions remains to be defined. Quantitative analysis of the borane reaction on the DNA oligo by HPLC-MS/MS confirms that pic-borane converts 5caC and 5fC to DHU with around 98% efficiency and has no activity against unmethylated cytosine, 5mC or 5hmC (Fig. 2B).

[0207] As a uracil derivative, DHU can be recognized by both DNA and RNA polymerases as thymine. Therefore, borane reduction can be used to induce both 5caC-to-T and 5fC-to-T transitions, and can be used for base-resolution sequencing of 5fC and 5caC, which we termed Pyridine borane Sequencing (“PS”) (Table 6). The borane reduction of 5fC and 5caC to T can be blocked through hydroxylamine conjugation (C. X. Song et al., Genome-wide profiling of 5-formylcytosine reveals its roles in epigenetic priming. *Cell* **153**, 678-691 (2013), incorporated herein by reference) and EDC coupling (X. Lu et al., Chemical modification-assisted bisulfite sequencing (CAB-Seq) for 5-carboxylcytosine detection in DNA. *J. Am. Chem. Soc.* **135**, 9315-9317 (2013), incorporated herein by reference), respectively (Fig. 6). This blocking allows PS to be used to sequence 5fC or 5caC specifically (Table 6).

Table 6. Comparison of BS and related methods versus PS for 5fC and 5caC sequencing.

Base	BS	fCAB-Seq /redBS-Seq	caCAB-Seq	fC-CET	PS	PS with 5fC blocking	PS with 5caC blocking
C	T	T	T	C	C	C	C
5mC	C	C	C	C	C	C	C
5hmC	C	C	C	C	C	C	C
5fC	T	C	T	T	T	C	T
5caC	T	T	C	C	T	T	C

[0208] Furthermore, TET enzymes can be used to oxidize 5mC and 5hmC to 5caC, and then subject 5caC to borane reduction in a process herein called TET-Assisted Pyridine borane Sequencing (“TAPS”) (Fig. 5A-B, Table 1). TAPS can induce a C-to-T transition of 5mC and 5hmC, and therefore can be used for base-resolution detection of 5mC and 5hmC.

[0209] In addition, β -glucosyltransferase (β GT) can label 5hmC with glucose and thereby protect it from TET oxidation (M. Yu et al., Base-resolution analysis of 5-hydroxymethylcytosine in the mammalian genome. *Cell* **149**, 1368-1380 (2012)) and borane reduction (Fig. 7), enabling the selective sequencing of only 5mC, in a process referred to herein as TAPS β (Fig. 5B, Table 1). 5hmC sites can then be deduced by subtraction of TAPS β from TAPS measurements. Alternatively, potassium perruthenate (KRuO₄), a reagent previously used in oxidative bisulfite sequencing (oxBS) (M. J. Booth et al., Quantitative Sequencing of 5-Methylcytosine and 5-Hydroxymethylcytosine at Single-Base Resolution. *Science* **336**, 934-937 (2012)), can be used to replace TET as a chemical oxidant to specifically oxidize 5hmC to 5fC (Fig. 7). This approach, referred to herein as Chemical-Assisted Pyridine borane Sequencing ("CAPS"), can be used to sequence 5hmC specifically (Fig. 5B, Table 1). Therefore, TAPS and related methods can in principle offer a comprehensive suite to sequence all four cytosine epigenetic modifications (Fig. 5B, Table 1, Table 6).

[0210] TAPS alone will detect the existing 5fC and 5caC in the genome as well. However, given the extremely low levels of 5fC and 5caC in genomic DNA under normal conditions, this will be acceptable. If under certain conditions, one would like to eliminate the 5fC and 5caC signals completely, it can also be readily accomplished by protecting the 5fC and 5caC by hydroxylamine conjugation and EDC coupling, respectively, thereby preventing conversion to DHU.

[0211] The performance of TAPS was evaluated in comparison with bisulfite sequencing, the current standard and most widely used method for base-resolution mapping of 5mC and 5hmC. *Naegleria* TET-like oxygenase (NgTET1) and mouse Tet1 (mTet1) were used because both can efficiently oxidize 5mC to 5caC *in vitro*. To confirm the 5mC-to-T transition, TAPS was applied to model DNA containing fully methylated CpG sites and showed that it can effectively convert 5mC to T, as demonstrated by restriction enzyme digestion (Fig. 8A-B) and Sanger sequencing (Fig. 9A). TAPS β and CAPS were also validated by Sanger sequencing (Fig. 12).

[0212] TAPS was also applied to genomic DNA (gDNA) from mouse embryonic stem cells (mESCs). HPLC-MS/MS quantification showed that, as expected, 5mC accounts for 98.5% of cytosine modifications in the mESCs gDNA; the remainder is composed of 5hmC (1.5%) and trace amounts 5fC and 5caC, and no DHU (Fig. 9B). After TET oxidation, about 96% of cytosine modifications were oxidized to 5caC and 3% were oxidized to 5fC (Fig. 9B). After borane reduction, over 99% of the cytosine modifications were converted into DHU

(Fig. 9B). These results demonstrate both TET oxidation and borane reduction work efficiently on genomic DNA.

[0213] Both TET oxidation and borane reduction are mild reactions, with no notable DNA degradation compared to bisulfite (Fig. 10A-D) and thereby provide high DNA recovery. Another notable advantage over bisulfite sequencing is that TAPS is non-destructive and can preserve DNA up to 10 kbs long (Fig. 10C). Moreover, DNA remains double stranded after TAPS (Fig. 10A-C), and the conversion is independent of the DNA length (Fig. 15A-B).

[0214] In addition, because DHU is close to a natural base, it is compatible with various DNA polymerases and isothermal DNA or RNA polymerases (Figs. 13A-B) and does not show a bias compared to T/C during PCR (Fig. 14).

[0215] Whole genome sequencing was performed on two samples of mESC gDNA, one converted using TAPS and the other using standard whole-genome bisulfite sequencing (WGBS) for comparison.

[0216] To assess the accuracy of TAPS, spike-ins of different lengths were added that were either fully unmodified, in vitro methylated using CpG Methyltransferase (M.SssI) or GpC Methyltransferase (M.CviPI) (using the above methods). For short spike-ins (120mer-1 and 120mer-2) containing 5mC and 5hmC, near complete conversion was observed for both modifications on both strands in both CpG and non-CpG contexts (Fig. 17A-B).

[0217] 100 ng gDNA was used for TAPS, compared to 200 ng gDNA for WGBS. To assess the accuracy of TAPS, we added three different types of spike-in controls. Lambda DNA where all CpGs were fully methylated was used to estimate the false negative rate (non-conversion rate of 5mC); a 2 kb unmodified amplicon was used to estimate the false positive rate (conversion rate of unmodified C); synthetic oligo spike-ins containing both a methylated and hydroxymethylated C surrounded by any other base (N5mCNN and N5hmCNN, respectively) were used to compare the conversion rate on 5mC and 5hmC in different sequence contexts. The combination of mTet1 and pyridine borane achieved the highest 5mC conversion rate (96.5% and 97.3% in lambda and synthetic spike-ins, respectively) and the lowest conversion rate of unmodified C (0.23%) (Fig. 18A-B and Fig. 16). A false negative rate between 2.7% and 3.5%, with a false-positive rate of only 0.23%, is comparable to bisulfite sequencing: a recent study showed 9 commercial bisulfite kits had average false negative and false positive rates of 1.7% and 0.6%, respectively (Holmes, E.E. et al. Performance evaluation of kits for bisulfite-conversion of DNA from tissues, cell lines, FFPE tissues, aspirates, lavages, effusions, plasma, serum, and urine. PLoS One 9, e93933 (2014)). The synthetic spike-ins suggest that TAPS works well on both 5mC and 5hmC, and that

TAPS performs only slightly worse in non-CpG contexts. The conversion for 5hmC is 8.2% lower than 5mC, and the conversion for non-CpG contexts is 11.4% lower than for CpG contexts (Fig. 18A).

[0218] WGBS data requires special software both for the alignment and modification-calling steps. In contrast, our processing pipeline uses a standard genomic aligner (bwa), followed by a custom modification-calling tool that we call “asTair”. When processing simulated WGBS and TAPS reads (derived from the same semi-methylated source sequence), TAPS/asTair was more than 3x faster than WGBS/Bismark (Fig. 18C).

[0219] Due to the conversion of nearly all cytosine to thymine, WGBS libraries feature an extremely skewed nucleotide composition which can negatively affect Illumina sequencing. Consequently, WGBS reads showed substantially lower sequencing quality scores at cytosine/guanine base pairs compared to TAPS (Fig. 18E). To compensate for the nucleotide composition bias, at least 10 to 20% PhiX DNA (a base-balanced control library) is commonly added to WGBS libraries (*see, e.g.*, Illumina’s Whole-Genome Bisulfite Sequencing on the HiSeq 3000/HiSeq 4000 Systems). Accordingly, we supplemented the WGBS library with 15% PhiX. This, in combination with the reduced information content of BS-converted reads, and DNA degradation as a result of bisulfite treatment, resulted in significantly lower mapping rates for WGBS compared to TAPS (Fig. 18D and Table 7).

Table 7. Mapping and sequencing quality statistics for WGBS and TAPS.

Measure	WGBS	TAPS
Total raw reads	376062375	455548210
Trimmed reads	367860813	453028186
Mapped reads (mm9+spike-ins+PhiX)	251940139	451077132
PCR deduplicated reads	232303596	398127851
Mapping rate (mapped reads/trimmed reads)	68.49%	99.57%
Unique mapping rate (unique reads [MAPQ>0 for TAPS]/trimmed reads)	68.49%	88.08%
Unique PCR deduplicated mapping rate (unique PCR deduplicated reads [MAPQ>0 for TAPS] /trimmed reads)	63.15%	81.31%

[0220] Therefore, for the same sequencing cost (one NextSeq High Output run), the average depth of TAPS exceeded that of WGBS (21× and 13.1×, respectively; Table 8). Furthermore, TAPS resulted in fewer uncovered regions, and overall showed a more even coverage distribution, even after down-sampling to the same sequencing depth as WGBS (inter-quartile range: 9 and 11, respectively; Fig. 19A and Table 8).

Table 8. Coverage statistics for TAPS, WGBS and TAPS down-sampled to have approximately the same mean coverage as WGBS. Here, coverage was computed for both strands at all positions in the genome.

Measure	WGBS	TAPS with down-sampling	TAPS without down-sampling
Mean	13.078	12.411	21.001
Variance	1988.242	482.242	1371.912
median	13	13	22
qtl25	7	8	15
qtl75	18	17	28
iqr	11	9	13
maximum	116084	37329	63526

For example, CpG Islands (CGIs) in particular were generally better covered by TAPS, even when controlling for differences in sequencing depth between WGBS and TAPS (Fig. 21A), while both showed equivalent demethylation inside CGIs (Fig. 22). Moreover, WGBS showed a slight bias of decreased modification levels in highly covered CpG sites (Fig. 23A), while our results suggest that TAPS exhibits very little of the modification-coverage bias (Fig. 23B). These results demonstrate that TAPS dramatically improved sequencing quality compared to WGBS, while effectively halving the sequencing cost.

[0221] The higher and more even genome coverage of TAPS resulted in a larger number of CpG sites covered by at least three reads. With TAPS, 88.3% of all 43,205,316 CpG sites in the mouse genome were covered at this level, compared to only 77.5% with WGBS (Fig. 21B and 19B). TAPS and WGBS resulted in highly correlated methylation measurements across chromosomal regions (Fig. 21D and Fig. 20). On a per-nucleotide basis, 32,755,271 CpG positions were covered by at least three reads in both methods (Fig. 21B). Within these sites, we defined “modified CpGs” as all CpG positions with a modification level of at least 10% (L. Wen *et al.*, Whole-genome analysis of 5-hydroxymethylcytosine and 5-methylcytosine at base resolution in the human brain. *Genome Biology* **15**, R49 (2014)). Using this threshold, 95.8% of CpGs showed matching modification states between TAPS and WGBS. 98.5% of all CpGs that were covered by at least three reads and found modified in WGBS were recalled as modified by TAPS, indicating good agreement between WGBS and TAPS (Fig. 21C). When comparing modification levels per each CpG covered by at least three reads in both WGBS and TAPS, good correlation between TAPS and WGBS was observed (Pearson $r = 0.63$, $p < 2e-16$, Fig. 21E). Notably, TAPS identified a subset of highly modified CpG positions which were missed by WGBS (Fig. 21E, bottom right corner). We further validated 7 of these CpGs, using an orthogonal restriction digestion and real-time

PCR assay, and confirmed all of them are fully methylated and/or hydroxymethylated (Table 9).

Table 9. Comparison of CmCGG methylation level in mESC gDNA quantified by TAPS, WGBS and HpaII-qPCR assay. Coverage and methylation level ($mC\%$) by TAPS and WGBS were computed for per strand. Ct value for HpaII digested sample (Ct_{HpaII}) or control sample (Ct_{Ctrl}) in the HpaII-qPCR assay was average of triplicates. $mC\%$ is calculated using following equation: $mC\% = 2^{-(Ct_{Ctrl} - Ct_{HpaII})} * 100\%$.

Position of CmCGG	TAPS		WGBS		HpaII-qPCR assay			Forward and reverse primer (5'-3')
	Coverage	$mC\%$	Coverage	$mC\%$	Ct_{HpaII}	Ct_{Ctrl}	$mC\%$	
chr6: 135868201	17	100%	11	0%	29.628	29.642	101.0%	GCTGCAGATTGGAGCC AAAG TTGATGGTGATGGTGG AGCC
chr3: 31339449	15	100%	10	0%	22.162	22.111	96.5%	TCAGTGCTCATGGACTC ATACT ATACCCTGGGAGCAAA GTTGTG
chr4: 128271030	12	100%	10	0%	31.304	31.279	98.3%	CCCACTAGACATGCTCT GCC CAAAATGTTGCTTGCCCT TCCG
chr1: 58635199	11	100%	8	0%	22.008	22.026	101.3%	TCCCTGAGCCCTGATCT AGT AATACTGGCTGACCGG TTCT
chr14: 36331351	11	100%	14	0%	21.228	21.053	88.6%	ACACCACAGCAGAAGA GAGC TAGGATTGTTGCACAG GCCA
chr19: 42893499	11	100%	18	0%	22.515	22.558	103.0%	GCTGAGCTGTATCCTTG AGGT ACACGTGGGTATTCCA CAGC
chr3: 113611193	10	100%	5	0%	22.439	22.545	107.6%	GTGGATCTTCAGTGGTG GCA ATGCTCCCTCATCCTTT GCA
Negative CCGG site								
chr19: 9043049	25	0%	17	0%	27.11	21.409	1.9%	AGCCTCTGAACTTGACT GCC GCCTGGAACCTCTGAC AGTC
Positive CCGG site								
chr15: 39335961	16	100%	4	100%	22.163	22.248	106.1%	GGTCCTTGATCCACCCA GAC ACATGGTGCTGGTCTA ACCG

Together, these results indicate that TAPS can directly replace WGBS, and in fact provides a more comprehensive view of the methylome than WGBS.

[0222] Finally, TAPS was tested with low input DNA and TAPS was shown to work with as little as 1 ng gDNA and in some instances down to 10 pg of gDNA, close to single-cell level. TAPS also works effectively with down to 1 ng of circulating cell-free DNA. These results demonstrate the potential of TAPS for low input DNA and clinical applications (Fig. 24A-C, Fig. 25A-B).

[0223] TAPS was tested on three circulating cell-free DNA samples (cfDNA) from one healthy sample, one Barrett's oesophagus (Barrett's) sample, and one pancreatic cancer sample that were obtained from 1-2 ml of plasma. Standard TAPS protocol was followed and each sample sequenced to ~10x coverage. Analysis of the cfDNA TAPS results showed that TAPS provided the same high-quality methylome sequencing from low-input cfDNA as from bulk genomic DNA, including high 5mC conversion rate (Fig. 26A), low false positive rate (conversion of unmodified cytosine, Fig. 26B), high mapping rate (Fig. 26C), and low PCR duplication rate (Fig. 26D). These results demonstrate the power of TAPS for disease diagnostics from cfDNA.

[0224] TAPS can also differentiate methylation from C-to-T genetic variants or single nucleotide polymorphisms (SNPs), therefore could detect genetic variants. Methylations and C-to-T SNPs result in different patterns in TAPS: methylations result in T/G reads in original top strand (OT)/original bottom strand (OB) and A/C reads in strands complementary to OT (CTOT) and OB (CTOB), whereas C-to-T SNPs result in T/A reads in OT/OB and (CTOB/CTOT) (Fig. 27). This further increases the utility of TAPS in providing both methylation information and genetic variants, and therefore mutations, in one experiment and sequencing run. This ability of the TAPS method disclosed herein provides integration of genomic analysis with epigenetic analysis, and a substantial reduction of sequencing cost by eliminating the need to perform standard whole genome sequencing (WGS).

[0225] In summary, we have developed a series of PS-derived bisulfite-free, base-resolution sequencing methods for cytosine epigenetic modifications and demonstrated the utility of TAPS for whole-methylome sequencing. By using mild enzymatic and chemical reactions to detect 5mC and 5hmC directly at base-resolution with high sensitivity and specificity without affecting unmodified cytosines, TAPS outperforms bisulfite sequencing in providing a high quality and more complete methylome at half the sequencing cost. As such TAPS could replace bisulfite sequencing as the new standard in DNA methylcytosine and hydroxymethylcytosine analysis. Rather than introducing a bulky modification on cytosines

in the bisulfite-free 5fC sequencing method reported recently (B. Xia et al., Bisulfite-free, base-resolution analysis of 5-formylcytosine at the genome scale. *Nat. Methods* **12**, 1047-1050 (2015); C. Zhu et al., Single-Cell 5-Formylcytosine Landscapes of Mammalian Early Embryos and ESCs at Single-Base Resolution. *Cell Stem Cell* **20**, 720-731 (2017)), TAPS converts modified cytosine into DHU, a near natural base, which can be “read” as T by common polymerases and is potentially compatible with PCR-free DNA sequencing. TAPS is compatible with a variety of downstream analyses, including but not limit to, pyrosequencing, methylation-sensitive PCR, restriction digestion, MALDI mass spectrometry, microarray and whole-genome sequencing. Since TAPS can preserve long DNA, it can be extremely valuable when combined with long read sequencing technologies, such as SMRT sequencing and nanopore sequencing, to investigate certain difficult to map regions. It is also possible to combine pull-down methods with TAPS to further reduce the sequencing cost and add base-resolution information to the low-resolution affinity-based maps. Herein, it was demonstrated that TAPS could directly replace WGBS in routine use while reducing cost, complexity and time required for analysis. This could lead to wider adoption of epigenetic analyses in academic research and clinical diagnostics.

Claims:

1. A method for identifying 5-methylcytosine (5mC) in a target nucleic acid comprising the steps of:
 - a. providing a nucleic acid sample comprising the target nucleic acid;
 - b. modifying the nucleic acid comprising the steps of:
 - i. adding a blocking group to the 5-hydroxymethylcytosine (5hmC) in the DNA sample;
 - ii. converting the 5mC in the nucleic acid sample to 5-carboxylcytosine (5caC) and/or 5-formylcytosine (5fC); and
 - iii. converting the 5caC and/or 5fC to dihydrouracil (DHU) to provide a modified nucleic acid sample comprising a modified target nucleic acid; and
 - c. detecting the sequence of the modified target nucleic acid;wherein a cytosine (C) to thymine (T) transition in the sequence of the modified target nucleic acid compared to the target nucleic acid provides the location of a 5mC in the target nucleic acid.
2. The method of claim 1, wherein the percentages of a T at each transition location provide a quantitative level of 5mC at each location in the target nucleic acid.
3. A method for identifying 5mC or 5hmC in a target nucleic acid comprising the steps of:
 - a. providing a nucleic acid sample comprising the target nucleic acid;
 - b. modifying the nucleic acid comprising the steps of:
 - i. converting the 5mC and 5hmC in the nucleic acid sample to 5-carboxylcytosine (5caC) and/or 5fC; and
 - ii. converting the 5caC and/or 5fC to dihydrouracil (DHU) to provide a modified nucleic acid sample comprising a modified target nucleic acid; and
 - c. detecting the sequence of the modified target nucleic acid;wherein a cytosine (C) to thymine (T) transition in the sequence of the modified target nucleic acid compared to the target nucleic acid provides the location of either a 5mC or 5hmC in the target nucleic acid.
4. The method of claim 3, wherein the percentages of a T at each transition location provide a quantitative level of 5mC or 5hmC at each location in the target nucleic acid.
5. A method for identifying 5mC and identifying 5hmC in a target nucleic acid comprising:
 - a. identifying 5-methylcytosine (5mC) in the target nucleic acid comprising the steps of:
 - i. providing a first nucleic acid sample comprising the target nucleic acid;

- ii. modifying the nucleic acid in the first sample comprising the steps of:
 - 1. adding a blocking group to the 5-hydroxymethylcytosine (5hmC) in the first nucleic acid sample;
 - 2. converting the 5mC in the first nucleic acid sample to 5caC and/or 5fC; and
 - 3. converting the 5caC and/or 5fC to dihydrouracil (DHU) to provide a modified first nucleic acid sample comprising a modified target nucleic acid;
 - iii. optionally amplifying the copy number of the modified target nucleic acid; and
 - iv. detecting the sequence of the modified target nucleic acid; wherein a cytosine (C) to thymine (T) transition in the sequence of the modified target nucleic acid compared to the target nucleic acid provides the location of a 5mC in the target nucleic acid.
- b. identifying 5mC or 5hmC in the target nucleic acid comprising the steps of:
- i. providing a second nucleic acid sample comprising the target nucleic acid;
 - ii. modifying the nucleic acid in the second sample comprising the steps of:
 - 1. converting the 5mC and 5hmC in the second nucleic acid sample to 5caC and/or 5fC; and
 - 2. converting the 5caC and/or 5fC to dihydrouracil (DHU) to provide a modified second nucleic acid sample comprising a modified target nucleic acid;
 - iii. optionally amplifying the copy number of the modified target nucleic acid; and
 - iv. detecting the sequence of the modified target nucleic acid from the second sample; wherein a cytosine (C) to thymine (T) transition in the sequence of the modified target nucleic acid compared to the target nucleic acid provides the location of either a 5mC or 5hmC in the target nucleic acid; and
- c. comparing the results of steps (a) and (b), wherein a C to T transitions present in step (b) but not in step (a) provides the location of 5hmC in the target nucleic acid.
6. The method of claim 5, wherein in step (a) the percentages of a T at each transition location provide a quantitative level of 5mC at each location in the target nucleic acid; in step (b), the percentages of a T at each transition location provide a quantitative level of 5mC or 5hmC at each location in the target nucleic acid; and in step (c) the differences in percentages for a C to T transition identified in step (b), but not in step (a) provide the quantitative level of a 5hmC at each location in the target nucleic acid.
 7. The method according to any one of claims 1-2 or claims 5-6, wherein the blocking group added to the 5hmC is a sugar.
 8. The method of claim 7, wherein the sugar is glucose or a modified glucose.

9. The method of claim 7, wherein the blocking group is added to the 5hmC by contacting the nucleic acid sample with UDP linked to a sugar in the presence of a glucosyltransferase enzyme.
10. The method of claim 9, wherein the glucosyltransferase enzyme is selected from the group consisting of T4 bacteriophage β -glucosyltransferase (β GT), T4 bacteriophage α -glucosyltransferase (α GT), and derivatives and analogs thereof.
11. The method according to any one of claims 1 to 6, wherein the step of converting the 5mC in the nucleic acid sample to 5caC and/or 5fC and the step of converting the 5mC and 5hmC in the nucleic acid sample to 5caC and/or 5fC comprises contacting the nucleic acid sample with a ten eleven translocation (TET) enzyme.
12. The method of claim 11, wherein the TET enzyme is selected from the group consisting of human TET1, TET2, and TET3; murine Tet1, Tet2, and Tet3; *Naegleria* TET (NgTET); *Coprinopsis cinerea* (CcTET) and derivatives or analogues thereof.
13. The method of claim 12, wherein the TET enzyme is NgTET or murine TET.
14. A method for identifying 5caC or 5fC in a target nucleic acid comprising the steps of:
 - a. providing a nucleic acid sample comprising the target nucleic acid;
 - b. converting the 5caC and 5fC to dihydrouracil (DHU) to provide a modified nucleic acid sample comprising a modified target nucleic acid;
 - c. amplifying the copy number of the modified target nucleic acid; and
 - d. detecting the sequence of the modified target nucleic acid;
 wherein a cytosine (C) to thymine (T) transition in the sequence of the modified target nucleic acid compared to the target nucleic acid provides the location of either a 5caC or 5fC in the target nucleic acid.
15. The method of claim 14, wherein the percentages of a T at each transition location provide a quantitative level of 5caC or 5fC at each location in the target nucleic acid.
16. A method for identifying 5caC in a target nucleic acid comprising the steps of:
 - a. providing a nucleic acid sample comprising the target nucleic acid;
 - b. adding a blocking group to the 5fC in the nucleic acid sample
 - c. converting the 5caC to dihydrouracil (DHU) to provide a modified nucleic acid sample comprising a modified target nucleic acid;
 - d. optionally amplifying the copy number of the modified target nucleic acid; and
 - e. detecting the sequence of the modified target nucleic acid;
 wherein a cytosine (C) to thymine (T) transition in the sequence of the modified target nucleic acid compared to the target nucleic acid provides the location of a 5caC in the target nucleic acid.
17. The method of claim 16, wherein the percentages of a T at each transition location provide a quantitative level of 5caC at each location in the target nucleic acid.
18. The method of claim 16, wherein the step of adding a blocking group to the 5fC comprises contacting the nucleic acid sample with an aldehyde reactive compound

- selected from hydroxylamine derivatives, hydrazine derivatives, and hydrazide derivatives.
19. The method of claim 18, wherein the hydroxylamine derivative is O-ethylhydroxylamine.
20. A method for identifying 5fC in a target nucleic acid comprising the steps of:
- providing a nucleic acid sample comprising the target nucleic acid;
 - adding a blocking group to the 5caC in the nucleic acid sample
 - converting the 5fC to dihydrouracil (DHU) to provide a modified nucleic acid sample comprising a modified target nucleic acid;
 - optionally amplifying the copy number of the modified target nucleic acid; and
 - detecting the sequence of the modified target nucleic acid;
- wherein a cytosine (C) to thymine (T) transition in the sequence of the modified target nucleic acid compared to the target nucleic acid provides the location of a 5fC in the target nucleic acid.
21. The method of claim 20, wherein the percentages of a T at each transition location provide a quantitative level of 5fC at each location in the target nucleic acid.
22. The method of claim 20, wherein the step of adding a blocking group to 5caC comprises contacting the nucleic acid sample with a carboxylic acid derivatization reagent, and an amine, hydrazine or hydroxylamine compound.
23. The method of claim 22, wherein the step of adding a blocking group to 5caC comprises contacting the nucleic acid sample with 1-ethyl-3-(3-dimethylaminopropyl)carbodiimide (EDC) and with ethylamine.
24. The method according to any one of claims 1 to 6 and 14 to 23, wherein the method further comprises the step of amplifying the copy number of the modified target nucleic acid.
25. The method of claim 24, wherein the step of amplifying the copy number of the modified target nucleic acid comprises performing the polymerase chain reaction (PCR) or primer extension.
26. The method according to any one of claims 1 to 6 and 14 to 23, wherein the step of converting the 5caC and/or 5fC to DHU comprises contacting the nucleic acid sample with a reducing agent.
27. The method of claim 26, wherein the reducing agent is selected from the group consisting of pyridine borane, 2-picoline borane (pic-BH₃), borane, sodium borohydride, sodium cyanoborohydride, and sodium triacetoxyborohydride.
28. The method of claim 27, wherein the reducing agent is pyridine borane or 2-picoline borane.
29. The method according to any one of claims 1 to 6 and 14 to 23, wherein the step of detecting the sequence of the modified target nucleic acid comprises one or more of chain termination sequencing, microarray, high-throughput sequencing, and restriction enzyme analysis.
30. The method of any one of claims 1-29, wherein the nucleic acid is DNA.

31. The method of any one of claims 1-29, wherein the nucleic acid is RNA.