Published online 14 January 2022

NAR Genomics and Bioinformatics, 2022, Vol. 4, No. 1 **1**
https://doi.org/10.1093/nargab/lqab119

Low guanine content and biased nucleotide distribution in vertebrate mtDNA can cause overestimation of non-CpG methylation

Takashi Okada [1], Xin Sun[2], Stephen McIlfatrick[1] and Justin C. St. John [1,*]

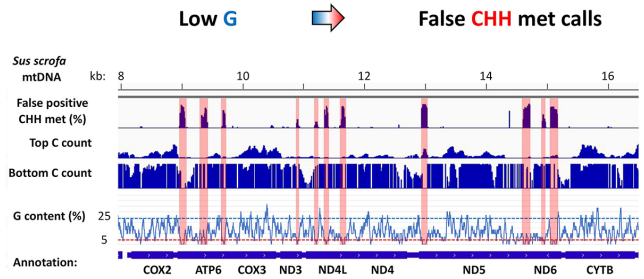
[1]Mitochondrial Genetics Group, Robinson Research Institute and School of Biomedicine, Faculty of Health and Medical Sciences, The University of Adelaide, Adelaide, SA 5000, Australia and [2]Centre for Cancer Research, Hudson Institute of Medical Research, Clayton, VIC 3168, Australia

Received August 26, 2021; Revised November 24, 2021; Editorial Decision December 07, 2021; Accepted January 09, 2022

ABSTRACT

Mitochondrial DNA (mtDNA) methylation in vertebrates has been hotly debated for over 40 years. Most contrasting results have been reported following bisulfite sequencing (BS-seq) analyses. We addressed whether BS-seq experimental and analysis conditions influenced the estimation of the levels of methylation in specific mtDNA sequences. We found false positive non-CpG methylation in the CHH context (fpCHH) using unmethylated *Sus scrofa* mtDNA BS-seq data. fpCHH methylation was detected on the top/plus strand of mtDNA within low guanine content regions. These top/plus strand sequences of fpCHH regions would become extremely AT-rich sequences after BS-conversion, whilst bottom/minus strand sequences remained almost unchanged. These unique sequences caused BS-seq aligners to falsely assign the origin of each strand in fpCHH regions, resulting in false methylation calls. fpCHH methylation detection was enhanced by short sequence reads, short library inserts, skewed top/bottom read ratios and non-directional read mapping modes. We confirmed no detectable CHH methylation in fpCHH regions by BS-amplicon sequencing. The fpCHH peaks were located in the D-loop, *ATP6*, *ND2*, *ND4L*, *ND5* and *ND6* regions and identified in our *S. scrofa* ovary and oocyte data and human BS-seq data sets. We conclude that non-CpG methylation could potentially be overestimated in specific sequence regions by BS-seq analysis.

GRAPHICAL ABSTRACT



INTRODUCTION

Since early studies (1–3) observed DNA methylation in mammalian mitochondrial DNA (mtDNA), there has been considerable interest in its role in regulating mtDNA gene expression and replication for more than four decades (4–7). The discovery of some DNA methyltransferases in mitochondria has been reported in different cell types and tissues in human and mouse, prompting epigenetic studies of the mitochondrial genome (5,8–10). However, the observed mtDNA methylation levels are rather low (1–5%) compared to nuclear DNA (1,2), questioning its significance and related biological function. In addition, the sensitivity and reproducibility of the various methylation detection methods and conflicting results demonstrated by different reports leaves no consensus about mammalian mtDNA methylation (11). Among these studies, bisulfite sequencing (BS-seq) has been frequently used to investigate mtDNA methylation. Several studies showed positive and significant levels of methylation in mammalian mtDNA, demonstrated by BS-amplicon sequencing and BS-pyrosequencing (7,8,10,12). More recently, significantly high levels (up to 44% on average) of non-CpG methylation in mammalian mtDNA have been reported by utilizing BS-seq combined with deep sequencing (13–15). However, other reports claimed no or insignificant levels of mtDNA methylation (<1%) detected by BS-seq methods, suggesting

*To whom correspondence should be addressed. Email: jus.stjohn@adelaide.edu.au

that the associated secondary structures of mtDNA contributed, at least in part, to incomplete BS-conversion and resulted in overestimation of mtDNA methylation (16–20).

Bisulfite sequencing (BS-seq) is one of the 'gold standard' methods to assess levels of methylation in epigenetic studies. BS treatment converts unmethylated cytosines to uracil (read as thymine after DNA amplification), whilst methylated cytosines are protected from BS conversion (21). This method has been used to study the methylation status of target sequence regions by BS-amplicon sequencing or BS-pyrosequencing and genome wide methylation in combination with Next Generation Sequencing (NGS). BS-seq NGS has the ability to generate outputs at the level of single base pair (bp) resolution across the whole genome (22). However, BS-seq does not distinguish between 5 methylcytosine (5mC) and 5 hydroxymethylcytosine (5hmC), and BS-seq library construction can be associated with biased representation of sequences caused by DNA fragmentation and biased PCR amplification (23,24). Furthermore, it has been shown that C-rich sequences are underrepresented in whole genome bisulfite sequencing (WGBS) data, which is frequently dependent on the library construction methods used (25).

BS-seq aligners map BS-converted reads to a reference sequence and calculate the percentage of methylation at each cytosine from the number of converted and unconverted Cs in the mapped reads (22). As cytosine methylation is not symmetrical, the two reference DNA strands need to be mapped separately. Regarding DNA strands obtained for the BS-seq reads, there are two types of BS-seq libraries, directional and non-directional. For example, conventional BS-seq directional libraries use adapter ligated DNA fragments for BS-treatment, which resulted in only original top (OT) or bottom (OB) strand sequence reads (26). On the other hand, non-directional libraries constructed through post-BS-treatment adapter ligation and subsequent PCR amplification resulted in sequencing of all four DNA strands, including strands complementary to the original top or bottom strands (CTOT or CTOB) (27).

In general, two types of algorithms are used for read mapping by BS-seq aligners, namely the 'wild card' and the 'three letter' approaches (28). In the 'wild card' approach, cytosines in the reference sequences are replaced with the wild card 'Y' (C or T), allowing for the alignment of Cs (methylated) and Ts (unmethylated). This approach is employed by BS-seq aligners such as BSMAP (29), GSNAP (30) and Last (31). In the 'three letter' approach, all Cs in the reference and sequence read data are converted to Ts, thus reducing the nucleotides in the sequence to A/T/G. BS-seq aligners such as Bismark (32) and BS-Seeker2 (33) use this algorithm. Many BS-seq aligners have been tested for their mapping accuracy, estimation of levels of methylation, and computational performance at a global level (34–36). However, it is still largely unknown whether BS-seq aligners could estimate the levels of methylation correctly for specific types of sequences and whether contradictory levels of mtDNA methylation are associated with BS-seq aligners and conditions employed during analysis.

Therefore, we aimed to investigate mtDNA methylation by BS-seq NGS using methylated and non-methylated *Sus scrofa* mtDNA controls and publicly available BS-seq data sets under various analysis conditions. We hypothesized that BS-seq aligner settings and specific types of sequences might influence the estimation of the levels of mtDNA methylation in BS-seq data analysis. We found that biased nucleotide distribution in vertebrate mitochondrial genomes caused the detection of false positive non-CpG methylation when using BS-seq aligners, especially in the asymmetric CHH context. Biased mtDNA nucleotide distribution was also associated with skewed representation of BS-seq reads, indicated by significantly low abundance of reads corresponding to the low guanine content sequences on the top/plus strand. Our data provide evidence for the overestimation of non-CpG methylation in vertebrate mtDNA by BS-seq analysis and support no or insignificant levels of mtDNA methylation in the data sets analysed.

MATERIALS AND METHODS

Isolation of mitochondria from *Sus scrofa* oocytes

Mitochondrial fractions were prepared from *S. scrofa* oocytes, as described (37). Briefly, *in vitro* matured metaphase II oocytes were denuded from cumulus oocyte complexes, and 15 oocytes were resuspended in 5 ml of mitochondrial isolation buffer (20 mM HEPES pH 7.6, 220 mM mannitol, 70 mM sucrose, 1 mM EDTA) containing 2 mg/ml BSA. Oocytes were homogenised by 10 strokes of a drill-fitted Potter-Elvehjem tissue grinder set (VWR International, PA, USA) on ice. The oocyte homogenate was centrifuged at 800 g for 10 min at 4°C to remove cell debris. Mitochondrial supernatants were centrifuged at 10 000 g for 20 min at 4°C to pellet the mitochondria. The mitochondrial pellet was resuspended in 700 μl of mitochondrial isolation buffer, then further centrifuged again at 10 000 g for 20 min. The supernatant was removed, and the mitochondrial pellet was resuspended in 5 μl of mitochondrial isolation buffer and used for DNA extraction.

Purification of mitochondrial DNA (mtDNA) from *Sus scrofa* ovarian tissue

Mitochondrial DNA was purified from *S. scrofa* ovarian tissue, as described (7). Briefly, about 100 mg of ovarian tissue was homogenized in 5 ml of mitochondrial isolation buffer containing 2 mg/ml BSA at 4°C using the Potter-Elvehjem tissue grinder set (VWR International, PA, USA) for 50 repetitions. The homogenate was centrifuged at 800 g for 10 min at 4°C to remove cell debris and nuclei. The supernatant was then centrifuged at 10 000 g for 20 min at 4°C to pellet the mitochondrial fraction. To further remove nuclear DNA, the pellet was resuspended in 200 μl of mitochondrial isolation buffer with 2 μl of Ambion TURBO DNaseI (Thermo Fisher Scientific, MA, USA) and incubated at 37°C for 30 min. Then, 1 ml of mitochondrial isolation buffer with BSA was added and the suspension was centrifuged at 10 000 g for 20 min at 4°C. The supernatant was discarded, and the mitochondrial pellet was resuspended in 200 μl of lysis buffer (50 mM Tris–HCl, pH 8.0, 10 mM EDTA, and 1% SDS) with 1 μl of proteinase K (20 mg/ml). The suspension was then incubated at 50°C

for 60 min. mtDNA was purified from the suspension using the DNeasy Blood & Tissue Kit (QIAGEN, VIC, Australia), according to the manufacturer's instructions. Purified mtDNA was assessed by agarose gel and quantitated by NanoDrop ND-2000 (Thermo Fisher Scientific, MA, USA).

Linearization of DNA and bisulfite sequencing (BS) library preparation

Purified ovarian mtDNA, lambda DNA (dam-: dcm-, Cat No. SD0021, Thermo Fisher Scientific, MA, USA), and pBR322 plasmid DNA (Cat No. N3033S, New England Biolabs, MA, USA) were linearized by EcoRV restriction enzyme digestion, which cut once in both *S. scrofa* mtDNA and pBR322 and 21 times in lambda DNA. Two to five μg of DNA were digested by EcoRV (New England Biolabs, MA, USA) for 4 h at 37°C. Then, digested DNA was subjected to 0.8% agarose gel electrophoresis, and mtDNA and pBR322 bands were excised from the gel for purification using the ISOLATE II PCR and Gel Kit (Meridian Bioscience, TN, USA). Digested lambda DNA was purified using the ISOLATE II PCR and Gel Kit directly after EcoRV digestion. Digested and purified DNA was validated by loading together with undigested DNA on an agarose gel. DNAs were also linearized by sonication using Bioruptor® Plus (Diagenode Inc, NJ, USA). Briefly, 1.5 to 3 μg of DNA in 50 μl of TE buffer (10 mM Tris pH 8.0, 1 mM EDTA) were added to 0.5 ml PCR tubes and sonicated for 30 s 4 times at 30 s intervals in a 4°C water bath. A few μl of sonicated DNA were loaded on an agarose gel for validation. After treatment and purification, the concentration of DNA was measured by Qubit Assay (Thermo Fisher Scientific, MA, USA). mtDNA, pBR322 and lambda DNAs were mixed in 10:5:1 molar ratios for each treatment sample and used for bisulfite sequencing library preparation. The DNA mix (10 μl containing 10–50 ng DNA) for each of the three treatments (undigested, EcoRV digested and sonicated) was used for bisulfite treatment and NGS library construction using the Pico Methyl-Seq Library Prep Kit (ZYMO RESEARCH, CA, USA), according to the manufacturer's instructions. Triplicate BS sample libraries were prepared for each treatment. NGS libraries were sequenced by the Illumina NextSeq platform using 150 bp paired-end sequencing v2.5 chemistry.

Generation of negative and positive controls for mtDNA bisulfite sequencing

To generate experimental controls for *S. scrofa* mtDNA bisulfite sequencing, mtDNA was amplified by long PCR to generate four overlapping fragments using the Platinum Taq High Fidelity system (Thermo Fisher Scientific, MA, USA) and four primer sets (Supplementary Table S1), as described (7). The resultant mtDNA amplicons were purified from excised gel bands using the QIAquick Gel Extraction Kit (QIAGEN, VIC, Australia). Four mtDNA fragments (Ssc_MT_frag1–4 in Supplementary Table S1) were mixed at the same molar ratios and used as a negative control for mtDNA bisulfite sequencing.

Similarly, the positive control was prepared by treating mixed mtDNA fragments with CpG methyltransferase *M.SssI* (Cat. No. M0226S, New England Biolabs, MA, USA), as described (7). Briefly, 1.6 μg of combined mtDNA fragments were treated with 16 units of *M.SssI* with 1× NEB buffer 2 and 160 μM of *S*-adenosylmethionine in a 100 μl reaction at 37°C for 3 h, followed by 20 min incubation at 65°C to terminate the reaction. Treated mtDNA was then purified by using QIAquick PCR Purification Kit (QIAGEN, VIC, Australia). CpG methylation of mtDNA was validated by digesting the fragments with MspI (methylation sensitive) or HpaII (insensitive) restriction enzymes (New England Biolabs, MA, USA) and checked on an agarose gel. Both negative and positive mtDNA controls were used for bisulfite sequence library preparation and Illumina sequencing, as described above.

Preparation of WGBS libraries

DNA was extracted from a pool of 40 metaphase II oocytes or mitochondrial fraction isolated from oocytes by using the QIAamp DNA Micro Kit (QIAGEN, VIC, Australia), according to the manufacturer's instructions. DNA was eluted in 20 μl of elution buffer. Preparation of WGBS libraries and NGS Illumina sequencing were conducted by SAHMRI Genomic Centre (Adelaide, SA, Australia). Briefly, 10 μl of total DNA solution were used for bisulfite treatment and NGS library construction using the Pico Methyl-Seq Library Prep Kit (ZYMO RESEARCH, CA, USA), according to the manufacturer's instructions. The Illumina NovaSeq S1 flow cell was used for oocyte total DNA WGBS library using 100 bp paired end sequencing chemistry and Illumina MiSeq was used for oocyte mitochondrial DNA BS-seq library using 75 or 150 bp paired end sequencing chemistry.

mtDNA bisulfite sequence data analysis

mtDNA bisulfite sequence data were analysed following the procedure, as described (7) with minor modifications. Firstly, adaptors and poor-quality reads were cleaned from raw sequences using the TrimGalore program v0.4.2 (https://github.com/FelixKrueger/TrimGalore) in the paired-end mode with the default adaptor trimming option and additional 10 bp trimming for both the 5' and 3' ends. In addition, various trimming sizes were used to investigate the effect of trimmed read sequence size; and recommended trimming sizes were used (https://rawgit.com/FelixKrueger/Bismark/master/Docs/Bismark_User_Guide.html) for publicly available BS-seq data sets depending on the type of BS-seq libraries.

The quality filtered and trimmed sequences were mapped to the *Sus scrofa* mitochondrial genome sequence (Accession No. NC_000845.1) using Bismark Package v0.22.3 (32). The following Bismark options were applied: *–bowtie2*, *-N* 1, *-L* 20, *–non_directional*, *–score_min* L,0,0 or L,0,-0.2 or L,-0.6,-0.6 for paired-end sequence data. Paired end reads were also analyzed by using the single-end read mapping mode with the same options. Output bam files for each sample set were deduplicated using the Bismark pack-

age '*deduplicate_bismark*' function for removal of PCR duplicates. Deduplicated mapped reads were used to extract methylation coverage for CpG, CHG and CHH contexts using the '*bismark_methylation_extractor*' with options (*–cytosine_report* and *–CX*). Cytosine sites covered by a minimum of ten reads were kept for further analysis. Methylation % data for each methylation sequence context were extracted from the CX report file and formatted as a bedGraph file using an in-house script for visualization by Integrative Genome Viewer (https://software.broadinstitute.org/software/igv/). Alternatively, methylation coverage data were extracted using the MethylDackel v0.5.0 program (https://github.com/dpryan79/MethylDackel) with minumum coverage of ten reads. SAMtools (38) were used to manipulate output bam files and obtain mapped sequence read statistics using utilities: '*sort*', '*merge*', '*view*', '*index*', '*stats*', '*idxstats*', '*depth*'.

Some BS-seq data sets were analysed by BSMAP version 2.90 (29) with aligner '*bsmap*' options: *-n* 0 for directional BS-seq libraries or 1 for non-directional libraries, *-s* 16, *-m* 0/50/100 or 150, and methyl extractor '*methratio.py*' options: *-u* and *-p*. BS-Seeker2 (33) was also used for *S. scrofa* mtDNA BS-seq data with aligner '*bs_seeker2-align.py*' options: *–aligner = bowtie2*, *-t* N for directional BS-seq libraries or Y for non-directional libraries, *-m* 2, *-I* 0 or 150 and methyl extractor '*bs_seeker2-call_methylation.py*' option: *-r* 10. Similar to the Bismark analysis, cytosine sites covered by a minimum of ten reads were kept for further analysis.

WGBS data analysis

WGBS data were analysed by following the procedure, as described (7) with minor modifications. Firstly, adaptors and poor-quality reads were cleaned from raw sequences using the TrimGalore program v0.4.2 (https://github.com/FelixKrueger/TrimGalore) in the paired-end mode with default adaptor trimming and 10 bp trimming for both ends. The quality filtered and trimmed sequences were mapped to the *Sus scrofa* reference genome sequence Sscrofa11.1 (Accession No. GCF_000003025.6) using Bismark Package v0.22.3 (32) and the following options: *–bowtie2*, *-N* 1, *-L* 20, *–non_directional*, *–score_min* L,0,-0.2 for paired-end sequence data. Output bam files for each data set were deduplicated using '*deduplicate_bismark*' function for removal of PCR duplicates. Then, deduplicated mapped reads were used to extract methylation coverage for the CpG, CHG and CHH contexts using by '*bismark_methylation_extractor*' with options *(–cytosine_report* and *–CX*). Genome wide methylation coverage data were visualized by Integrative Genome Viewer.

Analysis of bisulfite sequencing data in human

BS-seq data for human and mouse used in this study were obtained from the NCBI Sequence Read Archive (https://www.ncbi.nlm.nih.gov/sra) and listed in Supplementary Table S4. WGBS data were analysed using the procedure, as described above, aligning to the *Homo sapiens* mitochondrial genome sequence (Accession No. NC_0012920.1) as a reference.

Bisulfite amplicon sequencing

Induced methylated and non-methylated *Sus scrofa* mtDNA and purified mtDNA from ovary, as described above, were bisulfite treated using the EZ DNA Methylation-Gold Kit (Zymo Research, CA, USA). Bisulfite sequence primers were designed for four fpCHH regions (R1 to R4, Figure 4) by MethPrimer (39) and are listed in Supplementary Table S1. Target fpCHH regions were amplified in an MJ Research PTC-200 Thermal Cycler (Marshall Scientific, NH, USA) using Platinum™ Taq DNA Polymerase High Fidelity (Cat. No. 11304011, Thermo Fisher Scientific, MA, USA) by denaturing at 95°C for 3 min followed by 40–45 cycles of 95°C for 15 s, 55–59°C for 20 s, 65°C for 30 s. Triplicate technical replicates for MET and noMET mtDNAs and triplicate biological replicates for ovary mtDNA were used for bisulfite treatment and PCR amplification. To distinguish each replicate PCR amplicon, unique 3-bp nucleotides were added to the 5' end of the forward primers (Supplementary Table S1). Amplified PCR products for R1 to R4 from each sample were pooled and purified using the ISOLATE II PCR and Gel Kit (Meridian Bioscience, TN, USA). NGS libraries were made from purified and pooled amplicons using the TruSeq Nano DNA Library Prep Kit (Illumina Inc, CA, USA) and sequenced with Illumina MiSeq Nano 150 bp paired end sequencing chemistry. NGS library construction and NGS Illumina MiSeq sequencing were conducted by the Australian Genome Research Facility (VIC, Australia). Amplicon sequence data were analysed by Bismark to call methylation, as mentioned above and graphs were prepared using RStudio (http://www.rstudio.com/).

Statistical analysis and data visualization

Statistical analyses including ANOVA test, Tukey test and Pearson's correlation were conducted using RStudio and results were visualized by bar graph, histogram and scatter plot using *ggplot2* (40).

RESULTS

The circularity of mtDNA does not affect bisulfite conversion

In order to determine whether circular DNA requires linearisation prior to bisulfite treatment, we used a mixture of plasmid DNA pBR322, lambda DNA and *Sus scrofa* ovarian mtDNA with and without the use of linearisation. BS-Seq libraries were made from: (i) non-linearised; (ii) EcoRV digested and (iii) sonicated mtDNA. Non-linearised/undigested pBR322 showed significantly higher levels of non-converted cytosine/methylation in all three DNA methylation contexts, namely CpG, CHG and CHH methylation (Figure 1). This was found across the entire pBR322 sequence, whilst linearised pBR322 did not show any significant level of non-converted cytosines (<3%) except for a few CHG sites (Supplementary Figure S1A). This suggests that the highly supercoiled structure of pBR322 inhibited bisulfite conversion resulting in ∼15% non-converted cytosines.

On the other hand, undigested lambda DNA and mtDNA did not show any significant differences in the

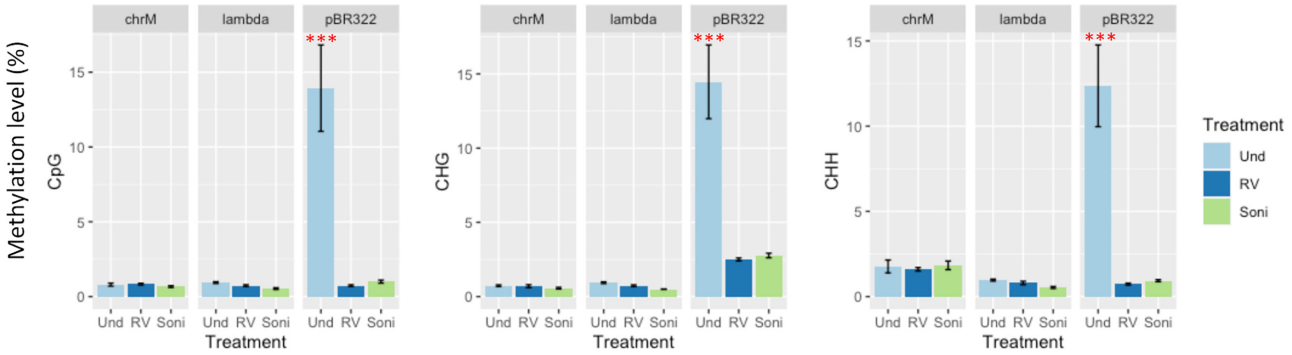


Figure 1. The effect of linearisation on the levels of cytosine conversion after bisulfite treatment. *Sus scrofa* ovary mtDNA (chrM), lambda DNA and plasmid DNA pBR322 were linearized by *EcoR*V digestion (RV) or sonication (Soni) prior to bisulfite conversion. Non-linearised DNAs (Und) were also used as a control. Bisulfite sequence data were analysed by Bismark, and the levels of unconverted cytosine are shown as % levels of methylation for the three methylation DNA contexts (CpG, CHG and CHH). Triplicate technical replicates were used for each treatment. Variation in levels of methylation were analysed by ANOVA test and significant differences between individual samples were determined using Tukey test (***$P < 0.001$). Error bars are S.D.

levels of bisulfite conversion from linearised samples (Figure 1 and Supplementary Figure S1B). Therefore, with the method used for mitochondrial enrichment, DNA extraction and BS-Seq library construction, we found no evidence to support inhibition of bisulfite treatment when mtDNA maintains its circularity. However, it is possible that the large circular mtDNA molecule is nicked during the purification process; or its more loosely coiled structure is not affected when bisulfite conversion is undertaken with the Pico Methyl-Seq Library Prep Kit. Consequently, we excluded linearisation of *S. scrofa* mtDNA in our remaining analyses.

Detection of CHH methylation in non-methylated mtDNA

We identified a peak of CHH methylation in the region spanning nt 12 940 to 13 050 of the porcine ovarian mitochondrial genome (NC_000845.1; Supplementary Figure S1B). To verify the presence of methylation in this region, we generated PCR products that spanned the whole *Sus scrofa* mitochondrial genome. These products were then treated with and without the CpG methyltransferase *M.SssI* to generate methylated (MET) and non-methylated (noMET) mtDNA samples, respectively. As anticipated, the MET sample showed extensive methylation in the CpG context demonstrating the potential for mtDNA to undergo DNA methylation (Supplementary Figure S2B). However, several peaks associated with CHH methylation were observed in both the MET and noMET samples including the region encompassing nt 12 940 to 13 050. To determine if this was an artifact associated with the BS-seq aligner, we investigated BS-seq data from the MET and noMET samples using three different BS-seq aligners, namely BSMAP (29), BSSeeker2 (33) and Bismark (32). BSMAP uses the 'wild card' approach in the read alignment algorithm, whilst BSSeeker2 and Bismark employ the 'three letter' approach, as mentioned above. Default settings employing the non-directional mapping mode for each BS-seq aligner were used for analysis. As expected, for the MET data, high levels of CpG methylation throughout the mitochondrial genome were detected in 762 out of the 784 potential CpG sites in both the top and bottom strands (Figure 2A). CpG methylation patterns detected by three BS-seq aligners were highly similar with only minor differences. Although the CpG methyltransferase does not methylate in the CHH context, we found peaks of CHH methylation again using all three BS-seq aligners for the MET data (Figure 2A). Similarly, we saw these CHH methylation peaks in the noMET, non-methylated mtDNA sample, data analysed by all three BS-seq aligners (Supplementary Figure S2A), suggesting false positive CHH (fpCHH) methylation calls made by each aligner. Since Bismark revealed fewer and lower fpCHH peaks than the other aligners (Figure 2A and S2A) and has more analytical options than the other BS-seq aligners (29,32,33), we performed further analysis using Bismark to investigate fpCHH methylation. We employed various alignment stringencies, including a value for accepting the number of sequence mismatches in the read alignment. However, tighter stringency did not improve the rate of fpCHH compared to the default Bismark setting (Figure 2B). Consequently, there might be specific sequences, which interfere with correct methylation calls when using BS-seq aligners and result in the detection of fpCHH methylation.

Low guanine content and biased nucleotide distribution associate with fpCHH

In order to account for the presence of fpCHH methylation, we assessed the sequences associated with the fpCHH regions. As fpCHH methylation was commonly detected between nt 12 940 and 13 050 in *Sus scrofa* mtDNA (Figure 2, Supplementary Figures S1 and S2), we analysed this region first. Within this 110 bp region, we found a highly biased nucleotide composition, showing only two guanine nucleotides in the top/plus strand (Figure 3A). When this region of sequence is bisulfite converted, the top/plus strand sequence would be extremely AT-rich. In contrast, the bottom/minus strand sequence has only two cytosines in this region, which would result in almost identical sequences to the original after bisulfite conversion. We hypothesized that this biased nucleotide distribution might have caused confused methylation calls amongst the BS-

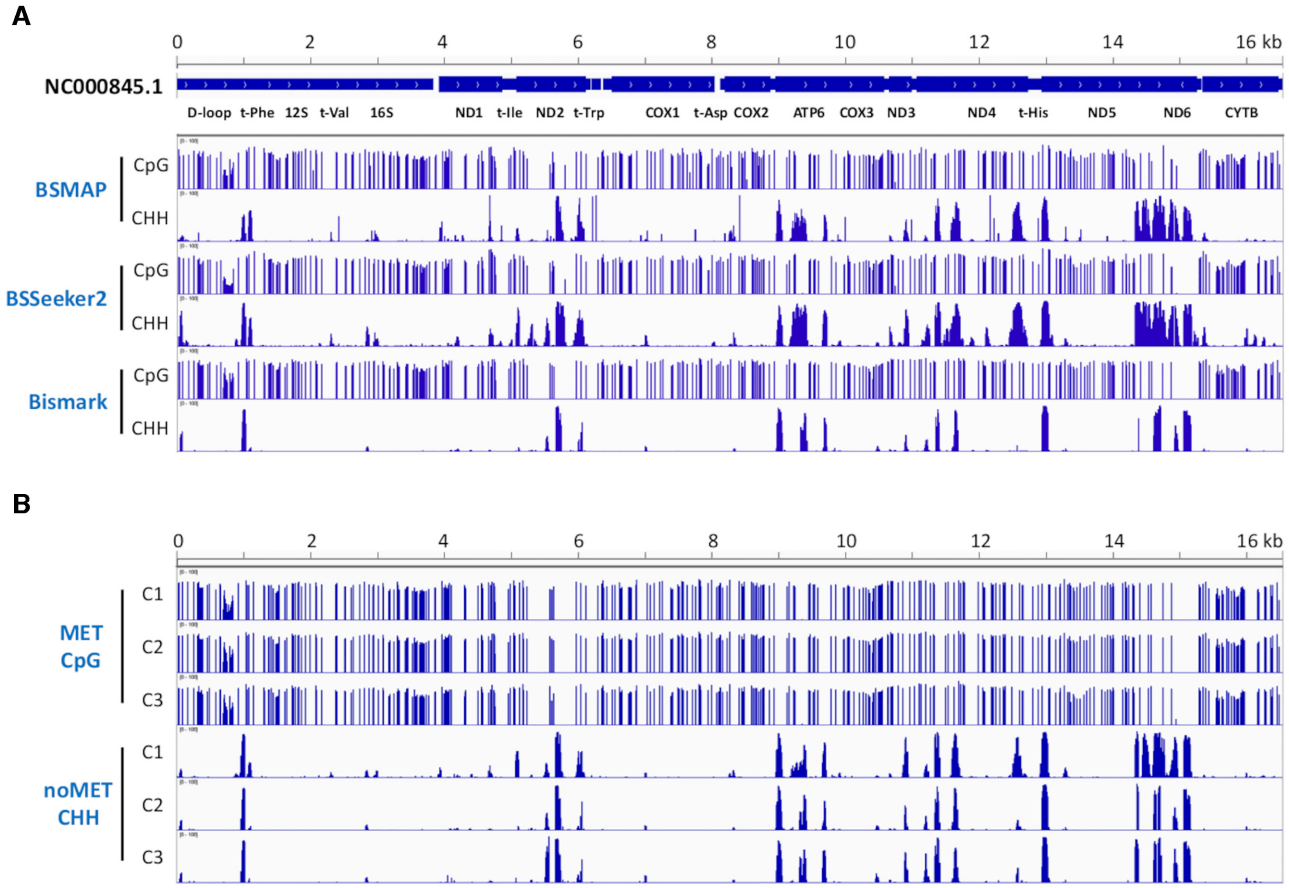


Figure 2. Detection of the levels of methylation by three BS-seq aligners in methylated mtDNA (**A**); and investigation of Bismark settings in methylated (MET) and non-methylated (noMET) mtDNA (**B**). (A) PCR amplified *Sus scrofa* mtDNA was methylated by the CpG methyltransferase *M.SssI* and used for BS-seq library construction, as described in the Methods. Bisulfite sequence data were analysed by BSMAP 1.0, BSSeeker2 v2.1.8 and Bismark v0.22.3 using default conditions. CpG and CHH methylation contexts are shown by bar graph (methylation range 0 - 100%) using IGV relative to the mtDNA sequence position on the x-axis. (B) Bismark aligner settings (–bowtie2, -N 1, -L 20, –non_directional) with different alignment stringency conditions were tested. The –score_min option in Bismark was used for investigating three conditions: C1, Bowtie2 default L,-0.6,-0.6 (least stringent); C2, Bismark default L,0,-0.2 (medium); C3, most stringent L,0,0. CpG methylation in methylated *Sus scrofa* mtDNA and CHH methylation in non-methylated mtDNA are presented.

seq aligners. To validate our hypothesis, we examined Bismark aligned reads containing methylated cytosine calls in the CHH context. Figure 3B shows an example of aligned paired end reads for this region. After trimming both ends, mate R1 only contained A/T/C, whilst mate R2 only possessed A/T/G, and the two mates partially overlapped as complementary sequences. This paired read was aligned to C to T converted sequences for the top strand and complementary sequences of C to T converted top strand, shown as G to A converted sequences, and methylation calls were made by Bismark as shown in Figure 3C. In this case, mate R1 was recognized as the bisulfite converted original top strand (OT) containing methylated Cs, indicated by 'H' on the top row. On the other hand, mate R2 was recognized as complementary to OT (CTOT), therefore, Gs in this read were called as methylated Cs. However, mate R2 exhibited possible evidence of C to T conversion from the original sequence at position 13 040 (indicated by the box in Figure 3C). This suggests that mate R2 was the original BS-converted bottom strand (OB) and mate R1 was complementary to OB (CTOB). If that is the case, all the methylated Cs in the CHH context would not have, as such, been called and should have been called as 'not a C' or 'irrelevant position'.

In the *Bismark* read mapping process, BS-seq reads are transformed into a C-to-T and G-to-A version, then each transformed read is aligned to an equivalently pre-converted reference genome using the short-read aligner Bowtie (32). Four parallel alignments determine the best unique alignment, assign origin of the strand, and make the call on methylation. This process of determination is disturbed when BS-seq reads have low complexity sequences as demonstrated in Supplementary Figure S3. As mentioned above, the sequence reads mate R1 and R2 in Figure 3C are most likely to be CTOB and OB strand reads, respectively. However, both R1 and R2 reads have a sequence mismatch (indicated by asterisk for R1, T to C change, and second from the 5' end called as 'h' for R2, G to A change, in Figure 3C), which resulted in alignment 1 as the best amongst the four parallel alignment results and, therefore, R1 and R2 were assigned as OT and CTOT, respectively (Supplementary Figure S3A). If R1 and R2's single mismatches

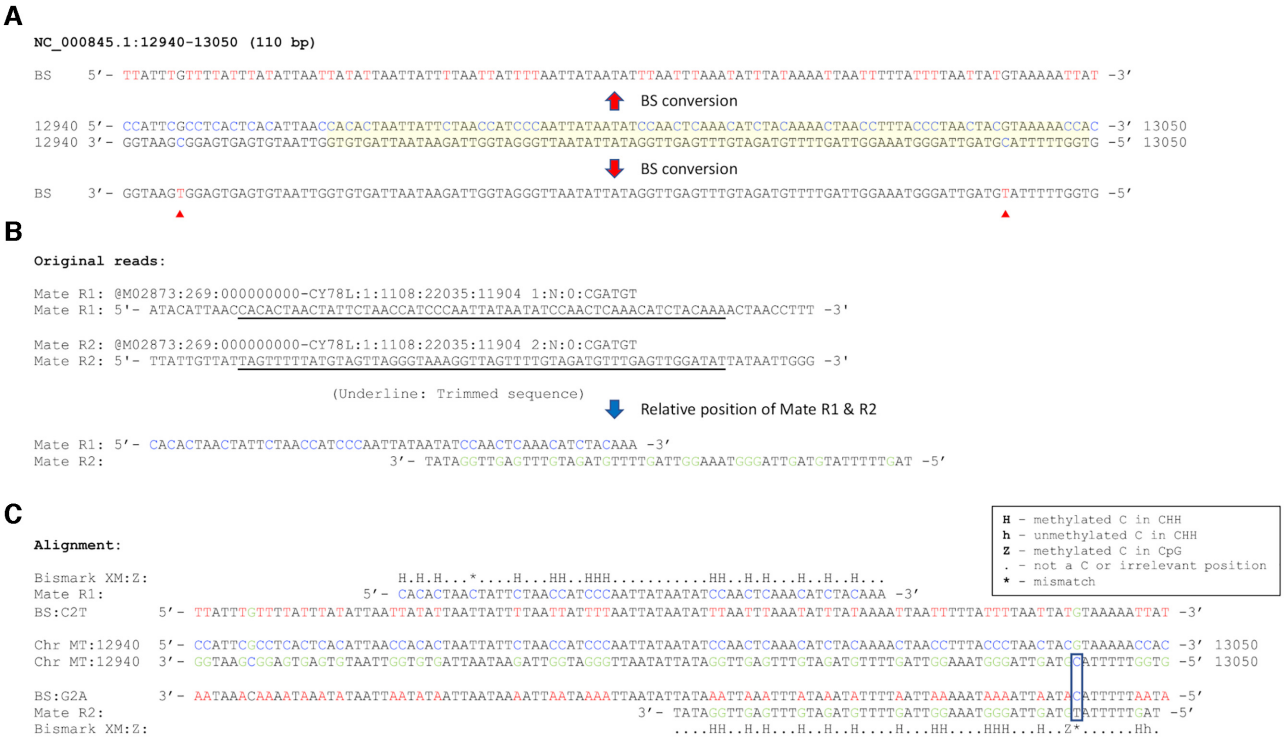


Figure 3. Example of mapped reads contributing to false positive CHH methylation calls in noMET BS-seq data. (**A**) Original double strand sequences of *Sus scrofa* mtDNA (accession number NC_000845.1) in the region 12 940–13 050, corresponding to region R4 in Figure 4, are indicated in the middle. Bisulfite (BS) converted sequences for each strand are presented at the top and bottom, respectively. Sequences corresponding to the mapped sequence example in (B and C) are indicated in yellow. Converted thiamines from cytosines in the bottom strand of this region are indicated by a red triangle. (**B**) Example of original paired end reads (mates R1 and R2) from noMET BS-seq data and trimmed sequences (underlined) used for BS-seq mapping. The relative position of the trimmed sequence mates R1 and R2 are shown below including the overlapping complementary sequence region. (**C**) Methylation calls, for example paired reads relative to original double strand sequences, are shown in the middle, highlighting the read pair mapped as the OT and CTOT strands. C to T converted sequences for the top strand are shown above the original mtDNA sequence and complementary sequences of C to T converted top strand are shown below the bottom strand as G to A converted sequences. The C to T conversion site in mate R2 is indicated by the box. Mapped mate R1 and R2 sequences and Bismark methylation calls for the corresponding sequences are also presented in appropriate positions with strings of methylation calls shown in the box.

were not present, then alignment 4 would have shown a perfect match, while alignment 1 would still have a single mismatch. Therefore, R1 and R2 would have been assigned correctly as CTOB and OB. However, if G at position 13 040 were not present in the original mtDNA sequence (Supplementary Figure S3A), both alignment 1 and 4 would have shown perfect match and Bismark may have not been able to call such sequence reads correctly. On the other hand, when the sequence read has sufficient complexity as introduced by sequence substitutions, as demonstrated in Supplementary Figure S3B, then correct strand assignment could be made for the same sequence region. This clearly demonstrates how low sequence complexity affects the processes of correct original strand assignment and sequence variations such as SNPs and sequence errors could influence final results. We analysed more fpCHH regions and numbers of aligned reads with methylated C calls in the CHH context and found similar patterns of fpCHH calling (Supplementary Figure S4). Therefore, fpCHH methylation calls are caused by a failure of correct strand assignment by BS-seq aligners in highly A/T/C-rich sequences in *Sus scrofa* mtDNA.

fpCHH regions correlate with low guanine content regions in *Sus scrofa* mtDNA

It is well known that mammal mtDNA has differences in nucleotide composition between the two strands, known as the heavy (H)- and light (L)-strands, with the H-strand containing more guanine nucleotides than the L-strand (41). From mammalian and other vertebrate mtDNA sequences deposited on NCBI database, it is evident that the top/plus strands have a significantly lower ratio of guanine nucleotides (12–16%) compared to other nucleotides (Table 1), whilst the bottom/minus strands contain more guanines. To avoid confusion over the use of H- and L-strands (42,43), we use the terms 'top/plus' and 'bottom/minus' strands when citing mtDNA sequences in the public database.

Given that we identified low levels of guanine nucleotides in the top/plus strand in fpCHH designated regions, we compared the guanine content and location of the fpCHH regions in *S. scrofa* mtDNA. We confirmed significant overlap between the fpCHH regions and low G content (<5% on average in the surrounding 30 bp) regions (Figure 4). If low G content in the top/plus strand is associated with fpCHH, as we predicted, we should be able to identify higher lev-

Table 1. Nucleotide composition in vertabrate mtDNA on the top/plus strand

Species	Common name	Accession ID	Size (bp)	A (bp)	%	T (bp)	%	G (bp)	%	C (bp)	%
Bos taurus	Cow	NC_006853.1	16 338	5457	33.4	4441	27.2	2202	13.5	4238	25.9
Sus scrofa	Pig	NC_000845.1	16 613	5766	34.7	4292	25.8	2202	13.3	4353	26.2
Homo sapiens	Human	NC_012920.1	16 568	5124	30.9	4094	24.7	2169	13.1	5181	31.3
Mus musculus	Mouse	NC_005089.1	16 299	5629	34.5	4681	28.7	2013	12.4	3976	24.4
Oryctolagus cuniculus	Rabbit	NC_001913.1	17 245	5429	31.5	4882	28.3	2350	13.6	4584	26.6
Ovis aries	Sheep	NC_001941.1	16 616	5594	33.7	4552	27.4	2181	13.1	4289	25.8
Danio rerio	Zebrafish	NC_002333.2	16 596	5301	31.9	4668	28.1	2658	16.0	3969	23.9
Gallus gallus	Chicken	NC_040902.1	16 784	5076	30.2	3982	23.7	2268	13.5	5458	32.5
Anolis carolinensis	Green Anole	EU747728.2	17 220	5580	32.4	5027	29.2	2403	14.0	4210	24.4
Xenopus tropicalis	Tropical Clawed frog	NC_006839.1	17 610	5461	31.0	4665	26.5	2544	14.4	4940	28.1

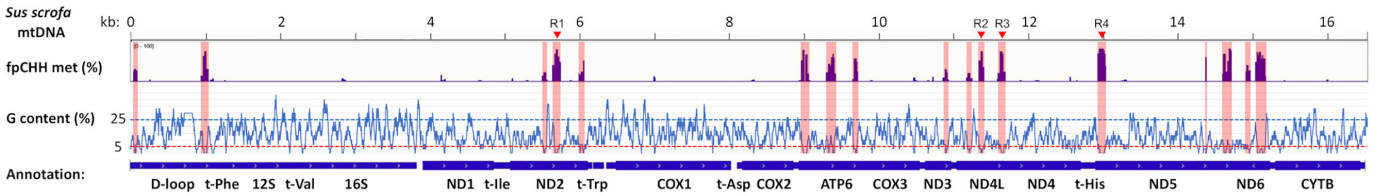


Figure 4. Guanine nucleotide content in *Sus scrofa* mtDNA and detection of false positive CHH (fpCHH) methylation by bisulfite sequencing of non-methylated (noMET) mtDNA. The average guanine nucleotide content per 30 bp region is shown on the y-axis relative to the *Sus scrofa* mtDNA sequence (Accession no. NC_000845.1) on the x-axis. Annotations of mtDNA encoded genes are presented at the bottom. Identified fpCHH methylation by Bismark analysis is shown above as a bar graph with methylation ranging from 0 to 100%. Regions of fpCHH methylation are highlighted in pink to visualize the corresponding position in the guanine (G) content graph below. Four regions (R1 to R4) investigated by bisulfite amplicon sequencing are indicated by red arrowheads.

els of CHH methylation in the top/plus rather than the bottom/minus strand. Thus, we looked at strand specific methylation levels in MET and noMET BS-seq data. For the CpG context in the MET data, we confirmed the presence of methylated CpG sites in both the top/plus and bottom/minus strands. On the other hand, for the CHH context, the majority of methylated CHH sites were found in the top/plus strand of the MET and noMET data (Figure 5A and B). Correlation analysis between levels of CHH methylation and G content revealed a significant and negative correlation for this strand ($r = -0.456$, $P = 2e{-}16$ in Figure 5C) whilst there was no significant correlation found in the bottom/minus strand. This further supports the hypothesis that biased nucleotide distribution and low G content in mtDNA is associated with fpCHH methylation calls in BS-seq data.

Short sequence reads and short BS-seq library inserts increase fpCHH

We identified many paired reads containing methylated Cs in the CHH context, which overlapped between the two mates of complementary sequences (Figure 3B). This indicates that a short BS-seq library insert has been employed and, it is, therefore, not long enough to cover a unique sequence region for the BS-seq aligner to assign the original strand correctly. Indeed, aligned paired reads with methylated Cs in the fpCHH regions have an average insert size of 51.2 bp, whilst the overall average insert size for the mapped noMET library reads was 109.5 bp (Supplementary Figure S5A and S5B). Therefore, it is a reasonable assumption that the length of sequence reads and insert size would affect the frequency of fpCHH detection. To test this hypothesis, we investigated various read and insert sizes relative to the level of fpCHH in noMET data. Bismark recommended 10 bp trimming of both ends of the read for data derived from the Zymo Pico Methyl Kit library (https://rawgit.com/FelixKrueger/Bismark/master/Docs/Bismark_User_Guide.html) and we investigated various sizes of base pair trimming. As predicted, larger base pair trimming resulted in more fpCHH peaks, especially when 20 bp were trimmed from each end (Supplementary Figure S5C).

Bismark has a read alignment option to set minimum insert size. In this context, increasing insert size significantly reduced detectable fpCHH (Supplementary Figure S5D). However, this comes at a cost. The number of mapped reads was reduced by 80% from 1147K in noMET_T10 (no minimum insert) to 244K in noMET_I150 (minimum insert size of 150 bp, Supplementary Table S2). Therefore, it is important to construct a BS-seq library with an appropriate insert size to obtain longer sequence reads that will minimise fpCHH. In some cases, however, longer inserts do not reduce fpCHH. We analysed the MET data with various insert sizes and found that fpCHH peaks were still present even when the minimum insert size was 150 bp (Supplementary Figure S6). This was caused by close proximity of the fpCHH regions to each other and the presence of each paired read being located in two different fpCHH regions. Consequently, fpCHH in vertebrate mtDNA is difficult to completely eliminate.

No false positive CHH methylation confirmed by BS amplicon sequencing

In order to validate the fpCHH detected in MET and noMET BS-seq data, we investigated fpCHH regions by BS amplicon sequencing. The four regions (R1 to R4) identi-

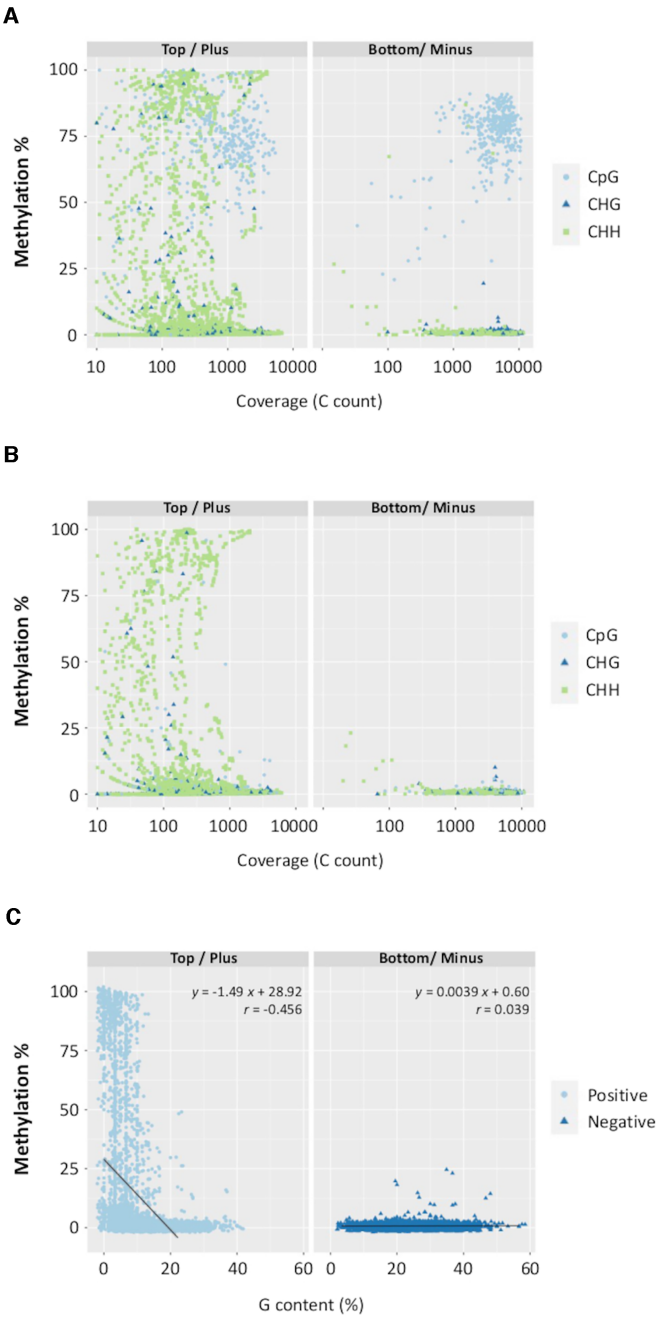


Figure 5. Strand specific mtDNA methylation relative to sequence read coverage and guanine content. Methylation levels in the top/plus (left panel) and bottom/minus (right panel) strands of methylated (MET) *Sus scrofa* mtDNA (**A**) and non-methylated (noMET) mtDNA (**B**) are presented as scatter plots. Coverage of cytosine (C) counts is shown on the x-axis using a base-10 log scale. The three sequence contexts (CpG, CHG and CHH) of DNA methylation are indicated in different shapes and colours in the graph. (**C**) Scatter plot showing levels of C methylation in the positive/top (left panel) and negative/bottom (right panel) strands relative to the guanine (G) content of the 30 bp sequence surrounding the target cytosine. Linear regression model and correlation coefficient (*r*) outcomes are shown at the top of each panel.

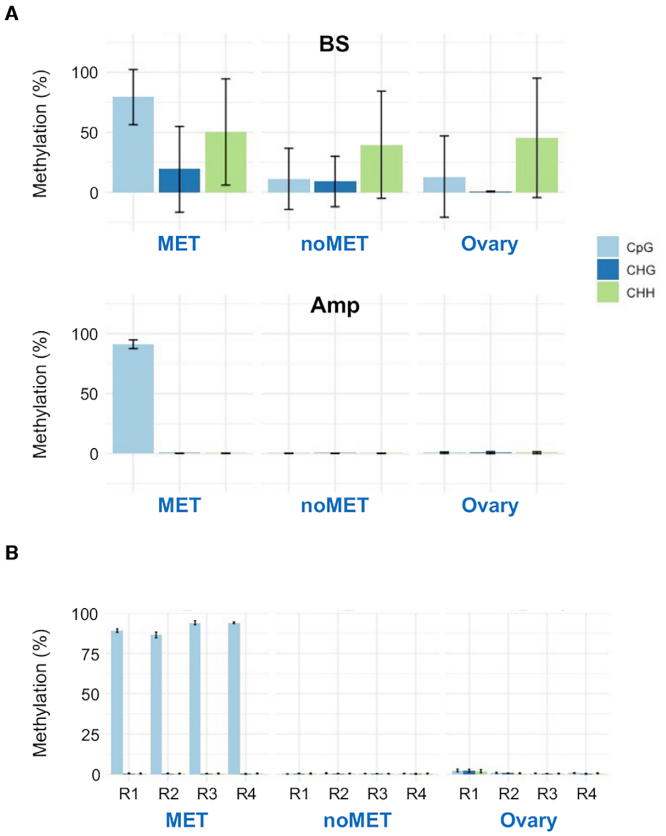


Figure 6. Validation of the methylation status of four fpCHH regions (R1 to R4 shown in Figure 4) by bisulfite amplicon sequencing. (**A**) Methylation levels for the R1 to R4 regions determined by bisulfite sequencing (BS, top) and bisulfite amplicon sequencing (Amp, bottom) in MET, noMET and Ovary samples. Averages for the R1 to R4 regions are shown for the CpG, CHG and CHH contexts in different colours. (**B**) Methylation levels for each fpCHH region. Error bars are S.D.

fied in Figure 4, including the nt 12 940–13 050 fpCHH region, were selected for BS-PCR followed by amplicon sequencing by NGS. In these four regions, 40 - 50% of cytosines in the CHH context were called as methylated in the MET, noMET and ovary mtDNA BS-seq data sets (Figure 6A). In contrast, BS amplicon sequencing did not show any significant CHG and CHH methylation (<1%) but confirmed CpG methylation in the MET data. Looking at the status of methylation by BS amplicon sequencing for the R1 to R4 regions individually (Figure 6B) and at single bp resolution (Supplementary Figure S7), there were no signs of CHH methylation in these regions. Therefore, we confirmed that methylated cytosine calls in the CHH context by BS-seq data analysis are false positive, caused by biased distribution of guanine nucleotides in *Sus scrofa* mtDNA.

No significant DNA methylation in *Sus scrofa* oocyte mtDNA

Based on potential fpCHH methylation regions in BS-seq analysis from MET and noMET data, we investigated actual mtDNA BS-seq data from *S. scrofa* oocytes. Oocytes have abundant copies of mtDNA (>150 000 copy per oocyte), and we constructed six BS-seq libraries from two independent experiments. Overall methylation levels

were <3% in all three DNA contexts in all oocyte BS-seq data, and no significant CpG methylation was identified (Supplementary Figure S8A). For methylation in the CHH context, all six data sets showed a peak in the nt 12 940–13 050 region (Supplementary Figure S8B), which is the most common fpCHH region in *Sus scrofa* mtDNA. This peak in oocyte mtDNA disappeared when the data were analysed using a minimum insert size of 150 bp, indicating this particular fpCHH methylation was caused by short insert reads mapped onto low G regions leading to incorrect methylation calls. Two oocyte BS-seq data sets (Oc #2–2 and #2–3) showed several additional broader CHH methylation peaks and low levels of CpG peaks throughout the mitochondrial genome (Supplementary Figure S8A and S8B), which was indicative of incomplete BS conversion.

Oocyte data showed a single major fpCHH peak in the nt 12 940–13 050 region, while noMET data revealed more than 10 fpCHH peaks (Supplementary Figure S2B). As demonstrated above, the longer BS-seq library insert size was associated with fewer number of detectable fpCHH peaks (Supplementary Figure S5D). Consistent with this, we found that the average oocyte BS-seq library insert size (164–202 bp) is bigger than the noMET data (110 bp in Supplementary Figure S5B). This may be one of the reasons for the single major fpCHH methylation peak detected in the oocyte data. We assumed that the results would be different if we analysed the same data in single-end mode instead of paired-end. Indeed, analysis of data by single-end mode increased the number of fpCHH peaks, which corresponded to fpCHH regions identified in the noMET data (Supplementary Figure S8C). This further highlights the importance of BS-seq data analysis in the paired-end mode and using a longer insert size to minimise the detection of fpCHH methylation.

fpCHH in nuclear-mtDNA pseudogenes (NUMT)

There are several reports that investigated whether BS-seq reads derived from NUMT affected mtDNA methylation results (13,19). Since coverage of NUMT derived reads are low and mtDNA derived reads are more abundant, it did not affect the status of mtDNA methylation. However, the effect of mtDNA derived BS-seq reads on the methylation status of NUMT regions has not been addressed so far. We generated *S. scrofa* oocyte whole WGBS data and looked at NUMT regions, especially regions with greater identity (>90%) to mtDNA. The largest NUMT in *S. scrofa* is found in chromosome 2 (115 563 900–115 574 158: 10.2 kb) with 90.5% identity to the equivalent mtDNA region (Supplementary Table S3). The overall average depth of Cs with at least one read in the genome was 1.98 in oocyte WGBS data, whilst we found much higher read depth (>100× in many regions) in the chromosome 2 NUMT region (Figure 7A). We also found several CHH methylation peaks in this region. We assumed higher read coverage in this region and CHH methylation could be due to reads derived from mtDNA, since oocytes have high numbers of mtDNA copy per cell. We looked at 128 reads mapped to one of the methylated regions on chromosome 2: 115,566,520 - 620 and found that there are two types of reads mapped in this region, based on SNPs. The first group of reads (only 2) seemed to be derived from genomic DNA as they had identical sequences to the NUMT region except for the G to A conversions (Figure 7B). The second group of reads (126) contained 4–6 SNPs and all these SNPs are identical to the mtDNA sequence (NC_000845.1) suggesting that those reads are most likely derived from mtDNA. Furthermore, this region is a low G content region (only 2 Gs in 103 bp), which is a typical feature of a fpCHH region and highly similar to the nt 8961–9,060 fpCHH region in the *Sus scrofa* mitochondrial genome (Figure 4). Therefore, these CHH methylation peaks are most likely false positive methylation calls arising from mtDNA derived BS-seq reads. We confirmed a similar pattern for the other CHH peaks in the other NUMTs in different chromosomes (e.g. Chr 8:118 587 710–790, Supplementary Figure S9). These results suggest that abundant mtDNA derived BS-seq reads could contribute to the fpCHH methylation calls in the NUMT genomic regions.

fpCHH methylation calls associate with non-directional library or alignment mode

We have demonstrated the presence of fpCHH methylation in *Sus scrofa* mtDNA BS-seq analysis data. Our BS-seq libraries were constructed using the Zymo Pico Methyl-Seq Library Prep Kit, which deliver non-directional BS-seq libraries. We addressed whether detection of fpCHH is dependent on (i) sample type; (ii) type of BS-seq library or (iii) method of analysis, by investigating publicly available human BS-seq data (Supplementary Table S4) for fpCHH methylation in mtDNA. For directional BS-seq libraries, e.g. standard library (adaptor ligation followed by BS treatment), TruSeq DNA Methylation Kit (also known as EpiGnome), and Swift Accel-NGS Methyl-Seq, only original top and bottom strands (OT and OB) are sequenced. We analysed different directional BS-seq libraries derived from various human cell or tissue types using the directional mapping mode in Bismark and found no sign of fpCHH methylation (Figure 8A). However, when the same data sets were analysed in the non-directional mapping mode, we found fpCHH peaks in low G regions in some cases in human mtDNA. These were located in the ATP6, ND4L and ND6 coding regions and also found in the *S. scrofa* mtDNA BS-seq noMET data (Figure 4). Raine *et al.* (44) investigated four different BS-seq library preparation methods using the same DNA material (SRA project ID: GSE89213). We used these data sets to test the effect of library preparation on the detection fpCHH methylation. Different libraries showed different levels of fpCHH when analysed using the non-directional mode (Figure 8B). The TruSeq BS-seq library (SRR4453294) showed high levels of methylation throughout the mitochondrial genome in both CpG and CHH contexts suggesting incomplete BS conversion (44). Nevertheless, it is critical to choose the correct BS-seq read alignment mode for directional BS-seq libraries to avoid detection of fpCHH methylation. Next, we analysed BS-seq data from non-directional libraries (Zymo Pico Methyl Kit and MDA-TruSeq; a combination of BS-converted DNA amplification and Illumina TruSeq DNA sample prep kit V2) using the non-directional mapping mode. We confirmed CHH methylation peaks in typical fpCHH methylated re-

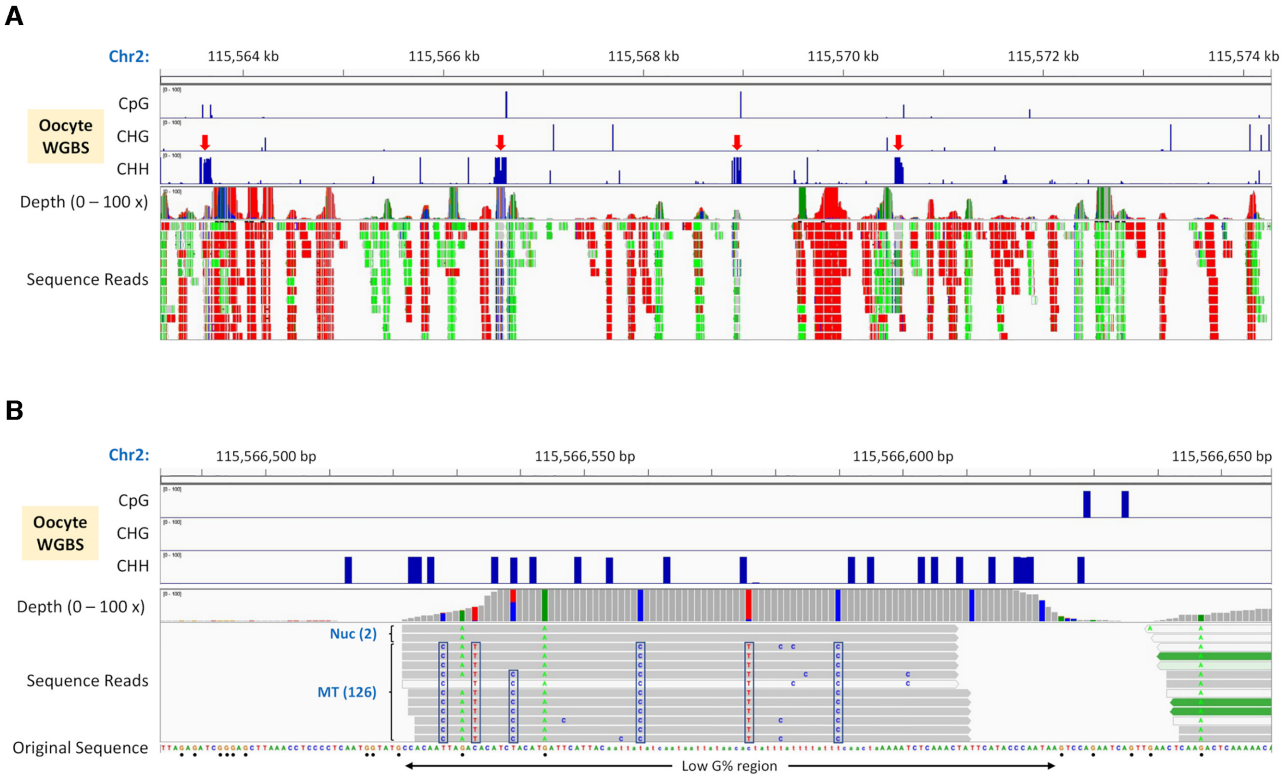


Figure 7. False positive CHH methylation calls in NUMT. (**A**) WGBS of *Sus scrofa* oocytes showing CHH methylation peaks (red arrows) in the chromosome 2 NUMT region. The CpG, CHG and CHH methylation contexts are shown by bar graph relative to the *Sus scrofa* v11.1 chromosome 2 reference sequence on the x-axis. Depth of mapped sequence reads and actual mapped reads are indicated in the bottom half of the panel. (**B**) Magnified image of the chromosome 2 NUMT region shown in (A). The original reference sequence is shown at the bottom and sequences with differences from the reference indicated on sequence reads by letters. Based on SNPs in the mapped reads, sequence reads derived from nuclear DNA (Nuc) and mtDNA (MT) can be identified and their coverage read number is also indicated. Sequence differences associated with mtDNA sequence are shown by boxes. Guanine nucleotides in this NUMT region are indicated by dots at the bottom of the sequence and the low guanine content region is also shown below.

gions suggesting that fpCHH was detectable in these data sets (Figure 8C). We also found low level peaks throughout the mitochondrial genome in some data sets. These outcomes suggest that non-directional BS-seq library and/or non-directional analysis mode is associated with the detection of fpCHH methylation.

Abundance of bottom strand mtDNA BS-seq reads and a lack of top strand reads associate with fpCHH methylation calls

It has been reported in mtDNA BS-seq data analysis that there is an abundance of bottom strand reads and an inverse relationship between the levels of mtDNA methylation and depth of C counts (13,14,19,45). We observed that detected fpCHH methylation has relatively low C counts, whilst non-methylated Cs have higher coverage (Figure 5), which confirms the inverse relationship previously reported. We also observed that the number of reads mapped to each strand was highly biased to the bottom strand in our data sets and other human BS-seq data (Supplementary Table S5). Theoretically, BS-seq reads would be mapped equally to top and bottom strands. However, we observed a significantly lower number of reads (normally 4- to 8-fold lower than bottom strands reads) mapped to the top strand in human BS-seq data and our *Sus scrofa* data. In the worst case, less than 1% of the reads were mapped to the top strand (SRR1035790 in Supplementary Table S5). Our analysed data sets showed a tendency to negative correlation between top to bottom strand (T/B) read ratios and the levels of fpCHH detected (Figure 8 and Supplementary Table S5) whilst oocyte WGBS data mapped to *S. scrofa* chromosomes or scaffolds were much closer to an equal ratio. This may be related to biased nucleotide distribution in vertebrate mitochondrial genomes, which resulted in lower representation of reads corresponding to top strand mtDNA in BS-seq libraries.

We further looked at the methylation call for C counts in a strand specific manner to evaluate the impact of skewed T/B read ratio. In the noMET data, the high depth of C counts throughout the mitochondrial genome bottom strand was confirmed (Figure 9A). In contrast, there were several regions in the top strand showing low depth of C counts, especially regions surrounding fpCHH methylation. Figure 9B showed a magnified image of fpCHH regions between nt 8000 and 12 000. fpCHH regions clearly showed low depth of top strand C counts, further confirming the inverse relationship between the levels of mtDNA methylation and depth of C counts. Low depth of top C count regions seemed to correlate with the pattern of G content for the top strand (Figure 9A). Indeed, the C count for the top strand correlated significantly with the G nucleotide content

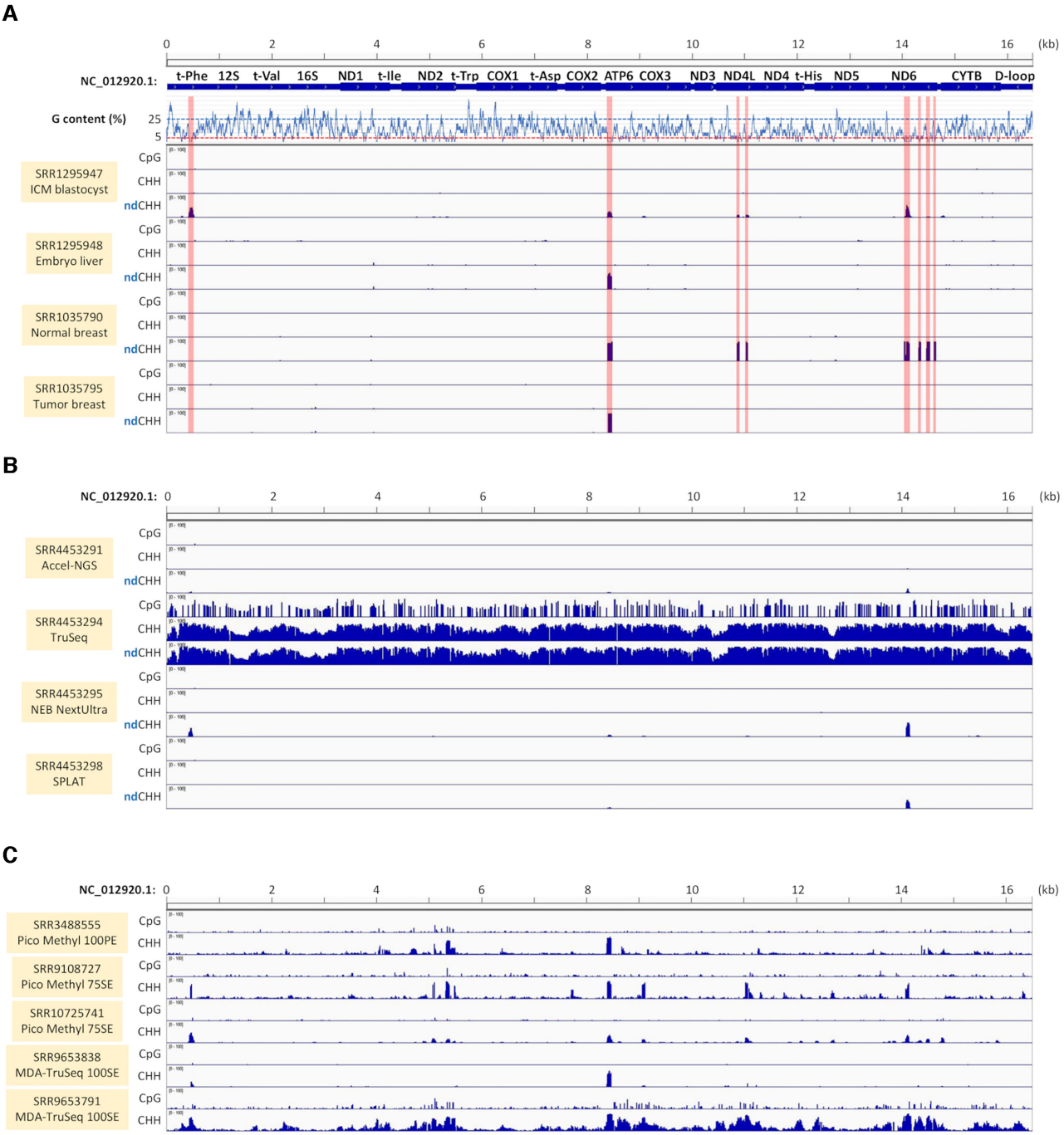


Figure 8. Detection of false positive CHH (fpCHH) methylation in human BS-seq data. (**A**) Association of fpCHH methylation with the non-directional BS-seq analysis option. Human BS-seq data were analysed by Bismark with the directional (CpG and CHH) or the non-directional (ndCHH) option. CpG and CHH methylation contexts were shown by bar graph relative to the mtDNA sequence position on the x-axis. Average guanine nucleotide contents of the 30 bp region are shown on the y-axis relative to the *Homo sapiens* mtDNA sequence (Accession no. NC_0012920.1) on the x-axis. Annotations of mtDNA encoded genes are presented at the top. (**B**) Association of fpCHH with BS-seq library preparation methods and quality. BS-seq data (Raine, 2016: GSE89213) obtained from four different library preparation methods, Accel-NGS Methyl-Seq DNA library kit (Swift BioSciences), TruSeq DNA Methylation Kit (Illumina), NEBNextUltra kit (New England Biolab) and Splinted adaptor tagging (SPLAT) protocol, were analysed. (**C**) Analysis of paired-end and single-end reads derived from a non-directional BS-seq library. Sequence Read Archive run data IDs used for analyses are shown in each panel.

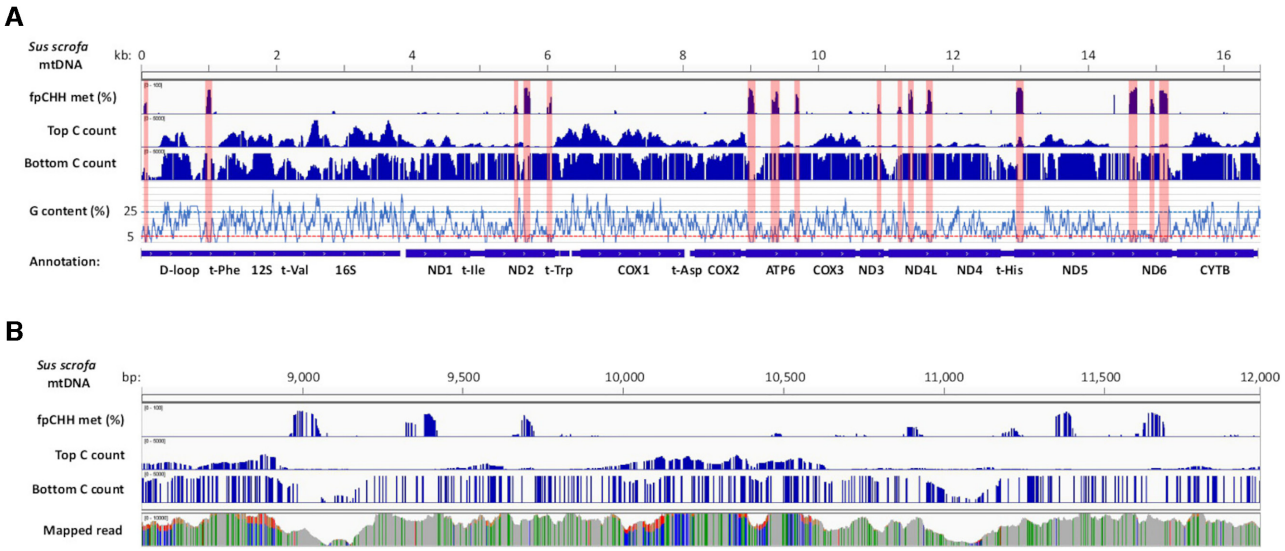


Figure 9. Biased distribution of mapped reads originated from the top and bottom strands. (**A**) Association of fpCHH regions with cytosine (C) methylation call depth. Levels of fpCHH methylation in noMET data are presented in the top row. C methylation call counts for top and bottom strands (range from 0 to 5000) are shown in the middle. Average guanine nucleotide contents are shown on the y-axis relative to the *Sus scrofa* mtDNA sequence on the x-axis and annotations of mtDNA encoded genes are presented at the bottom. The positions of the fpCHH methylated regions are highlighted in pink to visualize the relative position in the panel below. (**B**) Magnified image of the 8500–12 000 region in the panel (A). Depth of uniquely mapped reads for each nucleotide (range from 0 to 10 000) is shown in the bottom row.

($r = 0.41$, $P = 2.2e{-}16$), and it also weekly and negatively correlated with C nucleotide content ($r = -0.16$, $p = 2.2e{-}16$). These results suggest that biased nucleotide distribution in the mitochondrial genome is associated with regional low abundance of the top strand read coverage. This means a shortage of BS converted top strand reads in the fpCHH methylated regions, which is necessary for correct methylation calls. Furthermore, we have shown that a subset of bottom strand derived reads contributes to fpCHH methylation calls (Figure 4). Therefore, the abundance of bottom strand BS-seq reads and a lack of top strand reads are one of the major causes for the false positive methylation calls in mtDNA.

DISCUSSION

DNA methylation of the mammalian mitochondrial genome has been of great interest and would have significant implications for mtDNA gene expression and mtDNA replication (4–7). There are conflicting data about mtDNA methylation utilizing different methods to investigate its status (11,13). Bisulfite sequencing is an important tool for epigenetic analysis and WGBS provides single bp resolution of DNA methylation (21,22). In previous work using methylated DNA immunoprecipitation (MeDIP), we have demonstrated the use of methylated (MET) and non-methylated (noMET) mtDNA controls to determine the presence of mtDNA methylation in tumour cells (6). We have extended this approach in our analysis of mtDNA methylation through BS-seq of the *Sus scrofa* mitochondrial genome and shown that BS-seq aligners failed to correctly assign methylation calls in specific regions of the mitochondrial genome. We found in regions with low G content on the top strand of the mitochondrial genome that short sequence reads from the corresponding bottom strand sequences remained as almost original non-BS converted sequences. This is due to an insufficient number of C to T conversion sites on the original bottom (OB) strand that led to misrepresentation of the BS-converted OB strand as the complementary original top (CTOT) strand, which, in turn, resulted in false positive CHH methylation calls (Figure 3). We have clearly demonstrated how *Bismark*, the most popular BS-seq aligner, incorrectly assigned original strands due to the lack of sequence complexity (Supplementary Figure S3). To our knowledge, this is the first report describing a misassignment of the origin of BS-seq read strand by BS-seq aligners causing significant differences in DNA methylation calls. This appeared to be unavoidable with the three BS-seq aligners utilizing two different alignment approaches tested due to the acceptance of sequence mismatches in read alignment algorithms; and the low complexity of BS-converted sequences, which made it difficult to correctly assign the origin of each DNA strand. We have demonstrated that fpCHH methylation calls are enhanced by short BS-seq reads, short library inserts, single end mapping mode, and the non-directional read mapping mode by the aligners. Our results provide clear evidence that BS-seq aligners can overestimate the non-CpG methylation context in specific types of sequences under certain conditions.

One of the major causes for fpCHH methylation calls is the highly biased top/bottom BS-seq read coverage previously reported in mtDNA BS-seq analyses (13–15,19,45). This is associated with biased nucleotide content and distribution in the top/plus strand of mtDNA (Figure 9). The top strand of the *S. scrofa* mitochondrial genome contains 4353 Cs, while the bottom strand has 2202 Cs (Table 1). Bisulfite treatment converts unmethylated cytosines to uracils and

this deamination step associates with DNA fragmentation (23,46). Sequence regions containing more unmethylated Cs would likely be more fragmented. Indeed, it has been recently demonstrated that unmethylated C-rich sequences are more affected by BS-induced degradation than C-poor sequences. Therefore, the C-rich mtDNA top strand reads were significantly underrepresented over the bottom strand reads due to degradation (25). For example, there are 35 Cs in the 110 bp top strand sequence of *Sus scrofa* mtDNA (nt 12 940–13 050), while only two Cs are present in the bottom strand (Figure 3). Consequently, the top strand sequence for this region would be more fragmented, significantly reducing the recovery rate of DNA fragments from this region after BS-treatment for the BS-seq library construction process. Furthermore, this region of the top strand would become an extremely AT-rich sequence after BS conversion.

Generally, conventional BS-seq library construction protocols include a PCR amplification step. To this extent, it has been shown that the representation of AT-rich sequences in an amplified library is significantly reduced, depending on the type of denaturation undertaken prior to BS-treatment and the DNA polymerase used (25). Amplification of extremely AT-rich sequences requires a lower temperature (60–65°C) for the PCR extension step (47,48), whilst most BS-seq library construction protocols employ a normal extension temperature in the range of 68–72°C (e.g. https://www.genetargetsolutions.com.au/wp-content/uploads/2015/05/Accel-NGS-Methyl-Seq-DNA-Library-Kit-Manual.pdf, https://sapac.illumina.com/products/by-type/sequencing-kits/library-prep-kits/truseq-methyl-capture-epic.html and https://files.zymoresearch.com/protocols/_d5455_d5456_picomethylseq.pdf). Indeed, we had to employ an extension temperature of 65°C to amplify regions R1 to R4 for BS-amplicon sequencing (Figures 4, 6 and Materials and Methods). Our correlation analysis revealed that the top strand read number (C count in Figure 9) was less associated with C nucleotide content ($r = -0.16$), an indicator of the number of potential degradation sites by BS-conversion. However, it correlated more with low G content ($r = 0.41$), an indicator of an AT-rich sequence after BS-conversion, which suggests that PCR amplification bias contributed to a more skewed top/bottom read ratio.

Nevertheless, a combined effect of the BS-induced degradation and PCR amplification bias would have resulted in extremely skewed representation of top and bottom strand reads in the fpCHH regions (Supplementary Table S3 and Figure 9). A similar outcome has been demonstrated in a C-rich strand of a telomere repeat, which was hardly detectable in the raw reads of most BS-seq datasets (25). Telomeres have a tandem repeat sequence of (CCCTAA)n in mammals and (CCCTAAA)n in plants, which results in C-rich and low G sequences on one strand and resembles fpCHH sequences in mtDNA. Interestingly, BS-seq overestimated methylation levels, when the telomeric C-rich strand reads were in low abundance, as in the *Arabidopsis thaliana* BS-seq data (49). Underrepresentation of genomic regions enriched for unmethylated cytosines affects the final estimation of 5mC levels (25). In this respect, very low abundance of OT reads corresponding to fpCHH regions were outnumbered by a subset of highly abundant OB reads that were misannotated as CTOT by BS-seq aligners, which resulted in overestimation of methylation levels in fpCHH regions. This links to a tendency we observed for the negative correlation between T/B read ratios and the levels of fpCHH detected (Figure 8 and Supplementary Table S5) and is largely dependent on the methods and conditions employed by BS-seq library construction protocols. For example, the KAPA Uracil + DNA polymerase used for library amplification showed the least biased nucleotide representation (25). We further found that a modest T/B read ratio was associated with the lowest level of fpCHH peak detected (SRR4453291 in Supplementary Table S5 and Figure 8). We also showed (Figure 5) that this biased top/bottom representation of reads explains the inverse relationship between the levels of mtDNA methylation and coverage, and L-strand or top/plus strand biased detection of non-CpG methylation (13,14,18,19,45). In contrast, BS-seq aligners properly called methylation in the CpG context including fpCHH regions (Figure 2), thus this is a specific issue for non-CpG methylation, especially in the asymmetric CHH context.

Potential fpCHH methylation calls have previously been reported as a high ratio of unconverted Cs in some reads in the ND2, ND5 and ND6 regions associated with negative controls of mtDNA BS-seq data sets (18). These are the same typical fpCHH regions that we identified (Figure 2), with unconverted Cs likely arising from short sequence reads (50 bp PE) associated fpCHH methylation calls. In this respect, the authors hypothesised that those reads had escaped bisulfite conversion due to the DNA structures induced after denaturation and the local sequences of the top strand, which were then removed from their analysis using a filter. Others have also noted that BS-unconverted Cs in mtDNA were due to the secondary structure (19). We demonstrated that the linearisation of *Sus scrofa* ovarian mtDNA did not affect the methylation status under our experimental conditions and our MET and noMET data were obtained from linear PCR fragments treated by BS. Therefore, this is not due to incomplete BS conversion associated with mtDNA secondary structure, but a fundamental issue in BS-seq aligner algorithms for specific mtDNA sequence regions, as mentioned above. We also identified potential fpCHH methylation peaks at typical mtDNA locations in mammalian somatic cells and oocytes that were observed in previous publications. These analysed non-directional BS-seq libraries and/or were analysed using the non-directional read mapping mode (13,15).

The fpCHH regions we identified from *Sus scrofa* mtDNA were located in the D-loop, and the *ATP6, ATP8*, *ND2*, *ND3*, *ND4*, *ND4L*, *ND5* and *ND6* gene regions (Figure 4 and Supplementary Table S6). Some of these regions have been analysed by alternative methods such as BS-pyrosequencing and BS-amplicon sequencing (8,14,16,19,20,45). Both methods require PCR amplification from BS-treated DNA template. For instance, it has been demonstrated that primer specificity in pyrosequencing significantly affects detectable methylation levels and the presence of unconverted DNA could overestimate levels of methylation (20). As shown by biased representation of the top/bottom strand reads in BS-seq data, potential

Table 2. Proposed BS-seq analysis conditions for CHH methylation calls.

Subject	Step	Recommendation
BS-seq library	Library preparation	Directional (preferably with KAPA uracil + DNA polymerase)
Library insert size	Library preparation	>150 bp
Read type	NGS	Paired end
Read length	NGS & trimming	>100 bp
Read mapping	BS-seq alignment	Directional & paired end mode
Mapping insert size	BS-seq alignment	>150 bp
Positive CHH methylation	Methylation call	Check reference sequence (>1 G per 30 bp)

over fragmentation of the top strand and reduced amplification of AT-rich sequences could influence detectable levels of methylation by these methods in fpCHH regions. MeDIP has also been used as an alternative method to validate detected methylation by BS-seq analysis (7,13,14). Given the basal affinity of 5-mC antibody against unmethylated cytosines (50), biased nucleotide distribution in mtDNA could affect MeDIP results when overall levels of cytosine methylation in mtDNA are low. For example, we might expect over representation of C-rich sequences, as with the top strand sequence of fpCHH regions, in the mitochondrial genome of MeDIP data. Some MeDIP patterns for mtDNA resembled the fpCHH peaks (13,14,51). Therefore, as for whole mtDNA BS-seq, it is essential that MeDIP is performed using methylated and non-methylated control DNAs, no input and BS unconverted controls, and the results need to be normalized against controls accordingly, as previously shown (6).

Filtering out non-BS converted reads based on the number of unconverted Cs in non-CpG context is one option to avoid fpCHH associated overestimation. The use of self-designed filters (18) or Bismark's own option (*filter_non_conversion*) can eliminate mapped reads with high numbers of methylated Cs in the non-CpG context (25). Eliminating false positive non-CpG methylation calls in this manner results in no significant mtDNA methylation being detected (18,52). We also confirmed the effectiveness in our MET and noMET data sets to improve fpCHH detection (Supplementary Figure S2B). However, the majority of the reads removed by the filter did not contain unconverted Cs but they were misassigned to the origin of the strand in the read alignment step. More importantly, this filter cannot be used for some species and types of tissues and cells that have non-CpG methylation, such as plants which possess CHROMOMETHYLASE 2 and 3 for the maintenance of CHH and CHG methylation (53). Accumulated evidence supports the presence of non-CpG methylation with important functions in mammalian stem cells, cancer cells and neurons of the brain (54,55). Thus, simple elimination of reads with high levels of non-CpG methylation calls is a risky approach and could lead to misinterpreted BS-seq outcomes. Amplification-free BS-seq library preparation is proven to be the least biased approach for WGBS (25), but it is difficult to apply when the initial input quantity is very low, as with single cell epigenetic approaches (56,57). Therefore, detectable CHH methylation by BS-seq needs to be validated by alternative methods. We have demonstrated that there was no detectable CHH methylation in fpCHH regions by BS-amplicon sequencing (Figure 5). Unbiased methylation detection methods with the ability to analyse DNA methylation at single bp resolution, such as the PAC BIO (58,59) and Nanopore (60,61) platforms, are other possible options. Recent reports using the Nanopore method revealed low levels of global mtDNA methylation, except at specific CpG sites that exhibited elevated methylation levels, and far fewer non-CpG methylation sites were found in specific tissue types (60,61). The locations of Nanopore detectable CpG and non-CpG methylation were not similar to fpCHH methylation peaks or the high levels of methylation previously reported (13,14,51), and those previously reported mtDNA methylated regions in various tissues and cell types were not confirmed (61).

Alternatively, an additional filter in the BS-seq aligner strand assignment process could be introduced to improve fpCHH detection. Our data provided a basis for the presence of fpCHH regions in terms of length and the number of Gs (Supplementary Table S6) and could be used to develop a filter to flag genomic regions which require attention when making CHH methylation calls. For example, a 70 bp sequence containing two or fewer Gs would raise a flag, which would incorporate the surrounding genomic region (±30 bp). Sequence reads mapped in these regions could undergo additional filtering when undertaking parallel read mapping alignment. If the difference for the mismatch score between alignments is less than two (e.g. alignment 1 and 4 in Supplementary Figure S3A), then CHH methylation calls for these reads would be discarded or flagged for attention due to their potential ambiguity. More practically, BS-seq reads with, for example, >5 CHH methylation calls in the read could undergo this additional filtering process. Using this approach, it might be possible to retain reliable CHH methylation calls and effectively remove fpCHH methylation calls with the minimum loss of read data that was observed with other conditional changes we employed, for example when increasing insert size (Supplementary Table S2). With the currently available tools and options, we also provide proposed BS-seq analysis conditions for CHH methylation calls (Table 2) to minimise fpCHH detection in outputs, based on our findings. Each of these factors would help to improve outcomes, e.g. preparation of BS-seq library with longer insert size (>150 bp), paired-end reads with longer read length (>100 bp), and setting longer minimum insert size (>150 bp) for paired-end mapping (Table 2). However, directional BS-seq libraries and directional mapping mode are critical to avoid fpCHH and it would be good to check G content in the reference sequence when positive CHH methylation calls are identified.

In conclusion, we have demonstrated how false positive DNA methylation in the CHH context is detected in mtDNA BS-seq data and the conditions that influence

false detection. This is associated with biased nucleotide distribution in vertebrate mtDNA and BS-seq methodology both experimentally and analytically. By discounting fpCHH methylation from our oocyte BS-seq data and publicly available BS-seq data sets, we observed no or low levels of mammalian mtDNA methylation in the CpG and non-CpG contexts. Measuring levels of DNA methylation by BS-seq could potentially be overestimated in specific sequence regions. Therefore, detected asymmetric non-CpG methylation results using BS-seq need to be validated, desirably by unbiased alternative methods.

DATA AVAILABILITY

All sequence data have been deposited in NCBI Sequence Read Archive (SRA) under the BioProject ID PRJNA752230 and data accession IDs SRR15351362 - 82.

SUPPLEMENTARY DATA

Supplementary Data are available at NARGAB Online.

ACKNOWLEDGEMENTS

Author contribution: J.S.J. conceived the work. T.O., X.S. and J.S.J. planned and designed the research. T.O., X.S., S.McI. performed experiments and analysed data. T.O. and J.S.J. discussed results and interpretation of data. T.O. and J.S.J. wrote and edited the manuscript. J.S.J. obtained funding for the work. All authors reviewed the manuscript.

FUNDING

National Health and Medical Research Council [APP1160106].
Conflict of interest statement. None declared.

REFERENCES

1. Pollack,Y., Kasir,J., Shemer,R., Metzger,S. and Szyf,M. (1984) Methylation pattern of mouse mitochondrial DNA. *Nucleic Acids Res.*, **12**, 4811–4824.
2. Shmookler Reis,R.J. and Goldstein,S. (1983) Mitochondrial DNA in mortal and immortal human cells. Genome number, integrity, and methylation. *J. Biol. Chem.*, **258**, 9078–9085.
3. Vanyushin,B.F. and Kirnos,M.D. (1974) The nucleotide composition and pyrimidine clusters in DNA from beef heart mitochondria. *FEBS Lett.*, **39**, 195–199.
4. Pirola,C.J., Gianotti,T.F., Burgueno,A.L., Rey-Funes,M., Loidl,C.F., Mallardi,P., Martino,J.S., Castano,G.O. and Sookoian,S. (2013) Epigenetic modification of liver mitochondrial DNA is associated with histological severity of nonalcoholic fatty liver disease. *Gut*, **62**, 1356–1363.
5. Shock,L.S., Thakkar,P.V., Peterson,E.J., Moran,R.G. and Taylor,S.M. (2011) DNA methyltransferase 1, cytosine methylation, and cytosine hydroxymethylation in mammalian mitochondria. *Proc. Natl. Acad. Sci. U.S.A.*, **108**, 3630–3635.
6. Sun,X., Johnson,J. and St John,J.C. (2018) Global DNA methylation synergistically regulates the nuclear and mitochondrial genomes in glioblastoma cells. *Nucleic Acids Res.*, **46**, 5977–5995.
7. Sun,X., Vaghjiani,V., Jayasekara,W.S.N., Cain,J.E. and St John,J.C. (2018) The degree of mitochondrial DNA methylation in tumor models of glioblastoma and osteosarcoma. *Clin. Epigenetics*, **10**, 157.
8. Bellizzi,D., D'Aquila,P., Scafone,T., Giordano,M., Riso,V., Riccio,A. and Passarino,G. (2013) The control region of mitochondrial DNA shows an unusual CpG and non-CpG methylation pattern. *DNA Res.*, **20**, 537–547.
9. Chestnut,B.A., Chang,Q., Price,A., Lesuisse,C., Wong,M. and Martin,L.J. (2011) Epigenetic regulation of motor neuron cell death through DNA methylation. *J. Neurosci.*, **31**, 16619–16636.
10. Wong,M., Gertz,B., Chestnut,B.A. and Martin,L.J. (2013) Mitochondrial DNMT3A and DNA methylation in skeletal muscle and CNS of transgenic mouse models of ALS. *Front. Cell Neurosci.*, **7**, 279.
11. Maresca,A., Zaffagnini,M., Caporali,L., Carelli,V. and Zanna,C. (2015) DNA methyltransferase 1 mutations and mitochondrial pathology: is mtDNA methylated? *Front Genet*, **6**, 90.
12. Byun,H.M., Panni,T., Motta,V., Hou,L., Nordio,F., Apostoli,P., Bertazzi,P.A. and Baccarelli,A.A. (2013) Effects of airborne pollutants on mitochondrial DNA methylation. *Part. Fibre Toxicol.*, **10**, 18.
13. Dou,X., Boyd-Kirkup,J.D., McDermott,J., Zhang,X., Li,F., Rong,B., Zhang,R., Miao,B., Chen,P., Cheng,H. *et al.* (2019) The strand-biased mitochondrial DNA methylome and its regulation by DNMT3A. *Genome Res.*, **29**, 1622–1634.
14. Patil,V., Cuenin,C., Chung,F., Aguilera,J.R.R., Fernandez-Jimenez,N., Romero-Garmendia,I., Bilbao,J.R., Cahais,V., Rothwell,J. and Herceg,Z. (2019) Human mitochondrial DNA is extensively methylated in a non-CpG context. *Nucleic Acids Res.*, **47**, 10072–10085.
15. Sirard,M.A. (2019) Distribution and dynamics of mitochondrial DNA methylation in oocytes, embryos and granulosa cells. *Sci. Rep.*, **9**, 11937.
16. Hong,E.E., Okitsu,C.Y., Smith,A.D. and Hsieh,C.L. (2013) Regionally specific and genome-wide analyses conclusively demonstrate the absence of CpG methylation in human mitochondrial DNA. *Mol. Cell. Biol.*, **33**, 2683–2690.
17. Liu,B., Du,Q., Chen,L., Fu,G., Li,S., Fu,L., Zhang,X., Ma,C. and Bin,C. (2016) CpG methylation patterns of human mitochondrial DNA. *Sci. Rep.*, **6**, 23421.
18. Matsuda,S., Yasukawa,T., Sakaguchi,Y., Ichiyanagi,K., Unoki,M., Gotoh,K., Fukuda,K., Sasaki,H., Suzuki,T. and Kang,D. (2018) Accurate estimation of 5-methylcytosine in mammalian mitochondrial DNA. *Sci. Rep.*, **8**, 5801.
19. Mechta,M., Ingerslev,L.R., Fabre,O., Picard,M. and Barres,R. (2017) Evidence suggesting absence of mitochondrial DNA methylation. *Front Genet*, **8**, 166.
20. Owa,C., Poulin,M., Yan,L. and Shioda,T. (2018) Technical adequacy of bisulfite sequencing and pyrosequencing for detection of mitochondrial DNA methylation: sources and avoidance of false-positive detection. *PLoS One*, **13**, e0192722.
21. Frommer,M., McDonald,L.E., Millar,D.S., Collis,C.M., Watt,F., Grigg,G.W., Molloy,P.L. and Paul,C.L. (1992) A genomic sequencing protocol that yields a positive display of 5-methylcytosine residues in individual DNA strands. *Proc. Natl. Acad. Sci. U.S.A.*, **89**, 1827–1831.
22. Krueger,F., Kreck,B., Franke,A. and Andrews,S.R. (2012) DNA methylome analysis using short bisulfite sequencing data. *Nat. Methods*, **9**, 145–151.
23. Darst,R.P., Pardo,C.E., Ai,L., Brown,K.D. and Kladde,M.P. (2010) Bisulfite sequencing of DNA. *Curr. Protoc. Mol. Biol.*, https://doi.org/10.1002/0471142727.mb0709s91.
24. Warnecke,P.M., Stirzaker,C., Song,J., Grunau,C., Melki,J.R. and Clark,S.J. (2002) Identification and resolution of artifacts in bisulfite sequencing. *Methods*, **27**, 101–107.
25. Olova,N., Krueger,F., Andrews,S., Oxley,D., Berrens,R.V., Branco,M.R. and Reik,W. (2018) Comparison of whole-genome bisulfite sequencing library preparation strategies identifies sources of biases affecting DNA methylation data. *Genome Biol.*, **19**, 33.
26. Lister,R., Pelizzola,M., Dowen,R.H., Hawkins,R.D., Hon,G., Tonti-Filippini,J., Nery,J.R., Lee,L., Ye,Z., Ngo,Q.M. *et al.* (2009) Human DNA methylomes at base resolution show widespread epigenomic differences. *Nature*, **462**, 315–322.
27. Cokus,S.J., Feng,S., Zhang,X., Chen,Z., Merriman,B., Haudenschild,C.D., Pradhan,S., Nelson,S.F., Pellegrini,M. and Jacobsen,S.E. (2008) Shotgun bisulphite sequencing of the arabidopsis genome reveals DNA methylation patterning. *Nature*, **452**, 215–219.
28. Bock,C. (2012) Analysing and interpreting DNA methylation data. *Nat. Rev. Genet.*, **13**, 705–719.

29. Xi,Y. and Li,W. (2009) BSMAP: whole genome bisulfite sequence MAPping program. *BMC Bioinformatics*, **10**, 232.
30. Wu,T.D. and Nacu,S. (2010) Fast and SNP-tolerant detection of complex variants and splicing in short reads. *Bioinformatics*, **26**, 873–881.
31. Frith,M.C., Mori,R. and Asai,K. (2012) A mostly traditional approach improves alignment of bisulfite-converted DNA. *Nucleic Acids Res.*, **40**, e100.
32. Krueger,F. and Andrews,S.R. (2011) Bismark: a flexible aligner and methylation caller for bisulfite-seq applications. *Bioinformatics*, **27**, 1571–1572.
33. Guo,W., Fiziev,P., Yan,W., Cokus,S., Sun,X., Zhang,M.Q., Chen,P.Y. and Pellegrini,M. (2013) BS-Seeker2: a versatile aligning pipeline for bisulfite sequencing data. *BMC Genomics*, **14**, 774.
34. Grehl,C., Wagner,M., Lemnian,I., Glaser,B. and Grosse,I. (2020) Performance of mapping approaches for whole-genome bisulfite sequencing data in crop plants. *Front. Plant Sci.*, **11**, 176.
35. Tran,H., Porter,J., Sun,M.A., Xie,H. and Zhang,L. (2014) Objective and comprehensive evaluation of bisulfite short read mapping tools. *Adv. Bioinformatics*, **2014**, 472045.
36. Tsuji,J. and Weng,Z. (2016) Evaluation of preprocessing, mapping and postprocessing algorithms for analyzing whole genome bisulfite sequencing data. *Brief. Bioinform.*, **17**, 938–952.
37. Cagnone,G.L., Tsai,T.S., Makanji,Y., Matthews,P., Gould,J., Bonkowski,M.S., Elgass,K.D., Wong,A.S., Wu,L.E., McKenzie,M. *et al.* (2016) Restoration of normal embryogenesis by mitochondrial supplementation in pig oocytes exhibiting mitochondrial DNA deficiency. *Sci. Rep.*, **6**, 23229.
38. Li,H., Handsaker,B., Wysoker,A., Fennell,T., Ruan,J., Homer,N., Marth,G., Abecasis,G., Durbin,R. and Processing,GenomeProjectData, S. (2009) The sequence alignment/map format and SAMtools. *Bioinformatics*, **25**, 2078–2079.
39. Li,L.C. and Dahiya,R. (2002) MethPrimer: designing primers for methylation PCRs. *Bioinformatics*, **18**, 1427–1431.
40. Wickham,H. (2009) In: *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag, NY.
41. Anderson,S., Bankier,A.T., Barrell,B.G., de Bruijn,M.H., Coulson,A.R., Drouin,J., Eperon,I.C., Nierlich,D.P., Roe,B.A., Sanger,F. *et al.* (1981) Sequence and organization of the human mitochondrial genome. *Nature*, **290**, 457–465.
42. Barroso Lima,N.C. and Prosdocimi,F. (2018) The heavy strand dilemma of vertebrate mitochondria on genome sequencing age: number of encoded genes or g + t content? *Mitochondrial DNA A DNA Mapp. Seq. Anal.*, **29**, 300–302.
43. Taanman,J.W. (1999) The mitochondrial genome: structure, transcription, translation and replication. *Biochim. Biophys. Acta*, **1410**, 103–123.
44. Raine,A., Manlig,E., Wahlberg,P., Syvanen,A.C. and Nordlund,J. (2017) SPlinted ligation adapter tagging (SPLAT), a novel library preparation method for whole genome bisulphite sequencing. *Nucleic Acids Res.*, **45**, e36.
45. Morris,M.J., Hesson,L.B., Poulos,R.C., Ward,R.L., Wong,J.W.H. and Youngson,N.A. (2018) Reduced nuclear DNA methylation and mitochondrial transcript changes in adenomas do not associate with mtDNA methylation. *Biomark. Res.*, **6**, 37.
46. Raizis,A.M., Schmitt,F. and Jost,J.P. (1995) A bisulfite method of 5-methylcytosine mapping that minimizes template degradation. *Anal. Biochem.*, **226**, 161–166.
47. Dhatterwal,P., Mehrotra,S. and Mehrotra,R. (2017) Optimization of PCR conditions for amplifying an AT-rich amino acid transporter promoter sequence with high number of tandem repeats from arabidopsis thaliana. *BMC Res Notes*, **10**, 638.
48. Su,X.Z., Wu,Y., Sifri,C.D. and Wellems,T.E. (1996) Reduced extension temperatures required for PCR amplification of extremely A+T-rich DNA. *Nucleic Acids Res.*, **24**, 1574–1575.
49. Vega-Vaquero,A., Bonora,G., Morselli,M., Vaquero-Sedas,M.I., Rubbi,L., Pellegrini,M. and Vega-Palas,M.A. (2016) Novel features of telomere biology revealed by the absence of telomeric DNA methylation. *Genome Res.*, **26**, 1047–1056.
50. Jin,S.G., Kadam,S. and Pfeifer,G.P. (2010) Examination of the specificity of DNA methylation profiling techniques towards 5-methylcytosine and 5-hydroxymethylcytosine. *Nucleic Acids Res.*, **38**, e125.
51. Ghosh,S., Sengupta,S. and Scaria,V. (2014) Comparative analysis of human mitochondrial methylomes shows distinct patterns of epigenetic regulation in mitochondria. *Mitochondrion*, **18**, 58–62.
52. Duan,J.E., Jiang,Z.C., Alqahtani,F., Mandoiu,I., Dong,H., Zheng,X., Marjani,S.L., Chen,J. and Tian,X.C. (2019) Methylome dynamics of bovine gametes and in vivo early embryos. *Front Genet*, **10**, 512.
53. Stroud,H., Do,T., Du,J., Zhong,X., Feng,S., Johnson,L., Patel,D.J. and Jacobsen,S.E. (2014) Non-CG methylation patterns shape the epigenetic landscape in arabidopsis. *Nat. Struct. Mol. Biol.*, **21**, 64–72.
54. Guo,J.U., Su,Y., Shin,J.H., Shin,J., Li,H., Xie,B., Zhong,C., Hu,S., Le,T., Fan,G. *et al.* (2014) Distribution, recognition and regulation of non-CpG methylation in the adult mammalian brain. *Nat. Neurosci.*, **17**, 215–222.
55. Patil,V., Ward,R.L. and Hesson,L.B. (2014) The evidence for functional non-CpG methylation in mammalian cells. *Epigenetics*, **9**, 823–828.
56. Gu,C., Liu,S., Wu,Q., Zhang,L. and Guo,F. (2019) Integrative single-cell analysis of transcriptome, DNA methylome and chromatin accessibility in mouse oocytes. *Cell Res.*, **29**, 110–123.
57. Guo,F., Li,L., Li,J., Wu,X., Hu,B., Zhu,P., Wen,L. and Tang,F. (2017) Single-cell multi-omics sequencing of mouse early embryos and embryonic stem cells. *Cell Res.*, **27**, 967–988.
58. Rhoads,A. and Au,K.F. (2015) PacBio sequencing and its applications. *Genomics Proteomics Bioinformatics*, **13**, 278–289.
59. Soorni,A., Haak,D., Zaitlin,D. and Bombarely,A. (2017) Organelle_PBA, a pipeline for assembling chloroplast and mitochondrial genomes from pacbio DNA sequencing data. *BMC Genomics*, **18**, 49.
60. Aminuddin,A., Ng,P.Y., Leong,C.O. and Chua,E.W. (2020) Mitochondrial DNA alterations may influence the cisplatin responsiveness of oral squamous cell carcinoma. *Sci. Rep.*, **10**, 7885.
61. Goldsmith,C., Rodriguez-Aguilera,J.R., El-Rifai,I., Jarretier-Yuste,A., Hervieu,V., Raineteau,O., Saintigny,P., Chagoya de Sanchez,V., Dante,R., Ichim,G. *et al.* (2021) Low biological fluctuation of mitochondrial CpG and non-CpG methylation at the single-molecule level. *Sci. Rep.*, **11**, 8032.