nature protocols

PROTOCOL

https://doi.org/10.1038/s41596-018-0090-x

Resolution of the DNA methylation state of single CpG dyads using in silico strand annealing and WGBS data

Chenhuan Xu and Victor G. Corces*

Whole-genome bisulfite sequencing (WGBS) has been widely used to quantify cytosine DNA methylation frequency in an expanding array of cell and tissue types. Because of the denaturing conditions used, this method ultimately leads to the measurement of methylation frequencies at single cytosines. Hence, the methylation frequency of CpG dyads (two complementary CG dinucleotides) can be only indirectly inferred by overlaying the methylation frequency of two cytosines measured independently. Furthermore, hemi-methylated CpGs (hemiCpGs) have not been previously analyzed in WGBS studies. We recently developed in silico strand annealing (iSA), a bioinformatics method applicable to WGBS data, to resolve the methylation status of CpG dyads into unmethylated, hemi-methylated, and methylated. HemiCpGs account for 4–20% of the DNA methylome in different cell types, and some can be inherited across cell divisions, suggesting a role as a stable epigenetic mark. Therefore, it is important to resolve hemiCpGs from fully methylated CpGs in WGBS studies. This protocol describes step-by-step commands to accomplish this task, including dividing alignments by strand, pairing alignments between strands, and extracting single-fragment methylation calls. The versatility of iSA enables its application downstream of other WGBS-related methods such as nasBS-seq (nascent DNA bisulfite sequencing), ChIP-BS-seq (ChIP followed by bisulfite sequencing), TAB-seq, oxBS-seq, and fCAB-seq. iSA is also tunable for analyzing the methylation status of cytosines in any sequence context. We exemplify this flexibility by uncovering the single-fragment non-CpG methylome. This protocol provides enough details for users with little experience in bioinformatic analysis and takes 2–7 h.

Introduction

In WGBS and other related genome-wide methods, genomic DNA samples are subject to a process called bisulfite conversion to chemically distinguish methylated cytosines (Cs) from unmethylated Cs (unmethylated Cs are converted to uracil, whereas methylated Cs are resistant to conversion). The converted samples are used to construct DNA libraries and are sequenced using next-generation sequencing (NGS) technologies. After aligning the sequence reads to both a reference genome and a virtually converted one, the methylation state of each C mapped by the reads can be determined. When the same C is mapped by a sufficient number of different reads, its methylation frequency can be represented by the ratio of methylated events to all mapped events. This cytosine-centric nature of WGBS makes it impossible to directly measure the methylation status of CpG dyads, as the methylation information of the two Cs in a CpG dyad is no longer linked in the data. This missing information is critical to understanding the process of maintenance methylation, by which the methylation information of parental strands is copied to the nascent strands during DNA replication[1,2].

Development of the protocol

In WGBS and other related methods[1,3,4], the double-stranded DNA (dsDNA) fragments are end-repaired to be blunt at both ends before denaturation during bisulfite conversion, ensuring that the two DNA strands within the same fragments share the same genomic coordinates at both ends once sequenced and aligned to the reference genome. Based on this, we recently developed a bioinformatics method, iSA, to computationally resolve the preexisting WGBS datasets into single-CpG methylomes[1]. By searching pairs of alignments between Watson and Crick strands that share the same genomic coordinates, iSA unambiguously determines the methylation state of single CpG dyads

Department of Biology, Emory University, Atlanta, GA, USA. *e-mail: vgcorces@gmail.com

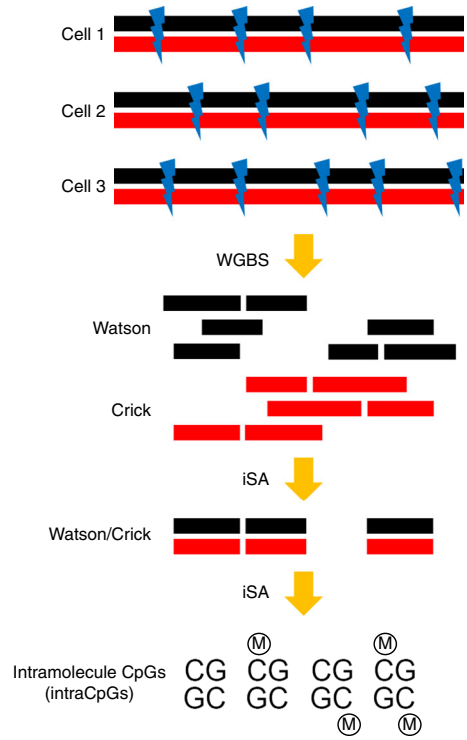

Fig. 1 | Experimental principles underlying the application of iSA. Genomic regions are randomly sheared to dsDNA fragments in different cells. Strand-specific analysis of WGBS data aligns reads to either the Watson (black) or the Crick (red) strand. iSA finds those pairs of alignments between the Watson and Crick strands sharing exactly the same two ends; these represent preexisting dsDNA fragments. The methylation state of intramolecule CpGs (intraCpGs) can be determined by reconstituting the methylation state of the two cytosines in a CpG dyad. M, methyl group.

(Fig. 1). Combining iSA with nasBS-seq and WGBS, we have shown that hemi-methylation is a substantial component of the DNA methylome in all cell types examined, and hemiCpGs flanking CCCTC-binding factor (CTCF)/cohesin co-occupied sites in pluripotent cells are inherited across cell divisions, suggesting a role as a form of stable epigenetic mark[1].

Applications of the method

In principle, iSA can be used downstream of WGBS and other related methods, such as nasBS-seq[1] and ChIP-BS-seq[3,4]. In the special case of RRBS (reduced representation bisulfite sequencing)[5], another commonly used method to measure genome-wide DNA methylation frequency, a certain (or a combination of) sequence-specific restriction enzyme is used to digest whole-genomic DNA, making all genomic fragments from the same location share exactly the same genomic coordinates at both ends after aligning. In this case, iSA is unable to distinguish and pair alignments from two strands that used to be in the same dsDNA fragment. However, this can be overcome by introducing a strategy of 'unique molecular identifiers' when performing RRBS to tag each genomic fragment with a unique barcode[6], and by taking the unique barcode into account when pairing alignments using iSA.

Other genome-wide methods have been developed by altering the step of bisulfite conversion to specifically convert a certain oxidized form of methylated Cs and thus to directly measure its frequency (Tet-assisted bisulfite sequencing (TAB-seq)[7]), or specifically not to convert an oxidized form and to indirectly infer its frequency by comparison with a more inclusive dataset (oxidative bisulfite sequencing (oxBS-seq)[8], 5fC chemically assisted bisulfite sequencing (fCAB-seq)[9]). Because the DNA fragment ends are well preserved before conversion, iSA can also be applied downstream of all these methods to resolve the data into oxidized methylomes of single CpGs.

Comparison with other methods

To our knowledge, the only available method that serves a purpose similar to that of iSA is hairpin–bisulfite sequencing[10,11], with only one report of its application in genome-wide studies[12]. Whereas hairpin–bisulfite sequencing requires very intensive de novo sequencing efforts to achieve a

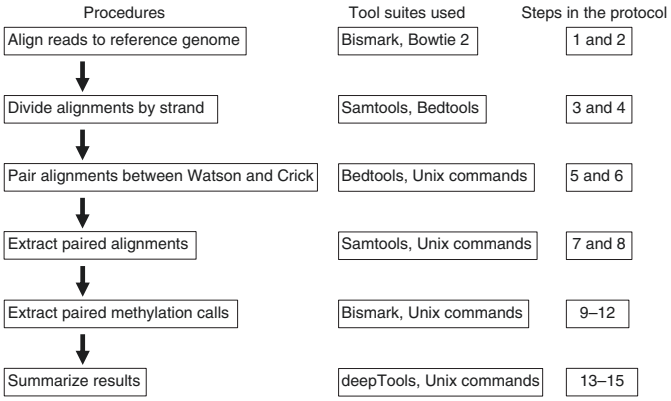

Fig. 2 | Overview of iSA. Key procedures performed in iSA, tool suites used, and the corresponding steps in this protocol.

high coverage of mammalian genomes[12], iSA can be applied downstream of preexisting WGBS datasets to computationally resolve the data into single-CpG methylomes. We recently modified the genome-wide hairpin–bisulfite sequencing method to enable its application downstream of a ChIP assay (ChIP–hairpinBS-seq)[1]. With much smaller sequencing efforts than those needed for genome-wide hairpinBS-seq, ChIP–hairpinBS-seq maps a single-CpG methylome of genomic regions occupied by a protein of interest and can be used as an independent method to verify the single-CpG methylomes from iSA-resolved WGBS datasets.

Overview of the protocol

In WGBS, paired-end reads are aligned to either the Watson or the Crick strand, using Bismark[13] (Fig. 2, Step 1). After cleaning the data by de-duplication (Fig. 2, Step 2), each alignment possesses a unique set of genomic coordinates (the genomic positions of the most 5′ base of the two mate reads) on either the Watson or the Crick strand. In iSA, an iterative search is performed between alignments on the Watson and Crick strands using Samtools[14] and Bedtools[15] (Fig. 2, Steps 3–5). Any pair of alignments between the Watson and Crick strands that share exactly the same genomic coordinates at both ends are retained for downstream analysis. It should be noted that random shearing of the bulk chromatin from a population of cells can probably yield some pairs of alignments between the Watson and Crick strands that share the same ends but are of distinct dsDNA origin. In this protocol, we use the mean count from pairing reads with −30-, −20-, −10-, 10-, 20-, and 30-bp distances between the ends of the two aligned strands to represent the level of random pairing (Fig. 3). To determine the possible interference from such random pairing, the same searching process is performed by counting the pairs of alignments between reads with the same distance of bases between the ends, and comparing the results with the number of same-end pairs, using Bedtools (Fig. 2, Step 6). In a typical WGBS dataset with sufficient genome coverage and sequencing depth (Experimental design), we usually observe a 20- to 100-fold enrichment of same-end over random pairing, suggesting a very small interference from the latter.

The DNA methylation calls are extracted from same-end paired alignments, using Samtools, Bismark, and Bedtools (Fig. 2, Steps 7–11). Each mapped C in a CpG context on one strand has a counterpart C mapped on the other strand. Guaranteed by a high fold enrichment over random pairing, these CpG dyads largely represent physically existing CpG dyads within dsDNA fragments during the early phases of WGBS and are termed ‘intraCpGs’ (intramolecule CpGs)[1]. Thus, the methylation status of intraCpGs can be determined to be one of the four types: unmethylated, hemi-methylated with C methylated on either the Watson or Crick strand, or methylated. In addition, the methylation state of Cs in a CHG context (in which ‘H’ stands for A or T or C) is extracted and paired between the Watson and Crick strands. The single-fragment methylation status of CAG/CTG can also be determined to be one of four types: unmethylated, hemi-methylated with C methylated on either the Watson or the Crick strand, or methylated.

Expertise needed to implement the protocol

This protocol requires only basic knowledge of NGS data analysis and beginner’s proficiency in working in a Unix shell terminal window. We encourage inexperienced users to become familiar with

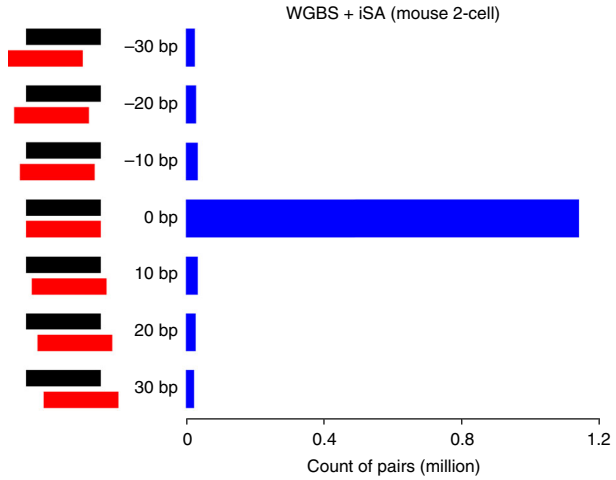

Fig. 3 | Fold enrichment of same ends over random pairing. Pairs of same-length strand alignments with differential end pairing were searched and counted between Watson (black) and Crick (red) strands in the WGBS dataset from mouse two-cell embryos. The center bar represents the number of same-end pairs of alignments with 0-bp distance between the Watson and Crick strands. The other bars represent the number of randomly paired alignments with arbitrarily chosen distances of −30, −20, −10, 10, 20, and 30 bp between the ends of the Watson and Crick strands.

the manual pages of the tool suites[13–18] and help information for Unix commands used in this protocol. An interactive shell script is available at https://github.com/chxu02/iSA as preliminary training for beginners. For users with higher proficiency, tuning of the script is encouraged for customizing purposes. As an example of this tunability, we also present commands (Step 12) for extracting the single-molecule non-CpG methylome from WGBS datasets using iSA.

Limitations

iSA requires information about the genomic coordinates of both ends of alignments provided by a pair of mate reads from paired-end sequencing. This information is missing in experiments using single-end sequencing strategies. Thus, iSA can be applied only downstream of WGBS and related datasets sequenced using paired-end strategies.

We have found that the performance of iSA improves when more sequencing reads are obtained with the same WGBS library. To illustrate this, we divided the 569 million paired-end reads in the example WGBS dataset into 20 equal fractions, each 5% of the total, and aligned them to the mouse reference genome in an additive manner: (i) align the first fraction (5% of reads); (ii) de-duplicate the alignments by removing redundant alignments originating from PCR amplification of the same DNA molecules; (iii) align the second fraction and merge with alignments from the last step; (iv) de-duplicate the alignments; (v) align the third fraction and so on for the remaining fractions. As expected, with more reads added, the duplication rate increases. Both the aggregate duplication rate (the percentage of all duplicated alignments out of all alignments in accumulated fractions of reads) and the real-time duplication rate (the percentage of duplicated alignments in a fraction out of all alignments in the accumulated fractions of reads) suggest improvement of sequencing depth. Interestingly, this trend is accompanied by two performance enhancements in iSA, increased pairing efficiency (fraction of alignments that can be paired) (Fig. 4a), and increased fold enrichment over random pairing (Fig. 4b). Thus, by having more DNA molecules in the library sequenced, there is a higher probability that an alignment on one strand can be paired with another alignment on the other strand, and a higher confidence that the paired alignments represent genuine dsDNA fragments.

We have also noticed that iSA usually has a low pairing efficiency (up to 3%), even when the input dataset has a very high sequencing depth (Fig. 4a). This is most likely attributable to the usually inevitable DNA degradation during the harsh bisulfite conversion process[19]. The degradation of Watson single-strand DNA and that of Crick single-strand DNA are independent of each other under denaturing conditions, further compounding the issue because iSA requires the integrity of both strands in a dsDNA fragment. The pairing efficiency of iSA may be improved in the future by the introduction of new ways to carry out bisulfite conversion during WGBS library preparation.

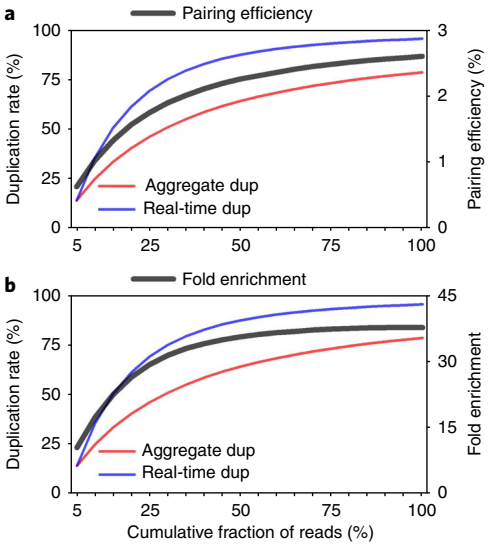

Fig. 4 | Enhanced performance of iSA with increased sequencing depth. a,**b**, The pairing efficiency (the fraction of alignments that can be paired) (**a**) and fold enrichment over random pairing (the ratio of the number of same-end alignments over the number of randomly paired alignments) (**b**) of iSA is enhanced when more reads are included in the analysis. Aggregate dup, the percentage of all duplicated alignments out of all alignments in accumulated fractions of reads; Real-time dup, the percentage of duplicated alignments in a fraction out of all alignments in accumulated fractions of reads. Aggregate dup is the common way of representing duplication rate by many software packages, whereas real-time dup is more informative when one needs to determine if it is cost-efficient to continue sequencing a previously sequenced library.

Experimental design

Choice of upstream experiments and sequencing format

iSA is not recommended for datasets of libraries prepared using the post-bisulfite adapter tagging method, in which bisulfite treatment precedes adapter tagging[20]. The substantial DNA degradation during bisulfite conversion leads to alteration of the terminal sequence context of most dsDNA fragments, which after tagging, sequencing, and aligning, leads to loss of information on genomic coordinates of the intact dsDNA fragments. In addition, single-end sequencing is not compatible with iSA (Limitations); paired-end sequencing is highly preferred over single-end sequencing when designing WGBS experiments, considering the comparable cost between the two (e.g., 50-bp paired ends versus 100-bp single ends).

Processing the read files

Although sequencing may generate bases with differential quality scores and may read through into sequencing adapters, Bismark assumes that all base content in the reads is of sample origin and has the highest quality score. Hence, it is a prerequisite to process the reads and eliminate unwanted bases before aligning. We encourage users to use very stringent cutoffs to search and remove any remnant adapter sequence in the reads (Equipment setup). For the 5′-ends of the reads, differential trimming by quality is highly discouraged. Doing so will blur the 5′-ends of the reads and lead to misperformance of two processes that rely on the accuracy of the 5′-end genomic coordinates of the alignments: (i) removal of duplicated alignments from duplicated reads that have the same sequence content but different quality scores at the 5′-end, and hence can escape from de-duplication after differential trimming of the 5′-end, and (ii) correctly pairing alignments through iSA, which may fail because of the compromised 5′-end coordinates after differential trimming. In cases of uniform trimming of the 5′-end, iSA provides an option to take this into account (see Step 4 of the Procedure). For the 3′-ends of reads, either differential or uniform trimming is tolerated.

Applying iSA to ChIP-BS-seq

We noticed that a lower fold enrichment may be observed when iSA is applied downstream of some genomic methods with enriched alignments at certain genomic regions, such as ChIP-BS-seq.

At genomic regions with relative enrichment of alignments, such as protein binding sites (peaks) in ChIP-BS-seq, the data are expected to have more randomly paired alignments, which lowers the fold enrichment in iSA. In this case, users are encouraged to validate results from iSA by using independent methods such as ChIP–hairpinBS-seq[1].

Materials

Equipment

Hardware

- 64-bit computer running the Linux operating system, with an eight-core processor (a 48-core processor is preferred) and 64 GB of RAM (256 GB is preferred)

Software ▲ **CRITICAL** The example dataset has been tested with the latest versions of the software specified below. We encourage users to install or upgrade to these versions.

- SRA Toolkit (http://ftp-trace.ncbi.nlm.nih.gov/sra/sdk/2.9.2/sratoolkit.2.9.2-ubuntu64.tar.gz), for retrieving sequence reads from the Sequence Read Archive (SRA)
- Trimmomatic[16] (http://www.usadellab.org/cms/uploads/supplementary/Trimmomatic/Trimmomatic-0.38.zip), for trimming reads to remove unwanted bases
- FastQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/fastqc_v0.11.7.zip), for reviewing the quality of reads
- Bismark[13] (http://www.bioinformatics.babraham.ac.uk/projects/bismark/bismark_v0.20.0.tar.gz), for aligning WGBS reads to a reference genome through Bowtie 2 and extracting DNA methylation calls
- Bowtie 2 (ref. [17], https://sourceforge.net/projects/bowtie-bio/files/bowtie2/2.2.9/)
- Bedtools[15] (https://github.com/arq5x/bedtools2/releases/download/v2.27.0/bedtools-2.27.0.tar.gz), for manipulation of files with genomic intervals
- Samtools[14] (https://github.com/samtools/samtools/releases/download/1.9/samtools-1.9.tar.bz2), for manipulation of alignment files in sam/bam format
- deepTools[18] (https://github.com/deeptools/deepTools/archive/2.5.4.tar.gz), for post-iSA visualization of data

Data

- Example WGBS dataset (GSM1386021) is available at SRA
- Mouse reference genome sequence files (Equipment setup)
- *lambda* reference genome sequence file (Equipment setup)

Equipment setup

Example dataset

In this protocol, we use an example WGBS dataset of mouse two-cell-stage embryos[21] for the following reasons. This dataset has 569 million 100-bp paired-end reads after trimming, providing a high sequencing depth and coverage of the mouse genome. Furthermore, under a high unique alignment rate of 74% (the fraction of unambiguously mapped reads out of all reads), the alignments piled up to reach an aggregate duplication rate of 79% (the fraction of redundant alignments out of all unambiguous alignments), enabling us to investigate the relationship between sequencing depth and performance of iSA (see discussion above).

Downloading and pre-processing the example dataset

This WGBS dataset has five sequencing runs (SRR1286778–SRR1286782). The `fastq-dump` command in the SRA Toolkit is used to retrieve the reads. Each sequencing run gives rise to two mate read files (e.g., SRR1286778_1.fastq and SRR1286778_2.fastq):

```
$ fastq-dump --split-3 SRR1286778 SRR1286779 SRR1286780 SRR1286781
SRR1286782
```

Reads from all sequencing runs (e.g., technical replicates) from the same biological replicate should be concatenated together before aligning and de-duplication, to avoid retaining duplicated alignments between technical replicates. To do so, type each of the following two

command lines into separate terminal windows to obtain two concatenated read files, m2C_1.fq and m2C_2.fq:

```
$ cat SRR1286778_1.fastq SRR1286779_1.fastq SRR1286780_1.fastq
SRR1286781_1.fastq SRR1286782_1.fastq > m2C_1.fq
$ cat SRR1286778_2.fastq SRR1286779_2.fastq SRR1286780_2.fastq
SRR1286781_2.fastq SRR1286782_2.fastq > m2C_2.fq
```

Use Trimmomatic to perform the trimming step and FastQC to review the trimming results:

```
$ java -jar /Trimmomatic-0.38/trimmomatic-0.38.jar PE -threads 8
m2C_1.fq m2C_2.fq m2C_1_trim.fq s1 m2C_2_trim.fq s2
ILLUMINACLIP:/Trimmomatic-0.38/adapters/TruSeq3-PE-2.fa:0:0:2
TRAILING:20 MINLEN:20
$ fastqc -t 2 --nogroup m2C_1_trim.fq m2C_2_trim.fq
```

m2C_1_trim.fq and m2C_2_trim.fq are the two mate read files after trimming. s1 and s2 are two small subsets of unpaired reads after trimming and can be discarded. FastQC generates an .html report for each mate read file. Review the 'Per base sequence quality' result in the reports to make sure that all remaining bases have a Phred quality score >20, and the 'Adapter Content' result to make sure that there is no adapter sequence contamination. If low-quality bases are still present (e.g., Phred score <20), re-run Trimmomatic, using the following command, and run FastQC to review the results:

```
$ java -jar /Trimmomatic-0.38/trimmomatic-0.38.jar PE -threads 8
m2C_1.fq m2C_2.fq m2C_1_trim.fq s1 m2C_2_trim.fq s2
ILLUMINACLIP:/Trimmomatic-0.38/adapters/TruSeq3-PE-2.fa:0:0:2
TRAILING:20 SLIDINGWINDOW:3:20 MINLEN:20
```

▲ **CRITICAL** All commands in this protocol are meant to be run from the Unix shell prompt in a terminal window.

Downloading the mouse reference genome sequence files and building index files

All files ending in '.fa.gz' under http://hgdownload.cse.ucsc.edu/goldenPath/mm10/chromosomes/ are downloaded to a new directory (e.g., 'mm10_bismark') and decompressed with this command:

```
$ gzip -d mm10_bismark/*.fa.gz
```

The index files needed for aligning WGBS reads to the mouse reference genome are built by Bismark as follows:

```
$ bismark_genome_preparation --verbose mm10_bismark
```

Downloading the *lambda* reference genome sequence file and building index files

The *lambda* genome sequence file can be downloaded from https://github.com/chxu02/iSA/blob/master/lambda.fa to a new directory (e.g., 'lambda_bismark'). The index files needed for aligning WGBS reads to the *lambda* reference genome are built by Bismark as follows:

```
$ bismark_genome_preparation --verbose lambda_bismark
```

Procedure

Align reads to the reference genome ● Timing 1–5 d, depending on the size of the dataset and the computational capability of hardware

1 Align the trimmed WGBS reads in the files 'm2C_1_trim.fq' and 'm2C_2_trim.fq' to the mouse reference genome:

```
$ bismark --multicore 3 --score_min L,0,-0.4 mm10_bismark -1
m2C_1_trim.fq -2 m2C_2_trim.fq
```

This step generates an alignment file with the name of 'm2C_1_trim_bismark_bt2_pe.bam' and a report file.

Alternatively, if the library was prepared with spike-in *lambda* DNA, run these two commands sequentially instead of the above one:

```
$ bismark --multicore 3 --un --ambiguous --score_min L,0,-0.4
mm10_bismark -1 m2C_1_trim.fq -2 m2C_2_trim.fq
$ bismark --multicore 3 --score_min L,0,-0.4 lambda_bismark -1
m2C_1_trim.fq_unmapped_reads_1.fq.gz -2 m2C_2_trim.fq_unmapped_
reads_2.fq.gz
```

The additionally generated report file 'm2C_1_trim.fq_unmapped_reads_1_bismark_bt2_PE_report.txt' states the false 'methylation' frequency of *lambda* DNA and can be used to estimate the bisulfite conversion rate.

▲ **CRITICAL STEP** This step is extremely time- and resource-consuming. Given sufficient computational resources on the hardware, increasing the number after `--multicore` will greatly reduce the running time. In addition, make sure that the current hard drive has sufficient free space (>700 GB in this case, proportional to the size of the dataset). If not, add the `--gzip` option to compress the temporary files (>300 GB of free space is still required in this case), or use the `--temp_dir` option to direct the writing of temporary files to another hard drive with sufficient free space. The `--score_min` option controls the quality of reads that can be successfully mapped.

? TROUBLESHOOTING

2 Remove the duplicated alignments:

```
$ deduplicate_bismark -p --bam m2C_1_trim_bismark_bt2_pe.bam
```

This step generates a file with the name 'm2C_1_trim_bismark_bt2_pe.deduplicated.bam'.

▲ **CRITICAL STEP** Steps 3 through 13 are wrapped in a shell script 'iSA.sh', available at https://github.com/chxu02/iSA. Run

```
$ iSA.sh
```

to obtain instructions on usage of the script and modifying parameters. Steps 3–13 of the Procedure break down the script into individual command lines to show what each step does and how to modify the parameters if necessary. The shell script uses the .bam file generated in Step 2 as input. Using the shell script or the individual command lines generates the same two output files 'm2C.intraCpG.bed' and 'm2C.intraCWG.bed', which contain the intramolecule methylation status of CpG and non-CpG, respectively. The shell script generates an additional report file with all statistics throughout iSA. First-time users are encouraged to use the shell script for the dedicated reporting feature.

Divide the alignments by strand ● Timing 0.5–2 h depending on the size of the dataset

3 Divide the alignments into two subsets. The two commands can be run simultaneously in two terminal windows to generate two output files, 'm2C.Wat.bam' and 'm2C.Cri.bam', containing alignments to the Watson and Crick strands, respectively:

```
$ samtools view -h m2C_1_trim_bismark_bt2_pe.deduplicated.bam | awk
'$3!="chrM" && $16!="XG:Z:GA"' | samtools view -bh -@ 8 - > m2C.Wat.bam
$ samtools view -h m2C_1_trim_bismark_bt2_pe.deduplicated.bam | awk
'$3!="chrM" && $16!="XG:Z:CT"' | samtools view -bh -@ 8 - > m2C.Cri.bam
```

4 Generate .bed files with genomic coordinates for each alignment from each strand:

```
$ bamToBed -bedpe -i m2C.Wat.bam | awk '{if($2<t1){$9=0} else{$9=
$2-t2}; print $1,$9,$6+t2,NR,$7}' OFS='/t' | sort -k1,1 -k2,2n -k3,
3n -u -S 64G > m2C.Wat.bed
```

```
$ bamToBed -bedpe -i m2C.Cri.bam | awk '{if($2<t2){$9=0} else{$9=
$2-t2}; print $1,$9,$6+t1,NR,$7}' OFS='/t' | sort -k1,1 -k2,2n -k3,3n
-u -S 64G > m2C.Cri.bed
```

▲ **CRITICAL STEP** If uniform trimming of the 5′-ends of the reads was performed before alignment, it must be taken into account at this step by replacing and with the number of bases uniformly trimmed from the 5′-ends of the mate 1 and mate 2 reads, respectively (Experimental design). Use 0 for t1 and t2 if no 5′-end trimming was performed.

Pair alignments between the Watson and Crick strands ● Timing 1–3 h, depending on the size of the dataset

5 For pairing alignments between Watson and Crick strands sharing exactly the same two ends, use the following:

```
$ intersectBed -a m2C.Wat.bed -b m2C.Cri.bed -wa -wb -f 1 -F 1 -sorted |
cut -f 1-5,9,10 > m2C.iSA.bed
$ wc -l m2C.iSA.bed
```

The first command generates a file, 'm2C.iSA.bed', containing the successfully paired alignments. The second command reports the number of such successful pairs.

? TROUBLESHOOTING

6 Use the following code to pair alignments between Watson and Crick strands with the same base distance at both ends. These commands can be run in different terminal windows simultaneously or in the same window sequentially.

```
$ awk '{if($2>=30) print $1,$2-30,$3-30}' OFS='/t' m2C.Wat.bed |
intersectBed -a - -b m2C.Cri.bed -wa -wb -f 1 -F 1 -sorted | wc -l
$ awk '{if($2>=20) print $1,$2-20,$3-20}' OFS='/t' m2C.Wat.bed |
intersectBed -a - -b m2C.Cri.bed -wa -wb -f 1 -F 1 -sorted | wc -l
$ awk '{if($2>=10) print $1,$2-10,$3-10}' OFS='/t' m2C.Wat.bed |
intersectBed -a - -b m2C.Cri.bed -wa -wb -f 1 -F 1 -sorted | wc -l
$ awk '{print $1,$2+10,$3+10}' OFS='/t' m2C.Wat.bed | intersectBed
-a - -b m2C.Cri.bed -wa -wb -f 1 -F 1 -sorted | wc -l
$ awk '{print $1,$2+20,$3+20}' OFS='/t' m2C.Wat.bed | intersectBed
-a - -b m2C.Cri.bed -wa -wb -f 1 -F 1 -sorted | wc -l
$ awk '{print $1,$2+30,$3+30}' OFS='/t' m2C.Wat.bed | intersectBed
-a - -b m2C.Cri.bed -wa -wb -f 1 -F 1 -sorted | wc -l
```

This step calculates the number of random pairs of alignments between Watson and Crick strands with −30-, −20-, −10-, 10-, 20-, and 30-bp distances between ends (Fig. 3). Each command performs a different pairing and prints the information to the screen. In the shell script, the mean value of these numbers is compared with the number of same-end pairs (Step 5) to estimate the putative interference from random pairing. Users are encouraged to proceed to the next step if the ratio of the number of same-end pairs over the number of random pairs is >10.

? TROUBLESHOOTING

Extract paired alignments ● Timing 20–60 min, depending on the size of the dataset

7 Extract line numbers of paired alignments for Watson or Crick strands from the file generated in Step 5. These line numbers will be used by commands in Step 8 to extract paired alignments from the file containing all alignments.

```
$ cut -f 4 m2C.iSA.bed | awk '{print $1*2-1"/n"$1*2}' > m2C.Wat.LN
$ cut -f 6 m2C.iSA.bed | awk '{print $1*2-1"/n"$1*2}' > m2C.Cri.LN
```

8 Run the following three commands sequentially to extract paired alignments for Watson or Crick strands. The files 'm2C.Wat.LN' and 'm2C.Cri.LN' from Step 7 are used to extract

paired alignments from the file containing all alignments. The two output files, 'm2C.Wat.iSA.bam' and 'm2C.Cri.iSA.bam', contain paired alignments for Watson and Crick strands, respectively.

```
$ samtools view -H m2C.Wat.bam > header
$ samtools view m2C.Wat.bam | awk 'FNR==NR {h[$1];next} (FNR in h)'
m2C.Wat.LN - | cat header - | samtools view -bh -@ 8 - > m2C.Wat.iSA.bam
$ samtools view m2C.Cri.bam | awk 'FNR==NR {h[$1];next} (FNR in h)'
m2C.Cri.LN - | cat header - | samtools view -bh -@ 8 - > m2C.Cri.iSA.bam
```

Extract DNA methylation calls ● Timing 20–60 min, depending on the size of the dataset

9 Extract DNA methylation calls from paired alignments for Watson or Crick strands (Step 8):

```
$ bismark_methylation_extractor -p --multicore 8 --no_header --gzip
m2C.Wat.iSA.bam
$ bismark_methylation_extractor -p --multicore 8 --no_header --gzip
m2C.Cri.iSA.bam
```

This step generates three methylation call files (Cs in CpG, CHG, or CHH context) for each strand. The names of the files with methylation calls start with 'CpG_OT', 'CHG_OT', or 'CHH_OT' for the Watson strand, and 'CpG_OB', 'CHG_OB', or 'CHH_OB' for the Crick strand.

▲ **CRITICAL STEP** Users are encouraged to review the M-bias report files (one for each strand) generated at this step. These files report the average methylation frequency of each base position throughout the reads. The bases closest to the 5′-ends of reads (especially on some mate 2 reads) frequently show abnormally low methylation frequency (2–30%) as compared with the average methylation frequency throughout the reads (60–80% for most mammalian cell types). If this is the case, re-run this step using the following two commands:

```
$ bismark_methylation_extractor -p --multicore 8 --no_header --gzip
--ignore_r2 4 m2C.Wat.iSA.bam
$ bismark_methylation_extractor -p --multicore 8 --no_header --gzip
--ignore_r2 4 m2C.Cri.iSA.bam
```

10 Extract alignment IDs for pairing between Watson and Crick strands using the file generated in Step 5. The alignment IDs are used by Step 11 to pair methylation calls between Watson and Crick strands.

```
$ awk '{print $5"/t"NR}' m2C.iSA.bed | sort -k1,1 > m2C.Wat.ID
$ awk '{print $7"/t"NR}' m2C.iSA.bed | sort -k1,1 > m2C.Cri.ID
```

11 Pair the two Cs in the same CpG dyads by their alignment IDs and genomic coordinates. Run these commands sequentially. The first and second commands extract methylation calls from the Watson and Crick strands, respectively. The last command pairs methylation calls between Watson and Crick when they have the same alignment ID and genomic coordinates.

```
$ gzip -cd CpG_OT_m2C.Wat.iSA.txt.gz | sort -k1,1 | join -j 1 - m2C.Wat.
ID | sed 's/ //t/g' | awk '{print $6"_"$3"_"$4,$3,$4-1,$4+1,$5}'
OFS='/t' | sort -k1,1 > m2C.Wat.me
$ gzip -cd CpG_OB_m2C.Cri.iSA.txt.gz | sort -k1,1 | join -j 1 - m2C.Cri.
ID | sed 's/ //t/g' | awk '{print $6"_"$3"_"$4-1,$3,$4-2,$4,$5}'
OFS='/t' | sort -k1,1 > m2C.Cri.me
$ join -j 1 m2C.Wat.me m2C.Cri.me | sed 's/ //t/g' | awk '{if($5=="z" &&
$9=="z"){print $2,$3,$4,1,0,0,0} else if($5=="z" && $9=="Z")
```

```
{print $2,$3,$4,0,1,0,0} else if($5=="Z" && $9=="z"){print $2,$3,
$4,0,0,1,0} else{print $2,$3,$4,0,0,0,1}}' OFS='/t' | sort -k1,1
-k2,2n | groupBy -g 1-3 -c 4,5,6,7 -o sum > m2C.intraCpG.bed
```

12 Pair the two Cs in the same CWG (the two CAG/CTG trinucleotides from two DNA strands) context by their alignment ID and genomic coordinates. Run these commands sequentially. The first and second extract methylation calls from Watson and Crick strands, respectively. The last command pairs methylation calls between Watson and Crick when they have the same alignment ID and genomic coordinates.

```
$ gzip -cd CHG_OT_m2C.Wat.iSA.txt.gz | sort -k1,1 | join -j 1 - m2C.
Wat.ID | sed 's/ //t/g' | awk '{print $6"_"$3"_"$4,$3,$4-1,$4+2,$5}'
OFS='/t' | sort -k1,1 > m2C.Wat.CHG.me
$ gzip -cd CHG_OB_m2C.Cri.iSA.txt.gz | sort -k1,1 | join -j 1 - m2C.
Cri.ID | sed 's/ //t/g' | awk '{print $6"_"$3"_"$4-2,$3,$4-3,$4,$5}'
OFS='/t' | sort -k1,1 > m2C.Cri.CHG.me
$ join -j 1 m2C.Wat.CHG.me m2C.Cri.CHG.me | sed 's/ //t/g' | awk '{if
($5=="x" && $9=="x"){print $2,$3,$4,1,0,0,0} else if($5=="x" &&
$9=="X"){print $2,$3,$4,0,1,0,0} else if($5=="X" && $9=="x")
{print $2,$3,$4,0,0,1,0} else{print $2,$3,$4,0,0,0,1}}' OFS='/t' |
sort -k1,1 -k2,2n | groupBy -g 1-3 -c 4,5,6,7 -o sum > m2C.intraCWG.bed
```

Summarize and visualize the results ● Timing ~5 min

13 Apply the two commands below to summarize intraCpGs/intraCWGs in each methylation state. Each .bed file generated contains the genomic coordinates and methylation status of intramolecule CpGs and CWGs, respectively:

```
$ awk '{To+=$4+$5+$6+$7; Un+=$4; HC+=$5; HW+=$6; Me+=$7} END
{print "You found "To" intraCpGs, of which:/n"Un" are unmethylated,/
n"HW" are hemi-Watson,/n"HC" are hemi-Crick,/n"Me" are methy-
lated."}' m2C.intraCpG.bed
 $ awk '{To+=$4+$5+$6+$7; Un+=$4; HC+=$5; HW+=$6; Me+=$7} END
{print "You found "To" intraCWGs, of which:/n"Un" are unmethylated,/
n"HW" are hemi-Watson,/n"HC" are hemi-Crick,/n"Me" are methy-
lated."}' m2C.intraCWG.bed
```

14 The frequency of CpGs in different methylation states at certain genomic features can be profiled using various published tool suites. As an example, we use deepTools[20] to profile the frequency of hemiCpGs at gene bodies in mouse two-cell-stage embryos.

The .bed file with intraCpGs generated in Step 11 is converted to .bw format, which is accepted by deepTools. To convert the file to .bw, use the following commands:

```
$ awk '{printf "%s/t%.0f/t%.0f/t%.2f/n",$1,$2,$3,($5+$6)/($4+$5+
$6+$7)}' m2C.intraCpG.bed > m2C.intraCpG-hemi.bdg
 $ bedGraphToBigWig m2C.intraCpG-hemi.bdg mm10_chromsize.txt m2C.
intraCpG-hemi.bw
```

15 Use the two commands in deepTools, `computeMatrix` and `plotProfile`, to visualize the result:

```
$ computeMatrix scale-regions -R mm10.refGene.bed -S m2C.intraCpG-
hemi.bw -m 3000 -a 1000 -b 1500 -out m2C.intraCpG-hemi.mtx -bs 50 -p
max
$ plotProfile -m m2C.intraCpG-hemi.mtx -o m2C.intraCpG-hemi.png
--yMin 0 --yMax 0.16
```

Troubleshooting

Troubleshooting advice can be found in Table 1.

Table 1 | Troubleshooting table

Step	Problem	Possible reason	Solution
1	Low rate (<50%) of aligned WGBS reads	Low-quality bases were not removed from reads, or bases from sequencing adapters were not removed from reads	Trim reads thoroughly by quality scores and adapter sequence content. Carefully review the trimming results by FastQC. More basically, when making WGBS libraries, tightly control the size range of fragmented genomic DNA to fit with your sequencing format (e.g., a library with fragments of 200–500 bp is ideal for 100-bp paired-end Illumina sequencing. Instead, a library with 100-bp mean fragment size will lead to substantial contamination of reads by adapter sequences when sequenced in 100-bp paired-end format)
5	Low pairing efficiency (<1%)	Low sequencing depth of library; severe loss of DNA during bisulfite conversion	The pairing efficiency increases with sequencing depth (Experimental design). Additional sequencing to obtain more reads from the same library will improve pairing efficiency. Technically, reduced time of bisulfite conversion may also help. Make sure that reduction of time does not lower the conversion rate by monitoring the conversion rate of spike-in *lambda* DNA
6	Low fold enrichment over random pairing	Differential trimming of 5′-end of reads; uniform trimming of 5′-end of reads not reflected at Step 5; low sequencing depth; suboptimal library preparation method	Avoid differential trimming of 5′-end of reads (Experimental design). If low-quality bases are abundant at 5′-end, perform uniform trimming of 5′-ends of reads and re-run Step 5 by taking this into account. Also, sequencing to obtain more reads from the same library improves fold enrichment over random pairing. For WGBS libraries prepared using the Tn5 transposition system, we have observed an overall low fold enrichment over random pairing, even when the sequencing depth is very high, possibly due to the local enrichment of adapter integration, which in turn leads to a higher chance of random pairing and lower fold enrichment. Last, contact sequencing service providers to see if trimming of reads was performed but not reported to users. In practice, we recommend that users proceed with iSA with datasets showing >10-fold enrichment, which can be translated into a >90% accuracy when calling intraCpGs/intraCWGs
	No fold enrichment over random pairing	Datasets from libraries prepared using PBAT (post-bisulfite adapter tagging) or a similar strategy (bisulfite conversion precedes ligation with adapter) will lead to loss of the information on the original 5′ genomic coordinates of most dsDNA fragments and will generally lead to a <1.5-fold enrichment based on our experience. The shell script provided does a quality check at the very beginning and will inform the user if the dataset is likely to have come from a PBAT experiment	The dataset cannot be analyzed by iSA

Timing

Equipment setup, building index files for mouse reference genome: 4–6 h, depending on the computational capability of the hardware.

Steps 1 and 2, alignment of reads to the reference genome: 1–5 d, depending on the size of the dataset and computational capability of hardware

Steps 3 and 4, division of the alignments by strand: 0.5–2 h, depending on the size of the dataset

Steps 5 and 6, pair alignments between Watson and Crick strands: 1–3 h, depending on the size of the dataset

Steps 7 and 8, extraction of paired alignments: 20–60 min, depending on the size of the dataset

Steps 9–12, extraction of DNA methylation state of intraCpGs/intraCWGs: 20–60 min, depending on the size of the dataset
Steps 13–15, summarization and visualization of results: ~5 min

The first step, aligning the WGBS reads to the reference genome, is the most time-consuming step and relies heavily on the computational capability of the hardware used. With a pre-processed alignment file (bypassing Steps 1 and 2), it usually takes iSA (Steps 3–13) 2–7 h to extract intraCpGs/intraCWGs and their methylation states, with the timing mainly dependent on the size of the dataset.

Anticipated results

The anticipated results from running the shell script 'iSA.sh' are shown below.

Pairing efficiency of iSA

At Step 5, the shell script prints the result in the terminal as shown below:

You found 1139561 pairs of alignments between Watson and Crick.

Pairing efficiency: 2.59% for Watson, 2.61% for Crick.

Fold enrichment over random pairing

At Step 6, the shell script prints the result as shown below:

24452 pairs of alignments found with −30 bp shift,

28825 pairs of alignments found with −20 bp shift,

34578 pairs of alignments found with −10 bp shift,

35353 pairs of alignments found with +10 bp shift,

30792 pairs of alignments found with +20 bp shift,

26536 pairs of alignments found with +30 bp shift,

You got 37.8-fold enrichment over random pairing.

The fold enrichment is calculated as a ratio of counts of same-end pairs over the mean counts of random pairs. The result can also be visualized by a histogram (Fig. 3).

Summary of intraCpGs and intraCWGs in each methylation state

The files 'm2C.intraCpG.bed' and 'm2C.intraCWG.bed' generated by iSA (Steps 11 and 12) contain information on genomic coordinates and methylation states of intraCpGs and intraCWGs, respectively. The different mapped events of the same CpG/CWG are summarized into one line of record. The content of the two files appears as shown below:

chr1	3006186	3006188	0	0	0	1
.....						
chr1	78034550	78034552	0	2	0	1
.....						

The columns represent:
First: chromosome name;
Second: start genomic coordinate of CpG/CWG;
Third: end genomic coordinate of CpG/CWG;
Fourth: number of mapped events of this CpG/CWG in state of unmethylation;
Fifth: number of mapped events of this CpG/CWG in state of hemi-methylation (C on Crick methylated);
Sixth: number of mapped events of this CpG/CWG in state of hemi-methylation (C on Watson methylated);

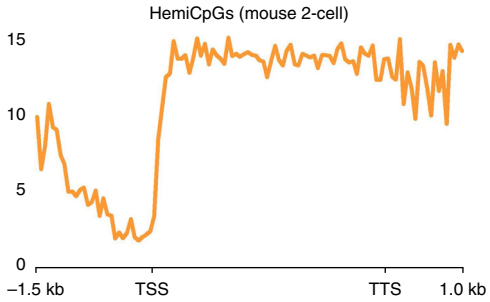

Fig. 5 | Frequency of hemiCpGs in mouse two-cell embryos. The DNA methylome is resolved into three components by iSA: unmethylated CpGs, fully methylated CpGs, and hemiCpGs. A metagene profile shows the frequency of hemiCpGs out of all three components of the DNA methylome around genic regions in mouse two-cell-stage embryos. TSS, transcription start site; TTS, transcription termination site.

Seventh: number of mapped events of this CpG/CWG in state of methylation.
At Step 13, the shell script prints the result as shown below:

You found 1780160 intraCpGs, of which:

880760 are unmethylated,

128622 are hemi-Watson,

126333 are hemi-Crick,

644445 are methylated.

You found 9055046 intraCWGs, of which:

8826253 are unmethylated,

112881 are hemi-Watson,

113932 are hemi-Crick,

1980 are methylated.

Profile of CpGs in different methylation states at genomic features

At Step 15, the results show that hemiCpGs are relatively depleted at promoter regions (Fig. 5), suggesting a role of hemiCpGs in the inhibition of promoter-based activities (e.g., transcription initiation).

Reporting Summary

Further information on research design is available in the Nature Research Reporting Summary.

Code availability

An interactive shell script is available at https://github.com/chxu02/iSA.

Data availability

WGBS dataset (GSM1386021) is available at SRA.

References

1. Xu, C. & Corces, V. G. Nascent DNA methylome mapping reveals inheritance of hemimethylation at CTCF/cohesin sites. *Science* **359**, 1166–1170 (2018).
2. Law, J. A. & Jacobsen, S. E. Establishing, maintaining and modifying DNA methylation patterns in plants and animals. *Nat. Rev. Genet.* **11**, 204–220 (2010).

3. Statham, A. L. et al. Bisulfite sequencing of chromatin immunoprecipitated DNA (BisChIP-seq) directly informs methylation status of histone-modified DNA. *Genome Res.* **22**, 1120–1127 (2012).
4. Brinkman, A. B. et al. Sequential ChIP-bisulfite sequencing enables direct genome-scale investigation of chromatin and DNA methylation cross-talk. *Genome Res.* **22**, 1128–1138 (2012).
5. Meissner, A. et al. Reduced representation bisulfite sequencing for comparative high-resolution DNA methylation analysis. *Nucleic Acids Res.* **33**, 5868–5877 (2005).
6. Kivioja, T. et al. Counting absolute numbers of molecules using unique molecular identifiers. *Nat. Methods* **9**, 72–74 (2011).
7. Yu, M. et al. Base-resolution analysis of 5-hydroxymethylcytosine in the mammalian genome. *Cell* **149**, 1368–1380 (2012).
8. Booth, M. J. et al. Quantitative sequencing of 5-methylcytosine and 5-hydroxymethylcytosine at single-base resolution. *Science* **336**, 934–937 (2012).
9. Song, C. X. et al. Genome-wide profiling of 5-formylcytosine reveals its roles in epigenetic priming. *Cell* **153**, 678–691 (2013).
10. Laird, C. D. et al. Hairpin-bisulfite PCR: assessing epigenetic methylation patterns on complementary strands of individual DNA molecules. *Proc. Natl. Acad. Sci. USA* **101**, 204–209 (2004).
11. Arand, J. et al. In vivo control of CpG and non-CpG DNA methylation by DNA methyltransferases. *PLoS Genet.* **8**, e1002750 (2012).
12. Zhao, L. et al. The dynamics of DNA methylation fidelity during mouse embryonic stem cell self-renewal and differentiation. *Genome Res.* **24**, 1296–1307 (2014).
13. Krueger, F. & Andrews, S. R. Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications. *Bioinformatics* **27**, 1571–1572 (2011).
14. Li, H. et al. The sequence alignment/map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
15. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).
16. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120 (2014).
17. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012).
18. Ramirez, F., Dundar, F., Diehl, S., Gruning, B. A. & Manke, T. deepTools: a flexible platform for exploring deep-sequencing data. *Nucleic Acids Res.* **42**, W187–W191 (2014).
19. Holmes, E. E. et al. Performance evaluation of kits for bisulfite-conversion of DNA from tissues, cell lines, FFPE tissues, aspirates, lavages, effusions, plasma, serum, and urine. *PLoS ONE* **9**, e93933 (2014).
20. Miura, F. et al. Amplification-free whole-genome bisulfite sequencing by post-bisulfite adaptor tagging. *Nucleic Acids Res.* **40**, e136 (2012).
21. Wang, L. et al. Programming and inheritance of parental DNA methylomes in mammals. *Cell* **157**, 979–991 (2014).

Acknowledgements

This work was supported by U.S. Public Health Service Award 5P01 GM085354. The content of this article is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Author contributions

C.X. and V.G.C. conceived the project; C.X. designed and streamlined the protocol; C.X. and V.G.C. wrote the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41596-018-0090-x.

Reprints and permissions information is available at www.nature.com/reprints.

Correspondence and requests for materials should be addressed to V.G.C.

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

Published online: 12 December 2018

Related link
Key reference using this protocol
Xu, C. & Corces, V. G. *Science* **359**, 1166–1170 (2018): https://doi.org/10.1126/science.aan5480