Methylation-Assisted bisulfite sequencing to simultaneously map 5fC and 5caC on a genome-wide scale for DNA demethylation analysis

Francesco Neri, Danny Incarnato, Anna Krepelova, Caterina Parlato & Salvatore Oliviero

Methylation-assisted bisulfite sequencing to simultaneously map 5fC and 5caC on a genome-wide scale for DNA demethylation analysis

VOL.11 NO.7 | 2016 pagg **1191-1205**
doi:10.1038/nprot.2016.063

TITLE

Methylation-assisted bisulfite sequencing (MAB-seq) to simultaneously map 5fC and 5caC on a genome-wide scale for DNA demethylation analysis

Francesco Neri[1-3], Danny Incarnato[1-3], Anna Krepelova[1-2], Caterina Parlato[1], and Salvatore Oliviero[1-2]*

[1] Human Genetics Foundation (HuGeF), via Nizza 52, 10126 Torino, Italy.

[2] Dipartimento di Scienze della Vita e Biologia dei Sistemi, Università di Torino, Via Accademia Albertina, 13 - 10123 Torino, Italy.

[3] Equal contribution

* Correspondence to: salvatore.oliviero@hugef-torino.org

KEYWORDS: DNA demethylation, 5fC, 5caC, single-base mapping, MAB-seq

EDITORIAL SUMMARY: Methylation-assisted bisulfite sequencing (MAB-seq) enables direct genome-scale mapping of 5fC and 5caC at single-base resolution to quantify DNA demethylation. Reduced representation (RRMAB-seq) also provides increased coverage on CpG-rich regions to reduce cost.

TWEET: MAB-seq for direct genome-scale mapping of 5fC and 5caC at single-base resolution

ABSTRACT

Active DNA demethylation is mediated by Ten Eleven Translocation (Tet) proteins, which progressively oxidize 5mC to 5-hydroxymethylcytosine (5hmC), 5-formylcytosine (5fC), and 5-carboxylcytosine (5caC). We have developed a methylation-assisted bisulfite sequencing (MAB-seq) method, which enables direct genome-scale mapping and quantitation of 5fC and 5caC marks together at single-base resolution. In bisulfite sequencing (BS)

unmethylated cytosine residues (Cs), 5fCs, and 5caCs, are converted to uracil and cannot be discriminated. The pre-treatment of the DNA with the CpG methylation enzyme M.SssI, which converts only the Cs to 5mCs, protects Cs but not 5fCs and 5caCs, enabling direct detection of 5fCs and 5caCs as uracils. Here we also describe an adapted version of the protocol to perform reduced representation MAB-seq (RRMAB-seq) that provides increased coverage on CpG-rich regions, thus reducing the execution costs and increasing the feasibility of the technique. The main advantage of MAB-seq is to reduce the number of chemical/enzymatic DNA treatments required prior to bisulfite treatment, and avoid the need for prohibitive sequencing coverage, thus making it more reliable and affordable than subtractive approaches. The method here presented is the ideal tool to study the DNA demethylation dynamics in all biological systems. Overall timing is ~3 days for library preparation.

INTRODUCTION

DNA methylation is the most studied epigenetic modification. It consists of the addition of a methyl group to the carbon-5 position of the cytosine, catalysed by DNA methyltransferase enzymes. In mammals it predominantly occurs in the context of CpG dinucleotides, and it is required for a correct development[1-3]. DNA methylation, particularly in CpG-rich promoter regions, has been shown to silence gene expression in a heritable manner[4,5]. The transcriptional silencing associated with 5-methylcytosine is required for fundamental physiological processes such as embryonic development, X-chromosome inactivation, genomic imprinting and protection against intragenomic parasites[1-3,6]. Deficient DNA methylation processes can lead to genetic diseases, such as ICF syndrome, and aberrant methylation patterns are found in the majority of human cancers[7]. In mammals, three principal active DNA methyltransferases exist, which are classified in two families that are structurally and functionally distinct. The DNMT3A and DNMT3B enzymes are responsible for de novo CpG methylation, whereas DNMT1 maintains the methylation pattern during chromosome replication and repair[1,4,6].

Recent studies have shown that DNA methylation can also be actively removed through multiple consecutive oxidative reactions, mediated by the TET proteins[8]. This involves the oxidation of 5mC to 5-hydroxymethylcytosine (5hmC), which can function as a new epigenetic mark, or as an intermediate toward further oxidized states; 5-formylcytosine

(5fC) and 5-carboxylcytosine (5caC)[9-11]. In embryonic and adult stem cells and tissues, 5hmC modification occurs at high and medium levels, while it is drastically reduced in cancers, especially in CRC (colon-rectal cancer) , where specific DNA hypermethylated regions are able to drive tumorigenesis[12,13]. 5fC and 5caC can be excised by the TDG protein to allow the restoration of unmodified C by the base excision and repair (BER) machinery, thus mediating an active DNA demethylation process[8,14,15]. Several previous studies reported that 5fC and 5caC could act as real epigenetic markers, due to the presence of a high number of nuclear proteins that are able to recognize and specifically bind these modifications[16-18], even though their relatively low abundance in the mammalian genome makes this hypothesis difficult to verify[19].

From this perspective, the development of streamlined and sensitive approaches for the accurate detection of these modified sites on a genome-wide scale has become a key need to facilitate a deeper understanding of their potential role in regulating cell's physiology and behaviour.

Development of the MAB-seq method

Bisulfite sequencing is the most widely used and reliable method to analyze the DNA methylation status of genomic cytosines. It consists of the chemical treatment of DNA with Sodium bisulfite to deaminate unmodified cytosines to uracils. The PCR amplification of the bisulfite-treated DNA will amplify the uracils as thymines, while methyl-cytosines (5mC) are protected from deamination by the methyl group, and will be amplified as cytosine. Sequencing of the PCR product will reveal the methylation status of the starting DNA at single-nucleotide resolution[20]. The accuracy of this method depends on the cytosine-to-uracil bisulfite conversion rate, that can easily reach more than 99% by adjusting incubation conditions, thus enabling accurate and reliable analyses[21]. Recent studies reported that 5hmC also is protected from bisulfite conversion, demonstrating that traditional bisulfite sequencing experiments are not able to discriminate between 5mC and 5hmC. 5fC and 5caC are however converted into uracils by bisulfite treatment[22,23]. Based on this evidence, we have developed a methylation-assisted bisulfite (MAB) sequencing method, that consists of an enzymatic pre-treatment of the genomic DNA with the M.SssI CpG methyltransferase before the bisulfite conversion, in order to protect the unmodified cytosines from deamination[24] (Fig. 1). M.SssI is a bacterial enzyme, that, in the presence of the S-

adenosylmethionine (SAM), methylates all unmodified cytosine residues (C5) within the double-stranded dinucleotide recognition sequence 5'...CG...3' [25]. Bisulfite sequencing of the M.SssI-treated DNA causes unmodified Cs, 5mCs, and 5hmCs to be read as Cs, while 5fC and 5caC residues will be read as thymines, thus allowing their identification at single-base resolution. A pilot experiment on PCR generated DNA fragments showed that the M.SssI methylation rate is 99.9% **[AU: Please add reference]**. Since some authors observed a detectable conversion of the 5mC to T[26-27] we tested several commercial kits for bisulfite conversion and protocols. The EpiTect Bisulfite Kit (Qiagen) together with the the conversion conditions recommended for WGBS or RRBS by Illumina (https://support.illumina.com/downloads/wgbs_for_methylation_analysis_guide_15021861.html), and freshly prepared bisulfite solutions gave the best results, achieving the rate of more than 99.9% of both unmodified cytosine conversion to uracils and 5mC/5hmC protection[24]. **[AU: Suggest adding sentence along the following lines to address outstanding reviewer concerns: The bisulfite conversion efficiency we obtain is higher than that obtained by other groups (provide example conversion rates and references, for example as cited by Reviewer); we think this is due to XXXX.]** We then applied the method to a genome-scale analysis to identify 5fC and 5caC along the genomic DNA. We developed both a genome-wide (MAB-seq) and a reduced representation (RRMAB-seq) version of the technique, which we will explain in detail in this protocol. Both types of experiments require high-throughput massive parallel sequencing to ensure a significant sequencing depth. We will describe here the protocol for MAB-seq for sequencing on Illumina platforms, but other sequencing platforms are also suitable by simply changing the adapters used for the library generation.

Applications of the method

One of the major advantages of MAB-seq is the ability to map 5fC/5caC, and to quantify their abundance in a single sequencing experiment. Single-base resolution makes this method suitable for the analysis of the DNA sequence context surrounding 5fC and 5caC, of their CpG symmetry and of their correlation with known nuclear proteins binding motifs. RRMAB-seq is a cheaper version of the MAB-seq approach, which increases the coverage in CpG-rich genomic regions, especially at gene promoters, thus providing a huge amount of information while only requiring a modest volume of sequencing. This and the relative ease

of library generation enable simultaneous comparative studies of more samples. To independently study the distribution of 5caC and 5fC residues it would be possible to further reduce 5fC to 5hmC by sodium borohydride treatment[26] in order to map only 5caC, and then perform subtractive analysis by comparison with a standard MAB-seq experiment to map 5fC indivudally.

MAB-seq is able to map the two final 5mC oxidization products together (5fC and 5caC), thus making the technique highly suitable for the study of the active DNA demethylation dynamics in normal or pathological biological processes. Thanks to the recent "omics" approaches, more and more genes were identified undergoing DNA hyper- or hypo-methylation on their promoter, especially during cancer development. The understanding of the mechanisms driving promoter DNA demethylation is becoming a key need to lay the foundation for possible future epigenetic therapies.

Comparison with other methods

Several methods have been developed for 5fC and 5caC genomic mapping in the last two years. Similar to 5mC and 5hmC analysis methods, approaches for mapping 5fC and 5caC fall into two major categories: (i) affinity-based precipitation of the modified cytosines using either specific antibodies, or chemically labelled modifications, followed by high-throughput sequencing (5fC-DP-seq, 5caC-DIP-seq, 5fC-DIP-seq, fC-seal-seq)[28-30]; (ii) Enzymatic or chemical treatment to distinguish the DNA modification after bisulfite sequencing (fCAB-seq, CAB-seq, redBS-seq, MAB-seq)[24,26,29,31] (Table 1). Methods falling in the first category cannot resolve the presence of 5fC and 5caC at single-nucleotide resolution and their relative level of modification **[AU: Please clarify – do you mean distinguish 5fC from 5caC]**, thus these methods lack the possibility to obtain relevant information on these modifications. As for all immunoprecipitation-based genome-wide experiments, the success of these methods highly depends on the relative level of the modification at each genomic site, on the number of modified CpGs in each genomic region, on the strength of affinity purification, and, in the case of chemical or enzymatic pre-treatment of the modification, on the reaction efficiency of the treatment. All these possible issues, coupled to the rarity and instability of 5fC and 5caC in the genome, could make these methods technically difficult to

perform and to interpret. Indeed, their application to mouse embryonic stem cells led to different results depending on the used approach[28-30].

The methods belonging to the second category are all variants of the traditional bisulfite sequencing technique, and therefore are not subject to these technical issues. The only possible pitfall could be the low efficiency rate of the enzymatic or chemical treatments prior to bisulfite conversion, which can generate false positive or negative calls. However, the fact that MAB-seq requires only a single DNA pre-treatment prior to bisulfite conversion, and a lower sequencing depth, strongly reduces these downsides. Conversely to fCAB-seq, CAB-seq, and redBS-seq, MAB-seq enables the direct identification, at single-base resolution, of 5fC and 5caC residues, and their relative quantitation at the level of individual cytosines, even though is not able to distinguish between them. Indeed, MAB-seq has the advantage of being able to discover both the modifications in a single experiment, thus reducing the economic costs and increasing the reproducibility. Remarkably, fCAB-seq, CAB-seq, and redBS-seq require the sequencing of a non-treated DNA control, and the calling of the modified cytosines derives from a subtractive analysis between the two sequencings. The main disadvantage in this kind of approach is the requirement for a higher sequencing depth to reduce the false discovery rate.

Limitations of the method

The major limitation of the MAB-seq is its inability to distinguish between 5fC and 5caC. This limitation can be overcome by performing an additional treatment with sodium borohydride[26] prior to bisulfite conversion, thus allowing the method to discriminate between 5fC and 5caC residues. However, current knowledge on these modifications suggests that since they are catalysed and removed by the same enzymes (TET and TDG proteins respectively), they can be essentially considered together when studying DNA demethylation dynamics. This assumption makes the MAB-seq the most economic, fast and reliable method currently available for the study of the DNA demethylation dynamics in mammalian genomes.

Experimental Design

The workflow of the MAB-seq protocol is represented in Figure 2. Some critical steps and recommendations are outlined below.

Genomic DNA extraction (Step 1). Genomic DNA (gDNA) can be derived from any cellular line, tissue or organism using the preferred extraction method. However, traditional phenol/chloroform extraction should be avoided because residues of these reagents could interfere with the M.SssI CpG methylation step, thus compromising the whole experiment. We recommend using the DNeasy Blood & Tissue Kit, following the instructions described in the Procedure. Non-degraded gDNA is recommended especially in the RRMAB-seq version of the protocol (see Fig. 3a).

Spike-in control (Step 4). The use of an unmethylated spike-in is a necessary step for calculating the M.SssI methylation fail rate. We recommend the use of unmethylated Lambda DNA, although other spike-in controls can be used, including *E.coli* gDNA or DNA fragments generated by PCR. We experimentally concluded that a 1:1000 ratio of the spike-in with respect to gDNA is sufficient to estimate the M.SssI fail rate, which, if underestimated, will cause false positive callings. Even though the bisulfite protocol is highly standardized, and a low bisulfite conversion efficiency only leads to an underestimate of the actual number of 5fC and 5caC residues, also evaluating the experimental bisulfite conversion rate is good practice. It can be easily calculated by measuring the cytosines in non-CpG contexts in the fully unmethylated spike-in control, that undergoes conversion to thymine. The use of spike-in controls generated by PCR containing modified cytosines (5mC, 5hmC, 5fC and 5caC) can be very useful for calculating the rate of eventual sporadic 5mC/5hmC deaminations, as well as 5fC/5caC conversion rates.

gDNA fragmentation and purification (Steps 12-15). Genomic DNA can be fragmented using the preferred instrument by following the manufacturer's protocol to generate DNA fragments averaging 200 bp in length (Fig. 3b). Whole or fragmented gDNA can be purified using any method, however we recommend avoiding ethanol precipitation because, after several steps of purification, the amount of salts could interfere with the following enzymatic reactions. We strongly recommend using the Agencourt AMPure XP Beads (Beckman Coulter).

M.SssI methylation (Steps 4-6 and 36-39). In vitro CpG methylation is the key step for the success of the MAB-seq experiment. For this purpose, three rounds of M.SssI methylation are performed on gDNA. To test the success of the *in vitro* CpG methylation step, methylated genomic DNA can be digested either with MspI, or its methylation-sensitive isoschizomer HpaII. MspI should completely digest the genomic DNA, while no smear should be present in the HpaII treated DNA (Fig. 3c). Three additional rounds of M.SssI methylation are performed after adapter ligation, in order to methylate eventual cytosines added during the end-repair step. This process could be avoided by trimming of the 5' reads (and of the 3' of the clipped reads) during the analysis, but we observed that M.SssI methylation of fragmented DNA further increased the overall accuracy of the MAB-seq experiment.

Bisulfite conversion (Steps 57-60). Several bisulfite conversion commercial kits are optimized for the conversion of cytosine to uracil. It has been recently demonstrated that 5fC conversion to uracil is slower than the unmodified cytosine[32]. To ensure full conversion 5fC to uracil we recommend a longer incubation time as described in the present protocol. This longer incubation has no effect on the final converted DNA yield. To avoid 5mC and 5hmC deamination we recommend to use only freshly prepared bisulfite mix and DNA protect buffer present in the EpiTect Bisulfite Kit.

Sequencing and data analysis (Steps 94-104). High quality reads **[AU: Edit OK]** could increase mapping percentage and avoid false callings of the modified cytosine. On Illumina platforms, we suggest to use a lower cluster density (about 70% of the maximum density recommended by Illumina) in the flowcell and, if possible, to mix the sample with a sequencing spike-in or other sequencing libraries (e.g. RNA-seq or ChIP-seq libraries), to balance the CG-content across each sequencing lane. These precautions are especially important when performing RRMAB-seq. The recommended sequencing coverage is about 1-1.5x10^9 sequencing reads for MAB-seq, and 3-4x10^8 reads for RRMAB-seq. Raw reads should be trimmed to remove low quality bases, clipped to remove adapter sequences and mapped on a specific reference genome. This allows to account for C->T single nucleotide **[AU: Please clarify what you mean]**, which may lead to false positive calls of modified 5fC/5caC residues. For samples derived from mice, strain-specific genome reference sequences are available from the Sanger Institute webpage

(http://www.sanger.ac.uk/resources/mouse/genomes/). For mouse embryonic stem cells (mESC) E14, we used a variant of the mouse mm9 reference assembly published from our lab[33]. The removal of low quality sequences and the mapping on a strain specific genome assembly will improve mapping accuracy and will decrease the number of false methylation status callings. The calling of 5fC and 5caC should be performed using a binomial test to take into account the M.SssI methylation fail rate and sequencing coverage of that base. Due to the low presence of 5fC and 5caC in mESCs, we considered as 5fC or 5caC only the cytosines covered at least 50X with a p-value < 0.001.

MATERIALS

REAGENTS

CRITICAL: All reagents must be kept DNase-free.

- Nuclease-free water (Life Technologies, cat. no. AM9930)
- Ethanol (Sigma Aldrich, cat. no. 02860)

CAUTION: Ethanol is highly flammable.

- E14 mouse embryonic stem cells (ATCC®, cat. no. CRL-1821™)
  CAUTION: The cell lines used in your research should be regularly checked to ensure they are mycoplasma-free.

- DNeasy Blood & Tissue Kit (Qiagen, cat. no. 69504)
- PfuTurbo Cx Hotstart DNA Polymerase (Agilent, cat. no. 600410)
- dNTP Set 100mM (Life Technologies, cat. no. 10297)
- MspI (New England Biolabs, cat. no. R0106M)
- M.SssI CpG Methyltransferase (New England Biolabs, cat. no. M0226M)
- HpaII (New England Biolabs, cat. no. R0171M)
- NEBNext® DNA Library Prep Master Mix Set for Illumina® (New England Biolabs, cat. no. E6040L)

- EpiTect Bisulfite Kit (Qiagen, cat. no. 59110)
- Sodium chloride (Hampton Research, cat. no. HR2-637)
- EDTA (Sigma Aldrich, cat. no. 03690-100ML)
- UltraPure™ 1M Tris-HCl, pH 8.0 (Life Technologies, cat. no. 15568-025)
- Agencourt AMPure XP Beads (Beckman Coulter, cat. no. A63880)
- 1 Kb Plus DNA Ladder (Life Technologies, cat. no. 10787-018)
- 100 bp DNA Ladder (Life Technologies, cat. no. 15628-019)
- MinElute Gel Extraction Kit (Qiagen, cat. no. 28604)
- MinElute PCR Purification Kit (Qiagen, cat. no. 28004)
- TRIS borate - EDTA buffer solution (Sigma Aldrich, cat. no. 93290)
- SYBR® Safe DNA Gel Stain (Life Technologies, cat. no. S33102)
- NEBNext® Multiplex Oligos for Illumina® (New England Biolabs, cat. no. E7535S)
- Unmethylated Lambda DNA (Promega, cat. no. D1521)
- Qubit dsDNA high-sensitivity (HS) assay (Life Technologies, cat. no. Q32851)
- Agarose (Sigma Aldrich, cat. no. A9539-500G)
- GelPilot DNA Loading Dye, 5x (Qiagen, cat. no. 239901)

SOFTWARE

- Any Linux/UNIX environment
- FastX toolkit (http://hannonlab.cshl.edu/fastx_toolkit/)
- BSMAP (https://code.google.com/p/bsmap/)[34]
- BSMAP Postprocess and 5fC/5caC BinTest (see Supplementary Material)

EQUIPMENT

- Tabletop microcentrifuge
- Electrophoresis cell
- Power supply (GE Healthcare, cat. no. 18-1130-02)
- Thermal cycler
- Microwave oven
- Heath block (or water bath)
- Sonicator (Covaris™ or Bioruptor®)
- NanoDrop spectrophotometer
- DynaMag™-2 Magnetic stand (Life Technologies, cat. no. 12321D)
- Freezers, -20°C and -80°C

- Advanced Analytical Fragment Analyzer™ (or Agilent Bioanalyzer)
- Blue-light transilluminator (Life Technologies, cat. no. G6600)
- Qubit fluorometer (Life Technologies, cat. no. Q32866)
- Illumina sequencer (Genome Analyzer II, HiSeq 2000 or greater, HiScan SQ, or NextSeq 500)

REAGENT SETUP

- **1-2% (wt/vol) Agarose gel.** Dissolve 1 gr (2 gr for 2% gels) of agarose powder, in 100 ml of 1X TRIS Borate-EDTA buffer. Add 10 µl of SYBR Safe DNA Gel Stain directly to the solution, and mix thoroughly by shaking. Heat solution in a microwave oven, with occasional shaking, until the agarose has completely dissolved.
- **Unmethylated Lambda DNA.** Quantify Lambda DNA using a NanoDrop spectrophotometer. Dilute to 1 ng/µl in nuclease-free water. This dilution can be stored at -20°C for 1 month.
- **M.SssI buffer (10X).** [100mM Tris-HCl pH 8.0; 500mM NaCl; 100mM EDTA]. This buffer can be stored at room temperature (RT) **[AU: Please specify what this is]** for 6 months.

  CRITICAL Alternatively, NEBuffer II can be used. Although in the presence of Mg^{++} ions the M.SssI enzyme exhibits a distributive instead of a processive activity, we obtained high methylation efficiencies in both conditions.

- **3.2mM S-adenosylmethionine stock (SAM).** SAM is provided as a 32 mM stock. Dilute 1 µl of the SAM stock in 9 µl of nuclease-free water, to obtain a 3.2mM stock. Store at -20°C for 1 month.
- **TE buffer** [10mM Tris pH 8.0; 1mM EDTA]. This buffer can be stored at RT for 6 months.

EQUIPMENT SETUP

- **Shearing settings for Bioruptor® sonicator.** [Time: 30s ON, 30s OFF; Power: High; Number of cycles: 10]

PROCEDURE

gDNA Methylation TIMING: ~9 h

1. Purify genomic DNA from your desired cell line (e.g. E14 mouse embryonic stem cells), using the DNeasy Blood & Tissue Kit. Follow manufacturer instructions for "Purification of Total DNA from Animal Blood or Cells (Spin-Column Protocol)"

   CAUTION: The cell lines used in your research should be regularly checked to ensure they are mycoplasma-free.

PAUSE POINT: gDNA can be stored at -80°C indefinitely.

2. Quantify gDNA using a NanoDrop spectrophotometer.
3. Inspect DNA integrity by running 100 ng of gDNA on a 1% agarose gel at 100 V for 30 min. For high quality DNA you would expect to see a single sharp thick band above the upper band of the 1 Kb Plus size ladder (see Fig. 3a).

CRITICAL STEP: If any smear appears, the gDNA is probably partially degraded. It is essential to have intact DNA, especially when performing the RRMAB-seq protocol, otherwise fragments other than the ones generated by MspI cutting will be adapter-ligated.

4. Setup the *in vitro* methylation reaction in a sterile 0.2 ml PCR tube, according to the following scheme:

Reagent	**Volume (µl)**
gDNA	variable (1 µg total)
Unmethylated Lambda DNA (1 ng/µl)	1
M.SssI buffer (10X)	2
SAM (3.2 mM)	1
CpG Methyltransferase M.SssI (20U/µl)	1
Nuclease-free water	variable (to 20 µl)

CRITICAL STEP: If performing whole-genome MAB-seq, setup two identical reactions in parallel, for a total of 2 µg of gDNA.

5. Incubate the tube in the thermal cycler for 2 h at 37°C.

6. Following incubation, add to the tube:

Reagent	**Volume (µl)**
M.SssI buffer (10X)	2
SAM (3.2 mM)	1
CpG Methyltransferase M.SssI (20U/µl)	1
Nuclease-free water	16

7. Incubate the tube in the thermal cycler for additional 2 h at 37°C**.**

8. During incubation, thaw Agencourt AMPure XP Beads at room temperature (RT) for at least 30 min. Prepare a fresh 80% (v/v) dilution of ethanol in nuclease-free water.

9. Following the 2 h incubation, transfer the whole reaction volume (40 µl) to a sterile 1.5 ml Eppendorf tube.

10. Purify using 2.8 volumes (112 µl) of Agencourt AMPure XP Beads (as described in Box 1). Elute in 16 µl nuclease-free water.

11. Repeat Steps 4-10 two additional times without adding Unmethylated Lambda DNA, for a total of 3 rounds of *in vitro* methylation.

PAUSE POINT: Methylated gDNA can be stored at -80°C for up to 6 months.

Fragmentation/digestion of methylated gDNA

12. Methylated gDNA should be fragmented by sonication for the whole-genome MAB-seq method using option A, or for RRMAB-seq, digest methylated gDNA following option B.

A. gDNA shearing (Whole-genome MAB-seq protocol) TIMING: ~1 h

i) Pool the two methylated gDNA samples (~2 µg total) in a single tube, and bring the final volume (32 µl) to 100 µl with TE buffer.

ii) Perform shearing using the selected sonicator, according to the settings specified in the Equipment Setup section.

iii) If using a Bioruptor® sonicator, perform three rounds of 10 sonication cycles, according to the specifications detailed in the Equipment Setup section.

CRITICAL STEP: Bioruptor® sonicators tend to overheat when performing more than 15 consecutive cycles. To avoid overheating, wait at least 10 minutes between each round of sonication. During this time, briefly spin samples, and store them on ice.

iv) Inspect sonication efficiency by running 1 µl of sheared gDNA on the Fragment Analyzer (or Agilent Bioanalyzer). Alternatively, you can run 5 µl of sheared gDNA on a 1% agarose gel at 100 V for 30 min. A smear should be present around 100-300 bp, with the maximum enrichment around 150-200 bp (See Fig. 3b).

B. gDNA digestion (RRMAB-seq protocol) TIMING: ~4 h

i) Transfer the methylated gDNA sample (~1 µg total) to a sterile 0.2 ml PCR tube. Bring the reaction volume (~16 µl) to 44 µl with nuclease-free water.

ii) Setup the digestion reaction according to the following scheme:

Reagent	**Volume (µl)**
gDNA	44
NEBuffer 4 (10X)	5
MspI (100U/µl)	1

iii) Incubate the tube in the thermal cycler for 2 h at 37°C.

iv) Following incubation, add to the tube:

Reagent	**Volume (µl)**
Nuclease-free water	44
NEBuffer 4 (10X)	5

MspI (100U/µl)	1

v) Incubate the tube in the thermal cycler for additional 2 h at 37°C. .

vi) Inspect digestion efficiency by running 1 µl of digested gDNA on the Fragment Analyzer (or Agilent Bioanalyzer). Alternatively, you can run 10 µl of digested gDNA on a 1% agarose gel at 100 V for 30 min. A smear should be present from the top to the bottom of the gel (See Fig. 3c, lane 2).

Fragmented/Digested gDNA purification TIMING: ~45 m

13. Repeat Step 8.

14. Transfer the whole reaction volume from Step 12 (~100 µl) to a sterile 1.5 ml Eppendorf tube.

15. Purify using 1.6 volumes (160 µl) of Agencourt AMPure XP Beads (see Box 1). Elute in 85 µl nuclease-free water.

PAUSE POINT: Fragmented/Digested gDNA can be stored at -80°C for up to 2 months.

End repair of gDNA TIMING: ~1h30m

16. Setup the end repair reaction according to the following scheme:

Reagent	**Volume (µl)**
Fragmented/Digested gDNA from Step 15	85
NEBNext End Repair Reaction Buffer (10X)	10
NEBNext End Repair Enzyme Mix	5

17. Incubate the tube in the thermal cycler at 20°C for 30 min.

18. During incubation, repeat Step 8.

19. Following incubation, transfer the whole reaction volume (100 µl) to a sterile 1.5 ml Eppendorf tube.

20. Purify using 1.6 volumes (160 µl) of Agencourt AMPure XP Beads (see Box 1). Elute in 42 µl nuclease-free water.

21. Transfer 42 µl of the cleared supernatant from the tube **[AU: Do you mean the 42ul eluted DNA from Step 20?]**to a sterile 0.2 ml PCR tube.

PAUSE POINT: End-repaired DNA can be stored at -20°C for up to 2 months.

dA-Tailing of end-repaired DNA TIMING: ~1h30m

22. Setup the dA-tailing reaction according to the following scheme:

Reagent	**Volume (µl)**
End-repaired DNA from Step 21	42
NEBNext dA-Tailing Reaction Buffer (10X)	5
Klenow Fragment (3'→5' exo$^-$)	3

23. Incubate the tube in the thermal cycler at 37°C for 30 min.

24. During incubation, repeat Step 8.

25. Following incubation, transfer the whole reaction volume (50 µl) to a sterile 1.5 ml Eppendorf tube.

26. Purify using 1.8 volumes (90 µl) of Agencourt AMPure XP Beads (see Box 1). Elute in 25 µl nuclease-free water.

27. Transfer 25 µl of the cleared supernatant from the tube **[AU: Do you mean the 25ul eluted DNA from Step 26?]**to a sterile 0.2 ml PCR tube.

PAUSE POINT: dA-Tailed DNA can be stored at -20°C for up to 2 months.

Adapter ligation TIMING: ~1h30m

28. Setup the ligation reaction according to the following scheme:

Reagent	**Volume (µl)**
dA-Tailed DNA from Step 27	25
Quick Ligation Reaction Buffer (5X)	10

NEBNext Methylated Adaptor for Illumina (15 µM)	10
Quick T4 DNA Ligase	5

29. Incubate the tube in the thermal cycler at 20°C for 15 min.

30. During incubation, repeat Step 8.

31. Immediately after the 15 min incubation, add 3 µl of USER™ Enzyme Mix to the reaction, and mix thoroughly by pipetting up and down at least 10 times.

32. Incubate the tube in the thermal cycler at 37°C for 15 min.

33. Following incubation, transfer the whole reaction volume (53 µl) to a sterile 1.5 ml Eppendorf tube.

34. Purify using 90 µl of Agencourt AMPure XP Beads (see Box 1). Elute in 16 µl nuclease-free water.

35. Transfer 16 µl of the cleared supernatant from the tube **[AU: Do you mean the 16ul eluted DNA from Step 34?]**to a sterile 0.2 ml PCR tube.

PAUSE POINT: Adapter-ligated DNA can be stored at -80°C indefinitely.

Methylation of DNA termini TIMING: ~9h

36. Using 16 µl of adapter-ligated DNA from Step 35 in the place of both gDNA and unmethylated lambda DNA, carry out the *in vitro* methylation reactions by repeating Steps 4-9. **[AU: Edit correct?]**

37. Purify using 1.6 volumes (64 µl) of Agencourt AMPure XP Beads (see Box 1). Elute in 16 µl nuclease-free water.

38. Transfer 16 µl of the cleared supernatant from the tube **[AU: Do you mean the 16ul eluted DNA from Step 37?]** to a sterile 0.2 ml PCR tube.

39. Repeat steps 36-38 an additional two times, for a total of 3 rounds of *in vitro* methylation.

PAUSE POINT: Methylated DNA can be stored at -80°C for up to 2 months.

Size-selection of DNA TIMING: ~2h

40. Dilute 1 µl of the 100 bp DNA Ladder with 7 µl of nuclease-free water, and then add 3 µl of 5x GelPilot DNA Loading Dye. Mix thoroughly.

41. Add to sample (~16 µl from Step 39) 4 µl of 5x GelPilot DNA Loading Dye, and mix thoroughly by pipetting (or vortexing).

42. Load the DNA ladder from Step 40 and sample from Step 41 on a 2% (wt/vol) agarose gel (see Reagent Setup). If processing more than one sample, skip at least one lane between each sample to avoid cross-contamination.

43. Run at 80V for 30 min in 1x TBE.

44. Visualize the gel on a blue-light transilluminator (see Fig. 3d). Using a sterile scalpel cut a gel slice corresponding to fragments in the range 200-400 bp, and transfer it to a sterile 2ml Eppendorf tube.

45. Weigh the gel slices, and add 6 volumes (variable) of Buffer QG. If the slice weight is >400mg, split the slice into two tubes.

46. Incubate samples on a rocking shaker at 50°C until the gel slices have completely dissolved. It should take approximately 20 min.

47. Add 1 gel volume of Isopropanol, and mix vigorously by inverting the tube (or by vortexing).

48. Transfer up to 700 µl of the dissolved gel slice to a MiniElute column. Centrifuge at 14,000*g* for 1 min at room temperature.

49. Discard the flow-through. Repeat Step 48 with the residual reaction volume.

50. Pipette 750 µl of PE buffer into the columns. Centrifuge at 14,000*g* for 1 min at room temperature. Discard the flow-through and transfer the columns to sterile 2ml collection tubes.

51. Centrifuge the columns at 14,000*g* for 2 min at room temperature with open lids to completely dry the membranes.

52. Carefully remove any residual ethanol using a fine pipet tip.

53. Transfer the columns to clean 1.5 ml Eppendorf tubes. Carefully pipette 20 µl of Buffer EB directly onto the center of the membrane, and incubate on the bench top for 5 min.

54. Centrifuge at maximum speed for 1 min at room temperature to collect the size-selected DNA.

55. Repeat Steps 53-54 once.

56. Transfer the entire eluate (~40 µl) to a sterile 0.2 ml PCR tube.

PAUSE POINT: Size-selected DNA can be stored at -80°C for up to 2 months.

Bisulfite conversion of size-selected DNA TIMING: ~15h

57. Dissolve one vial of Bisulfite Mix (Epitect Bisulfite Kit) by adding 800 µl of nuclease-free water, and incubate for 5 min 60°C.

CRITICAL STEP: Ensure that the Bisulfite Mix is completely dissolved before proceeding with the reaction setup, otherwise the conversion efficiency would be compromised. If any precipitate is still visible, mix thoroughly by vigorously vortexing.

CRITICAL STEP: Each time a MAB-seq experiment is performed, use a freshly dissolved Bisulfite Mix. Any residual Bisulfite Mix should be discarded after each experiment, and not re-used, otherwise the conversion efficiency would be compromised.

CRITICAL STEP: Each time a MAB-seq experiment is performed, ensure that the DNA protect buffer color is bright green. Color changes are caused by pH variations, and will compromise 5mC and 5hmC protection and total DNA yields.

58. Setup the bisulfite conversion reaction according to the following scheme:

Reagent	**Volume (µl)**
Size-selected DNA from Step 56	40
Bisulfite Mix	85
DNA Protect Buffer	15

59. Mix thoroughly by pipetting up and down the entire volume at least 20 times.

CRITICAL STEP: It is essential to mix thoroughly the entire reaction volume, until the entire solution appears uniformly blue. Incomplete mixing of the reaction components would compromise the conversion efficiency.

60. Perform the bisulfite conversion reaction following the conditions described in the table below:

Step	Temperature (°C)	Time
1	95	5 min
2	60	25 min
3	95	5 min
4	60	85 min
5-12	95	5 min
	60	180 min
13	4	Hold

Clean-up of bisulfite converted DNA TIMING: ~1h30m

61. Pre-heat a heat block (or water bath) at 56°C. During the clean-up, heat at least 60 µl of Buffer EB at 56°C for each sample being processed.

62. Transfer the entire bisulfite reaction volume (~140 µl from Step 60) to a sterile 1.5 ml Eppendorf tube, including any precipitate that might have formed.

63. Add 310 µl of Buffer BL, supplemented with 10 mg/ml of carrier RNA, then mix thoroughly by vortexing.

64. Add 250 µl of 100% ethanol to the sample, then mix thoroughly by vortexing for at least 30 s. Briefly spin the tube to collect solution.

65. Transfer the entire volume to an EpiTect spin column.

66. Centrifuge at 14,000*g* for 1 min at RT, and then discard the flow-through.

67. Add 500 µl of Buffer BW, centrifuge at 14,000*g* for 1 min at RT and then discard the flow-through.

68. Add 500 µl of Buffer BD to the spin column.

CRITICAL STEP: Buffer BD is basic, so it's susceptible to acidification from the CO_2 present in the air. After transferring 500 µl of Buffer BD to the spin column, immediately close the Buffer BD bottle and the column.

69. Incubate on the bench top for 20 min.

70. Centrifuge at 14,000*g* for 1 min at RT and then discard the flow-through.

71. Repeat Steps 68-70 once.

72. Add 500 µl of Buffer BW, centrifuge at 14,000*g* for 1 min at RT, and then discard the flow-through.

73. Repeat Step 72 once.

74. Place the spin column in a new 2 ml collection tube, and centrifuge at 14,000*g* for 2 min at RT to remove any residual ethanol.

75. Place the spin column with open lid in the pre-heated heat block (or water bath) from Step 61, and incubate for 5 min to completely dry the silica membrane.

76. Transfer the spin column to a sterile 1.5 ml Eppendorf tube, pipette 20 µl of pre-heated Buffer EB onto the center of the spin column, and incubate on the bench top for 1 min.

77. Centrifuge at 14,000*g* for 1 min at RT.

78. Repeat Steps 76-77 once, using the same 1.5 ml Eppendorf collection tube **[AU: Edit correct?]**.

79. To the eluted DNA (~40 µl), add 5 volumes of Buffer PB (200 µl), and incubate in a heat block (or water bath) at 37°C for 15 min.

80. Transfer the entire volume to a MinElute spin column, centrifuge at 14,000*g* for 1 min at RT, and then discard the flow-through.

81. Add 750 µl of Buffer PE, centrifuge at 14,000*g* for 1 min at RT, and then discard the flow-through.

82. Place the spin column in a new 2 ml collection tube, and centrifuge at 14,000*g* for 2 min to remove any residual ethanol.

83. Transfer the spin column to a sterile 1.5 ml Eppendorf tube, pipette 15 µl of pre-heated Buffer EB onto the centre of the spin column, and incubate on the bench top for 1 min.

84. Centrifuge at 14,000*g* for 1 min at RT .

PAUSE POINT: Bisulfite converted DNA can be stored at -20°C for up to one week.

Library enrichment TIMING: ~1h30m

85. Setup the PCR reaction according to the following scheme:

Reagent	Volume (µl)
Bisulfite converted DNA from Step 84	15
NEBNext Universal PCR Primer (10 µM)	5
NEBNext Index *n* Primer (10 µM)	5
dNTPs (10mM each)	1.25
Pfu Turbo Cx Reaction Buffer (10X)	5
Pfu Turbo Cx Hotstart DNA Polymerase (2.5 U/µl)	1
Nuclease-free water	17.75

CRITICAL STEP: The concentration of NEBNext primers has changed from 25 µM to 10 µM. If using NEBNext Multiplex Oligos for Illumina (Set 1) lots 0071402, or 0081407, scale the volume of each primer to 2 µl, and increase the water volume to 23.75 µl.

86. Mix thoroughly by pipetting up and down the entire reaction volume at least 6 times.

87. Perform the PCR reaction following the conditions described in the table below:

Step	Temperature (°C)	Time
1	95	5 min
	98	30 sec
2-46	65	30 sec
	72	30 sec
47	72	5 min
48	4	Hold

88. During incubation, repeat Step 8.

89. Once the thermal cycler reaches 4°C, transfer the whole reaction volume (50 µl) to a sterile 1.5 ml Eppendorf tube.

90. Purify using 1 volume (50 µl) of Agencourt AMPure XP Beads (see Box 1). Elute in 15 µl nuclease-free water.

91. Transfer 15 µl of the cleared supernatant from the tube **[AU: Do you mean the 15ul eluted DNA from Step 90?]**to a sterile 1.5 ml Eppendorf tube.

Library inspection TIMING: ~1h

92. Measure the concentration of the final library using a Qubit fluorometer. ?TROUBLESHOOTING

93. Visualize the fragments distribution by running 1 µl of the library on the Fragment Analyzer (or Agilent Bioanalyzer). Library should appear as a peak in the range of ~200-400 bp (see Fig. 3e). ?TROUBLESHOOTING

Library sequencing and reads filtering TIMING: variable

94. Load the library on a high-throughput sequencer. For Illumina platforms, we recommend to pool the MAB-seq library with other sequencing libraries (e.g. RNA-seq or ChIP-seq) to balance the CG-content across each sequencing lane or to use a low density cluster (about 70% of the maximum recommended cluster density). Use all the sequencer recommendations to obtain high quality score sequencing reads (e.g. freshly prepared reagents, flowcell not overloaded, etc).

95. Filter the sequencing reads for low quality score reads and clip out the adapter sequencer. For Illumina sequencing reads you can use the FastX-Toolkit (http://hannonlab.cshl.edu/fastx_toolkit/). Issue the commands:

$ *fastq_quality_filter –Q 33 -q 30 -p 90 -i <input FastQ file> -o <filtered FastQ file>*

$ *fastx_clipper -a <adapter sequence> -n -l 25 -Q 33 -M 10 -i <filtered FastQ file> -o <filtered clipped FastQ file>*

MAB-seq data analysis TIMING: variable

96. Download and install the BSMAP[34] software.

97. Download the reference genome sequence for your organism of interest, and the Lambda phage genome (GenBank: J02459) in FASTA format.

98. Map reads to the reference genome for the organism of interest using BSMAP. Choose the appropriate command line, depending on the type of MAB-seq protocol performed (option A for whole-genome MAB-seq or option B for RRMAB-seq).

(A) Whole-genome MAB-seq.

i) Issue the command:

$ bsmap -a *<filtered clipped FastQ file>* -d *<reference.fasta>* -s 16 -v 0.1 -n 1 -q 30 -r 0 -u -o *<output.sam>*

(B) Reduced representation MAB-seq.

i) Issue the command:

$ bsmap -a *<FastQ file>* -d *<reference.fasta>* -s 12 –D C-CGG -v 0.1 -n 1 -q 30 -r 0 -u -o *<output.sam>*

99. Post-process the output SAM file, using the *bsmap_postprocess* script (Supplementary Data 1). The script will produce a FastQ file containing all the unmappable reads from the SAM file, and a SAM file containing only the successfully mapped reads:

$ bsmap_postprocess -f *<Unmapped FastQ file>* -s *<Mapped SAM file>* *<output.sam>*

100. Map the unmapped reads from the FastQ file from Step 99 to the Lambda phage reference genome using BSMAP. Choose the appropriate command line, depending on the type of MAB-seq protocol performed (option A for whole-genome MAB-seq or option B for RRMAB-seq).

(A) Whole-genome MAB-seq.

i) Issue the command:

$ bsmap -a <*Unmapped FastQ file*> -d <*J02459.fasta*> -s 16 -v 0.1 -n 1 -q 30 -r 0 -o <*lambda.sam*>

(B) Reduced representation MAB-seq.

i) Issue the command:

$ bsmap -a <*Unmapped FastQ file*> -d <*J02459.fasta* > -s 12 –D C-CGG -v 0.1 -n 1 -q 30 -r 0 -o <*lambda.sam*>

101. (Optional) Perform methylation call on Lambda non-CpG sites (using file from Step 100) by using the *methratio.py* utility (provided in the BSMAP package), and calculate bisulfite conversion rate:

$ methratio.py -o non*CpG_lambda_methratio.txt* -d <*J02459.fasta*> -z -m 5 -x CHG,CHH <*lambda.sam*>

$ cut -f 7,8 *nonCpG_lambda_methratio.txt* | perl -e 'while(<>) { chomp(); my @x=split /\t/; $c += $x[0]; $tot += $x[1]; } print "Bisulfite conversion rate: " . sprintf("%.3f", 1 - $c/$tot) . "\n\n";'

?TROUBLESHOOTING

102. Perform methylation call on Lambda DNA CpG sites (using file from Step 100) by using the *methratio.py* utility, and calculate M.SssI fail rate:

$ methratio.py -o *CpG_lambda_methratio.txt* -d <*J02459.fasta*> -z -m 5 -x CG <*lambda.sam*>

$ cut -f 7,8 *CpG_lambda_methratio.txt* | perl -e 'while(<>) { chomp(); my @x=split /\t/; $c += $x[0]; $tot += $x[1]; } print "M.SssI fail rate: " . sprintf("%.3f", 1 - $c/$tot) . "\n\n";'

?TROUBLESHOOTING

103. Perform methylation call on reference genome (using SAM file containing the successfully mapped reads from Step 99) using the *methratio.py* utility:

$ methratio.py -o *CpG_organism_methratio.txt* -d *<reference.fasta>* -z -m 20 -x CG *<Mapped SAM file>*

104. Determine high-confidence 5fC/5caC sites (using the file from Step 103) using the *5fcaC_bintest* script (Supplementary Data 2). The script will perform a binomial test on all cytosines, and will return a BED file with the high-confidence 5fC/5caC sites:

$ 5fcaC_bintest -p *<M.SssI fail rate>* -c 0.001 -o 5fcaC_sites.bed *CpG_organism_methratio.txt*

TIMING

Steps 1-11, gDNA Methylation: ~9h

Steps 12, gDNA Shearing (Whole-genome MAB-seq protocol): ~1h

Steps 12, gDNA Digestion (RRMAB-seq protocol): ~4h

Step 13-15, Fragmented/Digested gDNA purification: ~45m

Steps 16-21, End repair of gDNA: ~1h30m

Steps 22-27, dA-Tailing of end-repaired DNA: ~1h30m

Steps 28-35, Adapter ligation: ~1h30m

Steps 36-39, Methylation of DNA termini: ~9h

Steps 40-56, Size-selection of DNA: ~2h

Steps 57-60, Bisulfite conversion of size-selected DNA: ~15h

Steps 61-84, Clean-up of bisulfite converted DNA: ~1h30m

Steps 85-91. Library enrichment: ~1h30m

Steps 92-93, Library inspection: ~1h

Steps 94-95, Library sequencing and reads filtering**:** variable

Steps 96-104, MAB-seq data analysis: variable

ANTICIPATED RESULTS

We have successfully employed this protocol to prepare MAB-seq libraries from mouse embryonic stem cells[24] and also from other cell lines. In our experience, typical whole-genome MAB-seq library yields are between 50 ng and 100 ng, while for reduced representation MAB-seq are between 150 ng and 300 ng of total material. For successful MAB-seq experiments, it is essential to achieve a M.SssI methylation efficiency >97%. Bisulfite conversion efficiency >99.9% is recommended in order to obtain accurate quantification of 5fC/5caC, although this is not mandatory since a lower efficiency would only cause an underestimate of the actual number of 5fC/5caC residues. In a single whole-genome MAB-seq experiment performed in mouse embryonic stem cells, we used 50 Gb of high-quality sequences to call around 300,000 5fC/5caC residues. The average modification level for 5fC/5caC residues should be between 10-20%. **[AU: Please add a reference for these figures]**

TROUBLESHOOTING

Troubleshooting advice can be found in **Table 2.**

Table 2. Troubleshooting

Step	Problem	Possible reason	Solution
92	Low library yield after PCR enrichment	Incomplete adapter ligation, or adapter ligation failure.	Ensure that no ethanol carryover occurs during the purification steps. Check the NEBNext kit expiry date. If performing whole-genome MAB-seq, we sometimes observed that repeating two times the end repair procedure greatly increases the library preparation efficiency.
93	Most of the PCR product is around 121 bp	Adapter dimers carryover is occurring. The gel slice has been cut lower than 200 bp during the size-selection step	Cut the gel slice at least 2mm above the 200 bp ladder's band during the size-selection step (since gDNA shearing produces fragments with a median length of ~150-200 bp, following adapter ligation most of the library fragments will be around 270-320 bp). If the library is visible above the

			adapter dimers, try to run it on a 3% (wt/vol) agarose gel at 80V constant for 35 min, to separate the PCR products from the adapter dimers, and then cut a gel slice from 200 to 400 bp.
101	Bisulfite conversion rate is < 99%	EpiTect Bisulfite kit is not working as expected.	Freshly dissolve the Bisulfite mix right before the experiment. Quantify the starting material after size-selection, and ensure that the total DNA amount is ≤ 2 µg. Try to split the input DNA into two separate bisulfite conversion reactions, or perform two consecutive rounds of bisulfite conversion.
102	M.SssI fail rate is > 5%	M.SssI treatment is not working as expected.	Ensure the purity of the gDNA. Following DNeasy extraction, it is possible to further purify the gDNA using 2.8X volumes of AMPure XP beads. Ensure that ≤ 1

µg of gDNA is used for each *in vitro* methylation reaction. Always prepare fresh buffers and SAM dilutions.

ACKNOWLEDGEMENTS

This work was supported by the Associazione Italiana Ricerca sul Cancro (AIRC) IG 2011 11982.

AUTHOR CONTRIBUTIONS

F.N. conceived the MAB-seq. D.I. and F.N. developed the protocol, the computational methods and analysed the data. A.K. and C.P. contributed to the MAB optimization. S.O. supervised the experiments. F.N, D.I. and S.O. wrote the manuscript. All authors read and approved the final manuscript.

COMPETING FINANCIAL INTERESTS

The authors declare that they have no competing financial interests.

SUPPLEMENTARY MATERIAL.

Supplementary Data 1.

Bsmap_postprocess (Step 99). This custom perl script elaborates the output SAM file derived from Bsmap software and it generates two independent files: a SAM file of the mapped reads and a FastQ file of the unmapped reads.

Supplementary Data 2.

5fcaC_bintest (Step 104). This custom perl script elaborates the methratio file generated in Step 103 and estimates the significance (p-value) of each cytosine to be 5fC/5caC taking into account the M.SssI methylation fail rate calculated in Step 102.

FIGURE LEGENDS

Figure 1. Schematic diagram of the methylation-assisted bisulfite sequencing (MAB-seq) protocol, compared to traditional bisulfite seqyuncing. In MAB-seq, M.SssI CpG methylation protects unmodified cytosine from bisulfite conversion to uracils, meaning that 5fC/5caC can be read as T.

Figure 2. Flowchart of the MAB-seq library preparation protocol.

Figure 3. Quality check steps during library preparation. (a) varying amounts of gDNA were run on a 1% agarose gel. A single sharp band should be visible above the 12 kb marker line demonstrating high quality, non-degraded template DNA. (b) gDNA was sonicated by as described in the Procedure (Step 12). Fragmented gDNA from 5 different cell lines are shown in the picture. 1, mouse embryonic stem cells (mESCs); 2, mouse embryonic fibroblasts (MEFs); 3, HEK-293 human cells; 4, HeLaS3 human cells; 5, Caco2 human cells. (c) Analysis of the M.SssI-dependent *in vitro* CpG methylation. 500 ng of M.SssI treated DNA was digested with the methylation insensitive (MspI, lane2) or sensitive (HpaII, lane3) digestion enzyme and run on 1% agarose gel. 2 µg of freshly extracted DNA was loaded as control in lane 1. (d) Fragmented and methylated gDNA was run on agarose gel and a gel slice corresponding to fragments in the range 200-400 was cut using a sterile scalpel. (e) Final library of MAB-seq was run on Fragment Analyzer instrument and it showed an enrichment around 300 bp. LM, lower marker; UM, upper marker.

BOX 1. DNA clean-up using Agencourt AMPure XP beads.

Purification of DNA using AMPure XP beads is used extensively throughout this protocol. All the steps are performed at room temperature.

1. Add the amount of AMPure XP Beads (specified at each step) to the solution containing DNA to be purified, and mix well by pipetting the entire volume up and down at least 15 times.
2. Let the mixed sample incubate for 15 minutes at room temperature.
3. Place the sample on the magnetic stand for 5 min (or until the solution is clear).
4. Using a P200 pipette, discard the supernatant from the tube leaving approximately 5 µl to avoid disturbing the beads. If processing more than one sample, change the tip after each sample.
5. With the tube remaining on the stand, pipette 200 µl of freshly prepared 80% (v/v) ethanol, without disturbing the beads.
6. Incubate the tube at room temperature for 30 sec, and then discard all of the supernatant from the tube.
7. Repeat steps 5-6 once.
8. Using a P10 pipette, get rid of all the residual ethanol from the bottom of the tube.
9. While keeping the tube on the magnetic stand with open lids, let the sample air-dry at room temperature for 15 min.
10. Remove the tube from the magnetic stand, and resuspend the dried pellet in the specified volume of nuclease-free water. Always consider an additional volume of 2.5 µl (e.g. if the final eluate volume should be 15 µl, resuspend the beads in 17.5 µl). Gently pipette the entire volume up and down at least 10 times.
11. Place the tube in a heat block (or water bath) for 5 min at 37°C to facilitate elution of DNA from the beads.
12. Briefly spin sample down by pulsing in a centrifuge to remove any condensed material from the lid.

13. Place the tube back on the magnetic stand, and incubate at room temperature for 5 min (or until the solution is clear).

14. Transfer the cleared supernatant (exclusive of the additional 2.5 µl used for beads resuspension) from the tube to a sterile tube. **[AU: I think this step should be removed as it is repeated in the Procedure at the appropriate places, i.e. Steps 21, 27, 35, 38, 91?]**

TABLE 1. Comparison between the currently available methods for whole-genome identification of 5fC and 5caC residues. MAB-seq allows single-base resolution 5fC/5caC identification in a single experiment.

method	modification	base resolution	affinity	treatment	sequencings required	reference number
5fC-DP	5fC	no	biotin	biotin chemical labeling	**2** (treated and input)	28
fC-Seal	5fC	no	biotin	5fC red to 5hmC andchemical labeling	**3** (treated, untreated and input)	29
5fC-DIP	5fC	no	antibody	none	**2** (Ab and input)	30
5caC-DIP	5caC	no	antibody	none	**2** (Ab and input)	30
fCAB	5fC	yes	none	5fC chemical protection	**2** (treated and untreated BS-seq)	29
caCAB	5caC	yes	none	5caC chemical protection	**2** (treated and untreated BS-seq)	31
redBS	5fC	yes	none	5fC chemical reduction to 5hmC	**2** (treated and untreated BS-seq)	26
MABseq	5fC/5caC	yes	none	enzymatic protection of unmodified C	**1** (treated)	24

REFERENCES

1. Bestor, T. H. The DNA methyltransferases of mammals. *Human Molecular Genetics* **9,** 2395–2402 (2000).
2. Reik, W. & Walter, J. Genomic imprinting: parental influence on the genome. *Nat Rev Genet* **2,** 21–32 (2001).
3. Li, E. Chromatin modification and epigenetic reprogramming in mammalian development. *Nat Rev Genet* **3,** 662–673 (2002).
4. Okano, M., Bell, D. W., Haber, D. A. & Li, E. DNA methyltransferases Dnmt3a and Dnmt3b are essential for de novo methylation and mammalian development. *Cell* **99,** 247–257 (1999).
5. Smith, Z. D. & Meissner, A. DNA methylation: roles in mammalian development. *Nat Rev Genet* **14,** 204–220 (2013).
6. Neri, F. *et al.* Dnmt3L antagonizes DNA methylation at bivalent promoters and favors DNA methylation at gene bodies in ESCs. *Cell* **155,** 121–134 (2013).
7. Rhee, I. *et al.* DNMT1 and DNMT3b cooperate to silence genes in human cancer cells. *Nature* **416,** 552–556 (2002).
8. Kohli, R. M. & Zhang, Y. TET enzymes, TDG and the dynamics of DNA demethylation. *Nature* **502,** 472–479 (2013).
9. Tahiliani, M. *et al.* Conversion of 5-methylcytosine to 5-hydroxymethylcytosine in mammalian DNA by MLL partner TET1. *Science* **324,** 930–935 (2009).
10. Ito, S. *et al.* Role of Tet proteins in 5mC to 5hmC conversion, ES-cell self-renewal and inner cell mass specification. *Nature* **466,** 1129–1133 (2010).
11. Neri, F. *et al.* Genome-wide analysis identifies a functional association of Tet1 and Polycomb repressive complex 2 in mouse embryonic stem cells. *Genome Biol.* **14,** R91 (2013).
12. Neri, F. *et al.* TET1 is a tumour suppressor that inhibits colon cancer growth by derepressing inhibitors of the WNT pathway. *Oncogene* **34,** 4168–4176 (2015).
13. Neri, F. *et al.* TET1 is controlled by pluripotency-associated factors in ESCs and downmodulated by PRC2 in differentiated cells and tissues. *Nucleic Acids Research* **43,** 6814–6826 (2015).
14. Cortázar, D. *et al.* Embryonic lethal phenotype reveals a function of TDG in maintaining epigenetic stability. *Nature* **470,** 419–423 (2011).
15. He, Y. F. *et al.* Tet-Mediated Formation of 5-Carboxylcytosine and Its Excision by TDG in Mammalian DNA. *Science* **333,** 1303–1307 (2011).
16. Spruijt, C. G. *et al.* Dynamic readers for 5-(hydroxy)methylcytosine and its oxidized derivatives. *Cell* **152,** 1146–1159 (2013).
17. Iurlaro, M. *et al.* A screen for hydroxymethylcytosine and formylcytosine binding proteins suggests functions in transcription and chromatin regulation. *Genome Biol.* **14,** R119 (2013).
18. Bachman, M. *et al.* 5-Formylcytosine can be a stable DNA modification in mammals. *Nat. Chem. Biol.* **11,** 555–557 (2015).
19. Ito, S. *et al.* Tet Proteins Can Convert 5-Methylcytosine to 5-Formylcytosine and 5-Carboxylcytosine. *Science* **333,** 1300–1303 (2011).
20. Harris, R. A. *et al.* Comparison of sequencing-based methods to profile DNA methylation and identification of monoallelic epigenetic modifications. *Nat Biotechnol* **28,** 1097–1105 (2010).
21. Gu, H. *et al.* Genome-scale DNA methylation mapping of clinical samples at

single-nucleotide resolution. **7,** 133–136 (2010).
22. Huang, Y. *et al.* The Behaviour of 5-Hydroxymethylcytosine in Bisulfite Sequencing. *PLoS ONE* **5,** e8888 (2010).
23. Yu, M. *et al.* Base-Resolution Analysis of 5-Hydroxymethylcytosine in the Mammalian Genome. *Cell* **149,** 1368–1380 (2012).
24. Neri, F. *et al.* Single-Base Resolution Analysis of 5-Formyl and 5-Carboxyl Cytosine Reveals Promoter DNA Methylation Dynamics. *CellReports* **10,** 674–683 (2015).
25. Xu, M., Kladde, M. P., Van Etten, J. L. & Simpson, R. T. Cloning, characterization and expression of the gene coding for a cytosine-5-DNA methyltransferase recognizing GpC. *Nucleic Acids Research* **26,** 3961–3966 (1998).
26. Booth, M. J., Marsico, G., Bachman, M., Beraldi, D. & Balasubramanian, S. Quantitative sequencing of 5-formylcytosine in DNA at single-base resolution. *Nat Chem* **6,** 435–440 (2014).
27. Jin, S.G., KAdam, S. & Pfeifer, G.P. Examination of the specificity of DNA methylation profiling techniques towards 5-methylcytosine and 5-hydroxymethylcytosine. *Nucleic Acids Research* **38,** e125 (2010)
28. Raiber, E.-A. *et al.* Genome-wide distribution of 5-formylcytosine in embryonic stem cells is associated with transcription and depends on thymineDNA glycosylase. *Genome Biol.* **13,** R69 (2012).
29. Song, C.-X. *et al.* Genome-wide Profiling of 5-Formylcytosine Reveals Its Roles in Epigenetic Priming. *Cell* **153,** 678–691 (2013).
30. Shen, L. *et al.* Genome-wide Analysis RevealsTET- and TDG-Dependent 5-Methylcytosine Oxidation Dynamics. *Cell* **153,** 692–706 (2013).
31. Lu, X. *et al.* Chemical modification-assisted bisulfite sequencing (CAB-Seq) for 5-carboxylcytosine detection in DNA. *J. Am. Chem. Soc.* **135,** 9315–9317 (2013).
32. Booth, M. J. *et al.* Quantitative Sequencing of 5-Methylcytosine and 5-Hydroxymethylcytosine at Single-Base Resolution. *Science* **336,** 934–937 (2012).
33. Incarnato, D., Krepelova, A. & Neri, F. High-throughput single nucleotide variant discovery in E14 mouse embryonic stem cells provides a new reference genome assembly. *Genomics* **104,** 121–127 (2014).
34. Xi, Y. & Li, W. BSMAP: whole genome bisulfite sequence MAPping program. *BMC Bioinformatics* **10,** 232 (2009).