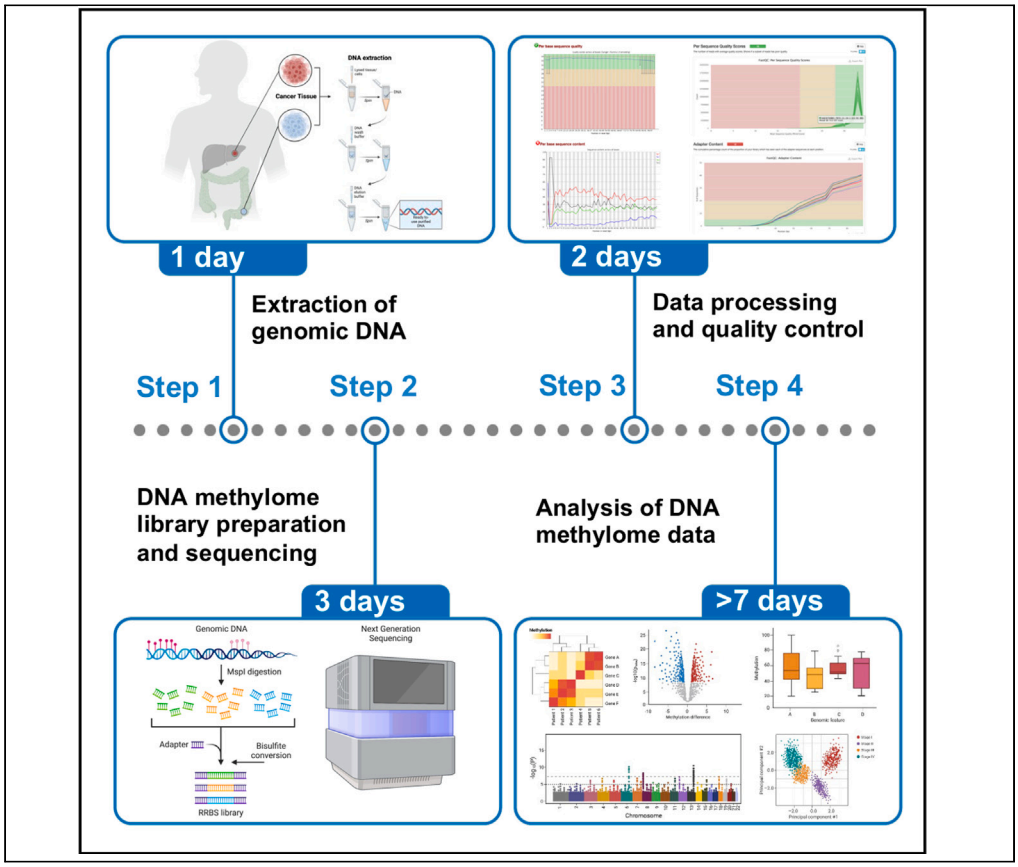
STAR Protocols

CellPress
OPEN ACCESS

Protocol

Protocol for generating high-quality genome-scale DNA methylation sequencing data from human cancer biospecimens



Aberrant DNA methylation is a universal feature of cancer. Here, we present a protocol for generating high-quality genome-scale DNA methylation sequencing data from a variety of human cancer biospecimens including immortalized cell lines, fresh-frozen surgical resections, and formalin-fixed paraffin-embedded tissues. We describe steps for DNA extraction considerations, reduced representation bisulfite sequencing, data processing and quality control, and downstream data analysis and integration. This protocol is also applicable for other human diseases and methylome profiling in other organisms.

Publisher's note: Undertaking any experimental protocol requires adherence to local institutional guidelines for laboratory safety and ethics.

Euan J. Rodger, Peter A. Stockwell, Suzan Almomani, Michael R. Eccles, Aniruddha Chatterjee

euan.rodger@otago.ac.nz (E.J.R.)
aniruddha.chatterjee@otago.ac.nz (A.C.)

Highlights

Processing of different biospecimens for optimal cancer DNA methylome analysis

Generation of sequencing-based genome-scale DNA methylation data from human cancers

Key considerations for optimal DNA methylation data output, quality, and interpretation

Computational workflow gives step-by-step guideline for data processing and analysis

Rodger et al., STAR Protocols 4, 102714
December 15, 2023
https://doi.org/10.1016/j.xpro.2023.102714

STAR Protocols

CellPress
OPEN ACCESS

Protocol

Protocol for generating high-quality genome-scale DNA methylation sequencing data from human cancer biospecimens

Euan J. Rodger,[1,3,*] Peter A. Stockwell,[1] Suzan Almomani,[1] Michael R. Eccles,[1] and Aniruddha Chatterjee[1,2,4,*]

[1]Department of Pathology, Dunedin School of Medicine, University of Otago, Dunedin, New Zealand

[2]Honorary Professor, UPES University, Dehradun, India

[3]Technical contact: euan.rodger@otago.ac.nz

[4]Lead contact

*Correspondence: euan.rodger@otago.ac.nz (E.J.R.), aniruddha.chatterjee@otago.ac.nz (A.C.)
https://doi.org/10.1016/j.xpro.2023.102714

SUMMARY

Aberrant DNA methylation is a universal feature of cancer. Here, we present a protocol for generating high-quality genome-scale DNA methylation sequencing data from a variety of human cancer biospecimens including immortalized cell lines, fresh-frozen surgical resections, and formalin-fixed paraffin-embedded tissues. We describe steps for DNA extraction considerations, reduced representation bisulfite sequencing, data processing and quality control, and downstream data analysis and integration. This protocol is also applicable for other human diseases and methylome profiling in other organisms.

For complete details on the use and execution of this protocol, please refer to Rodger et al. (2023).[1]

BEFORE YOU BEGIN

We have used the protocol described here for several human cancers[1–5] and have also successfully used it for DNA methylation analysis of other human diseases,[6–8] normal human cells,[9,10] mice,[11] zebrafish,[12–15] and mangrove rivulus.[16]

Before starting sample processing, reduce the risk of contamination by ensuring adherence to good laboratory practice. This includes: 1. Organizing a workspace with a flow of work that occurs in one direction (from 'clean' pre-PCR areas to 'dirty' post-PCR areas); 2. Aliquoting reagents to avoid multiple freeze-thaws and contamination of master stocks; 3. Clearly labeling and adding dates to all reagent and reaction tubes; 4. Routinely decontaminating surfaces and equipment such as pipettes; 5. Pipetting all reagents and samples using filter tips. Samples and reagents should be securely stored at the appropriate temperature (e.g., fresh-frozen tissue samples stored at −80°C, reagents such as enzymes stored according to the supplier's recommendations).

We sequence our bisulfite sequencing libraries on the Illumina platform, which is commonplace in many core sequencing facilities across the world. Before generating libraries, we recommend consulting with a sequencing facility to set up a contract and discuss library submission guidelines. This includes checking the following: 1. Requirements for multiplexing, library or library pool quality, concentration, and buffer volume; 2. Compatibility of adapter sequences with the current Illumina chemistry and specific sequencing machine; 3. Sequencing and output specifications (e.g., read length and configuration, either single- or paired-end, number of reads per library, genomic coverage, and data delivery).

Quality control and analysis of the sequencing data requires sufficient bioinformatics expertise and adequate computational processing capacity and storage. We recommend a Unix/Linux system with at least 16 GB of RAM and 5 cores. For data storage, we typically obtain raw FASTQ sequencing files that are ~2 GB per library (gzip compressed format); consideration also needs to be given for the storage of multiple intermediate and processed files. If there are limitations in storage capacity due to large cohort sizes, then we recommend only storing compressed FASTQ files for long-term storage.

Institutional permissions

Ethical approval, written informed consent, and de-identification of personal information are mandatory for working with patient-derived biospecimens in accordance with institutional and national guidelines and regulations. We obtained ethical approval for the use of tissue from cancer patients by the Human Ethics Committee, University of Otago. We also follow all institutional health and safety guidelines for working with biohazardous materials.

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Chemicals, peptides, and recombinant proteins		
Nuclease-free water	Thermo Fisher Scientific	Cat#: AM9937
Deparaffinization solution for FFPE tissue	QIAGEN	Cat#: 19013
Ethanol, analytical grade	Merck	Cat#: 100983
PBS, pH 7.4	Thermo Fisher Scientific	Cat#: 10010023
MspI restriction endonuclease (20 U/μL)	NEB	Cat#: R0106S
NuSieve GTG agarose	Lonza	Cat#: 50081
KAPA HiFi HotStart Uracil+ ReadyMix	Roche	Cat#: KK2801
Quick-Load Purple low molecular weight DNA ladder	NEB	Cat#: N0557S
DNA gel loading dye	Thermo Fisher Scientific	Cat#: R0611
Ethidium bromide, 10 mg/mL	Thermo Fisher Scientific	Cat#: 15585011
TAE buffer (Tris-acetate-EDTA)	Thermo Fisher Scientific	Cat#: B49
Critical commercial assays		
QIAamp DNA Mini or DNA FFPE Tissue Kit	QIAGEN	Cat#: 51304 Cat#: 56404
TruSeq Nano DNA LT Sample Prep Kit	Illumina	Cat#: TG-202-1001
EZ DNA Methylation-Direct Kit	Zymo Research	Cat#: D5021
Deposited data		
Raw and analyzed RRBS data of primary colorectal adenocarcinomas and liver metastases from the same patients	Rodger et al.[1]	GEO: GSE213402
Human reference genome NCBI build 37, GRCh37	Genome Reference Consortium	http://www.ncbi.nlm.nih.gov/projects/genome/assembly/grc/human/
Software and algorithms		
FastQC	Simon Andrews	https://www.bioinformatics.babraham.ac.uk/projects/fastqc/
Bowtie 2	Langmead and Salzberg[17]	https://bowtie-bio.sourceforge.net/bowtie2/index.shtml
SAMtools	Li et al.[18]	http://www.htslib.org/
Cutadapt	Marcel Martin	https://cutadapt.readthedocs.io/en/stable/
Trim Galore	Felix Krueger	https://www.bioinformatics.babraham.ac.uk/projects/trim_galore/
Bismark	Krueger et al.[19]	https://www.bioinformatics.babraham.ac.uk/projects/bismark/
DMAP	Stockwell et al.[20]	https://github.com/peterstockwell/DMAP
R packages	RStudio	https://posit.co

(*Continued on next page*)

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
BEDTools	Quinlan et al.[21]	https://bedtools.readthedocs.io/en/latest/
Other		
Rotary microtome	Leica Biosystems	https://www2.leicabiosystems.com
DNA LoBind 1.5 mL microcentrifuge tubes or equivalent	Eppendorf	Cat#: 022431021
Hemocytometer	Hausser Scientific	Cat#: 3100
Polypropylene pellet pestles or equivalent	DWK Life Sciences	Cat#: 749520-0000
QIAquick and MinElute Purification Kits or equivalent	QIAGEN	Cat#: 28104 Cat#: 28004
5425 Microcentrifuge or equivalent	Eppendorf	Cat#: 5405000646
MAXYMum Recovery 0.2 mL PCR tubes or equivalent	Axygen	Cat#: PCR-02-L-C
Qubit fluorometer and dsDNA BR Assay Kit or equivalent	Thermo Fisher Scientific	https://www.thermofisher.com/
N60 NanoPhotometer or equivalent	Implen	https://www.implen.de/nanophotometer/
2100 Bioanalyzer instrument and High Sensitivity DNA Kit or equivalent	Agilent	https://www.agilent.com/
T100 96-well thermal cycler or equivalent	Bio-Rad	Cat#: 1861096
Thermomixer with 1.5 mL SmartBlock or equivalent	Eppendorf	Cat#: 2231001005
EpiMag 96-well magnetic separator or equivalent	Epigentek	Cat#: Q10002-1
Mupid-exU gel electrophoresis system or equivalent	Takara	Cat#: AD140
Uvidoc HD6 UV gel imaging system or equivalent	Uvitec	https://www.uvitec.co.uk/

STEP-BY-STEP METHOD DETAILS

Extraction of genomic DNA

Timing: 1 day

The generation of reliable DNA methylome data is highly dependent on the quality and amount of the input DNA. For fresh-frozen cancer tissue and cell lines, this is relatively straightforward if there is a sufficient amount of starting material and best practice is used to ensure minimal degradation of DNA. Extraction of high-quality DNA from FFPE cancer tissue is more complex – the fixation process can cause crosslinking of nucleic acids, random breakages in nucleotide sequences, and introduce sequence artifacts. Obtaining reliable methylome data from FFPE tissue is achievable, as we have shown previously,[22] but results are highly dependent on how the tissue was fixed, how it was stored, and the age of the block. The DNA extraction protocol here is adapted from the QIAGEN QIAamp guidelines, with amendments to improve yield and quality. We provide specific recommendations for optimization of FFPE tissue, fresh-frozen tissue, and cell lines.

1. For FFPE cancer tissue:
   a. Using a microtome, cut 5–10 μm thick sections from the sample block.
      i. Discard the first 2–3 sections (due to air exposure).
      ii. Place the next 4–8 sections into a 1.5 mL DNA LoBind microcentrifuge tube.
   b. Add 320 μL deparaffinization solution.
      i. Mix thoroughly by vortexing.
      ii. Briefly centrifuge so content is at the bottom of tube.
      iii. Incubate at 56°C for 3 min.
      iv. Allow to cool at room temperature.
   c. Add 180 μL QIAamp ATL buffer.
      i. Mix by vortexing.
      ii. Centrifuge for 1 min at 11,000 × *g*.
   d. Add 20 μL proteinase K to the lower phase (clear liquid).

i. Mix gently by pipetting up and down.
ii. Incubate cell preparation on Thermomixer overnight (16–20 h) at 56°C, with moderate shaking (300–400 rpm).
iii. If the cells have not fully digested, add 20 μL proteinase K and incubate for a further 1–2 h.
iv. Remove tube from the Thermomixer.
v. Increase the temperature to 90°C, and then incubate at 90°C for 1 h.
vi. Briefly centrifuge to remove drops from inside the lid.
vii. Carefully pipette the lower phase into a new 1.5 mL DNA LoBind tube.

e. Add 4 μL RNase A (100 mg/mL).
i. Mix by vortexing.
ii. Incubate for 2 min at room temperature.

f. Add 200 μL QIAamp AL buffer to the sample.
i. Mix again by vortexing.
ii. Incubate at 70°C for 10 min.
iii. Briefly centrifuge to remove drops from inside the lid and proceed to step 4.

2. For fresh-frozen cancer tissue:
a. Retrieve the frozen tissue and equilibrate to room temperature for a short time.
i. Cut the tissue (up to 20 mg) into small pieces.
ii. Place in a 1.5 mL DNA LoBind microcentrifuge tube with 180 μL of QIAamp ATL buffer.

b. Use a pellet pestle to grind the tissue pieces.
i. Add 20 μL proteinase K.
ii. Mix by vortexing.
iii. Briefly centrifuge so content is at the bottom of tube.
iv. Incubate cell preparation on Thermomixer overnight at 56°C, with moderate shaking (300–400 rpm).
v. If the cells have not fully digested, add 20 μL proteinase K and incubate for a further 1–2 h.
vi. Briefly centrifuge to remove drops from inside the lid.

c. Add 4 μL RNase A (100 mg/mL).
i. Mix by vortexing.
ii. Incubate for 2 min at room temperature.
iii. Add 200 μL QIAamp AL buffer to the sample.
iv. Mix again by vortexing.
v. Incubate at 70°C for 10 min.
vi. Briefly centrifuge to remove drops from inside the lid and proceed to step 4.

3. For cancer cell lines:
a. To harvest adherent cultured cells, discard media and wash with 5 mL PBS.
i. Add ~3 mL trypsin to detach the cells from the surface of the flask.
ii. Incubate at 37°C for 2–3 min.
iii. Gently shake or tap the bottom of the flask to help detach cells and incubate longer if necessary.
iv. When the cells have detached, add 10 mL of media.
v. Transfer cell suspension to a 15 mL Falcon conical centrifuge tube.
vi. Count the number of cells using a hemocytometer.
vii. To pellet the cells, centrifuge the cells at 200 × *g* for 5 min.
viii. Discard the supernatant.
ix. The cell pellet can be frozen at −80°C for future use or further processed immediately.

b. In a 1.5 mL DNA LoBind tube, resuspend cell pellet (up to 5×10^6 cells) in PBS to a total volume of 200 μL.
i. Add 20 μL QIAGEN protease solution and 200 μL QIAamp AL buffer to cell suspension.
ii. Mix by vortexing.
iii. Briefly centrifuge so content is at the bottom of tube.
iv. Incubate cell preparation on Thermomixer overnight at 56°C, with moderate shaking (300–400 rpm).

v. If the cells have not fully digested, add 20 μL QIAGEN protease solution and incubate for a further 1–2 h.
vi. Briefly centrifuge to remove drops from inside the lid and proceed to step 4.

4. Following on from either step 1 (for FFPE tissue), step 2 (for fresh-frozen tissue) or step 3 (for cell lines):
   a. Add 200 μL ethanol (100%).
      i. Vortex to mix.
      ii. Carefully apply the mixture to a QIAamp mini spin column (in a 2 mL collection tube).
      iii. Centrifuge at 20,000 × *g* for 1 min.
      iv. Put the spin column in a clean 2 mL collection tube and discard the filtrate.
   b. Add 500 μL QIAamp AW1 buffer to the spin column.
      i. Centrifuge at 20,000 × *g* for 1 min.
      ii. Put the spin column in a clean 2 mL collection tube and discard the filtrate.
   c. Repeat with 500 μL QIAamp AW2 buffer.
      i. Centrifuge at 20,000 × *g* for 1 min.
      ii. Put the spin column in a clean 2 mL collection tube.
      iii. Centrifuge again at 20,000 × *g* for 3 min.
   d. Put the spin column in a 1.5 mL DNA LoBind tube.
      i. Immediately add 50 μL QIAamp ATE buffer (pre-warmed to 30°C) directly onto the spin column membrane.
      ii. Incubate at room temperature for 5 min.
      iii. Centrifuge at 6000 × *g* for 1 min.
      iv. Immediately re-apply the 50 μL filtrate onto the spin column membrane.
      v. Centrifuge again at 6000 × *g* for 1 min.
      vi. Retain the eluate containing purified DNA for the next step.
   e. Check the DNA quality (A260/A280 ratio 1.7–1.9) on a NanoPhotometer and determine the concentration using a Qubit Fluorometer and dsDNA BR Assay Kit. Troubleshooting.

Note: DNA can be stored short-term at 4°C and long-term at −20°C.

⚠ CRITICAL: Particularly for FFPE sections, which should be freshly cut, it is important to process the samples as soon as possible. Exposure of sections to air negatively impacts DNA quality, which gets worse with time.

⚠ CRITICAL: QIAGEN guidelines indicate a protease/proteinase K digestion time of 1 h, but a much longer time is required for DNA used in methylome sequencing libraries to enable complete dissociation of genomic DNA from contaminating nucleosomal or DNA binding proteins.

⚠ CRITICAL: Using the QIAamp mini spin columns requires careful pipetting and liquid handling. Avoid wetting the rim of the spin column or touching the membrane with the pipette tip. After the wash steps with QIAamp AW1 and AW2 buffers, the column is centrifuged for a further 3 min to completely dry the membrane and remove any ethanol carry-over. However, the membrane should not be left dry for too long before the elution buffer is added.

DNA methylome library preparation

Timing: 3 days

Here we present a detailed protocol for the generation of libraries for reduced representation bisulfite sequencing (RRBS), a technique for comprehensive profiling of genome-scale DNA methylation patterns.[23–25] RRBS library preparation involves digesting genomic DNA with MspI restriction

enzyme, which cuts at 5′-C↓CGG-3′ irrespective of the methylation status, to enrich for CpG-dense regions of the genome. Adapter-ligated fragments are bisulfite-converted and amplified by PCR, which allows for sequence-based distinction between methylated and unmethylated CpGs, and then 40–220 bp fragments are size-selected. Although the 4 million CpG sites that RRBS covers is less than for whole genome bisulfite sequencing, it is significantly more than the <850 k CpG sites covered by Infinium MethylationEPIC arrays. For this reason, RRBS provides ample data for robust genome-scale methylation profiling while also remaining cost-effective.

5. Set up MspI restriction endonuclease digest in MAXYMum Recovery 0.2 mL PCR tube.
   a. Incubate overnight (16–20 h) at 37°C on a 96-well thermal cycler.

MspI restriction digest

Reagent	Final amount	Volume
Genomic DNA (from step 4)	100–500 ng	x μL (up to 28 μL)
MspI enzyme (20 U/μL)	160 U	8 μL
10× CutSmart buffer	1×	4 μL
Nuclease-free water	N/A	28 - x
Total	**N/A**	**40 μL**

Note: if the volume of genomic DNA exceeds 28 μL, the total volume can be increased and the other components increased proportionately.

6. QIAquick spin column purification:
   a. Add 5 volumes of QIAquick PB buffer to 1 volume of the restriction digest from step 5.
      i. Mix by inversion or by pipetting (e.g., 200 μL of PB buffer to 40 μL restriction digest) in a 1.5 mL DNA LoBind tube.
   b. Apply mixture to a QIAquick spin column.
      i. Centrifuge at 20,000 × *g* for 1 min and discard the filtrate.
   c. To wash, add 700 μL QIAquick PE buffer to the spin column.
      i. Centrifuge at 20,000 × *g* for 1 min.
   d. Discard the filtrate, place the spin column into a new 2 mL collection tube.
      i. Centrifuge the column for an additional 1 min.
   e. Place spin column in a clean well-labeled 1.5 mL DNA LoBind tube.
      i. To elute DNA, add 60 μL QIAquick EB buffer (pre-warmed to 30°C) directly to the center of the spin column membrane.
      ii. Incubate at room temperature for 10 min.
      iii. Centrifuge at 20,000 × *g* for 1 min.
   f. Re-apply the 60 μL eluate directly on the spin column membrane.
      i. Incubate at room temperature for 1 min.
      ii. Centrifuge at 20,000 × *g* for 1 min.
      iii. Retain the eluate for the next step.

Pause point: DNA can be stored short-term at 4°C and long-term at −20°C.

7. Set up End Repair reaction in a 1.5 mL LoBind tube.
   a. Add 40 μL thawed TruSeq Nano ERP2 to the 60 μL of purified MspI-digested DNA from step 6.
      i. Incubate for 30 min at 30°C on a Thermomixer (hold at 4°C).
8. MinElute spin column purification:
   a. Add 500 μL of MinElute PB buffer to the 100 μL end repair reaction.
      i. Mix by inversion.
      ii. Apply the mixture to a MinElute spin column.
      iii. Centrifuge at 20,000 × *g* for 1 min, and discard the filtrate.

b. To wash, add 700 µL MinElute PE buffer to the spin column.
   i. Centrifuge at 20,000 × *g* for 1 min.
c. Discard filtrate and place the spin column into a new 2 mL collection tube.
   i. Centrifuge the column for an additional 1 min at 20,000 × *g*.
d. Place spin column in a clean well-labeled 1.5 mL LoBind tube.
   i. To elute DNA, add 18 µL MinElute EB buffer (pre-warmed to 30°C) directly to the center of the spin column membrane.
   ii. Incubate at room temperature for 10 min.
   iii. Centrifuge the column at 20,000 × *g* for 1 min.
e. Re-apply the 18 µL eluate directly on the spin column membrane.
   i. Incubate at room temperature for 1 min.
   ii. Centrifuge at 20,000 × *g* for 1 min.
   iii. Retain the eluate for the next step.

Pause point: DNA can be stored short-term at 4°C and long-term at −20°C.

9. Set up A-Tailing reaction in a MAXYMum Recovery 0.2 mL PCR tube.
   a. Add 12 µL thawed TruSeq Nano ATL to the 18 µL of purified DNA from step 8.
      i. Incubate on a 96-well thermal cycler:

A-Tailing conditions	
Temperature	Time
37°C	30 min
70°C	5 min
4°C	5 min
4°C	Hold

10. Set up Adapter Ligation reaction in a MAXYMum Recovery 0.2 mL PCR tube.
    a. Incubate for 11 min at 30°C on a 96-well thermal cycler.
    b. Immediately following the reaction, add 5 µL TruSeq Nano STL.

Adapter ligation reaction	
Reagent	Volume
End-repaired DNA (from step 9)	30 µL
TruSeq Nano RSB buffer	2.5 µL
TruSeq Nano Adapter index	2.5 µL
TruSeq LIG2	2.5 µL
Total	**37.5 µL**

Note: for low-input RRBS libraries (<100 ng), use adapter index and ligation mix which has been diluted 1 in 5 to prevent excessive adapter dimers in subsequent steps.

Note: there are 24 different TruSeq Nano Adapter indexes to allow for multiplexing of samples. We recommend using the Illumina Experiment Manager software to determine the best multiplexing adapter combinations.

11. MinElute spin column purification:
    a. Add 215 µL of MinElute PB buffer to the 37.5 µL adapter ligation reaction from step 10.
       i. Mix by inversion.
       ii. Apply the mixture to a MinElute spin column.
       iii. Centrifuge at 20,000 × *g* for 1 min and discard the filtrate.

b. To wash, add 700 µL MinElute PE buffer to the spin column.
   i. Centrifuge at 20,000 × *g* for 1 min.
   ii. Discard filtrate and place the spin column into a new 2 mL collection tube.
   iii. Centrifuge the column for an additional 1 min and place spin column in a clean well-labeled 1.5 mL LoBind tube.
c. To elute DNA, add 20 µL MinElute EB buffer (pre-warmed to 30°C) directly to the center of the spin column membrane.
   i. Incubate at room temperature for 10 min.
   ii. Centrifuge the column at 20,000 × *g* for 1 min.
   iii. Re-apply the 20 µL eluate directly on the spin column membrane.
   iv. Incubate at room temperature for 1 min.
   v. Centrifuge at 20,000 × *g* for 1 min.
   vi. Retain the eluate for the next step.

Pause point: DNA can be stored short-term at 4°C and long-term at –20°C.

12. Bisulfite conversion:
    a. Prepare EZ DNA Methylation-Direct CT conversion reagent.
       i. Add 790 µL of M-Solubilization buffer and 300 µL of M-Dilution Buffer to a tube of CT Conversion Reagent.
       ii. Mix thoroughly by vortexing for 1 min.
       iii. Incubate CT reagent mixture on Thermomixer for 1 h at room temperature with high shaking (>2000 rpm). By visual inspection, the powder should be completely dissolved.
       iv. Add 160 µL of M-Reaction Buffer.
       v. Mix for an additional 1 min.
    b. Set up Bisulfite conversion reaction in a MAXYMum Recovery 0.2 mL PCR tube.
       i. Add 130 µL of the CT reagent to the 20 µL of adapter-ligated DNA from step 11.
       ii. Incubate on a 96-well thermal cycler:

Bisulfite conversion	
Temperature	Time
98°C	8 min
64°C	3.5 h
4°C	Hold

13. EZ DNA spin column purification:
    a. Add 600 µL of EZ DNA M-Binding Buffer to the 150 µL bisulfite conversion reaction from 12 to a Zymo-Spin IC Column in a 2 mL collection tube.
       i. Mix by inversion.
       ii. Centrifuge at 20,000 × *g* for 1 min and discard the filtrate.
    b. To wash, add 100 µL of EZ DNA M-Wash Buffer to the spin column.
       i. Centrifuge at 20,000 × *g* for 1 min.
    c. Add 200 µL of EZ DNA M-Desulfonation Buffer to the spin column.
       i. Incubate at room temperature for 15–20 min.
       ii. Centrifuge at 20,000 × *g* for 1 min.
    d. To wash, add 200 µL of EZ DNA M-Wash Buffer to the spin column.
       i. Centrifuge at 20,000 × *g* for 1 min.
       ii. Repeat by adding a further 200 µL of EZ DNA M-Wash Buffer to the spin column.
       iii. Centrifuge at 20,000 × *g* for 1 min.
       iv. Discard the collection tube and place spin column in a clean well-labeled 1.5 mL LoBind tube.

e. To elute DNA, add 15 μL EZ DNA M-Elution Buffer (pre-warmed to 30°C) directly to the center of the spin column membrane.
   i. Incubate at room temperature for 10 min.
   ii. Centrifuge the column at 20,000 × *g* for 1 min.
   iii. Re-apply the 15 μL eluate directly on the spin column membrane.
   iv. Incubate at room temperature for 1 min.
   v. Centrifuge at 20,000 × *g* for 1 min.
   vi. Retain the eluate for the next step.

Pause point: bisulfite-converted DNA can be stored for 2–3 days at 4°C and <1 month at −20°C.

14. Set up a semi-quantitative PCR reaction in a MAXYMum Recovery 0.2 mL PCR tube.
   a. Mix by flicking the tube.
   b. Briefly centrifuge to bring content to bottom of tube.

Semi-quantitative PCR reaction

Reagent	Volume
KAPA HiFi HotStart Uracil+ ReadyMix	12.5 μL
TruSeq PCR Primer Cocktail	3 μL
Bisulfite-converted DNA (from step 13)	3 μL
Nuclease-free Water	6.5 μL
Total	**25 μL**

   c. Aliquot 12 μL of the reaction mix into two 0.2 mL PCR tubes and run separately on a 96-well thermal cycler (one for 15 and the other for 20 cycles of PCR):

Semi-quantitative PCR cycling conditions

Steps	Temperature	Time	Cycles
Initial Denaturation	98°C	45 s	1
Denaturation	98°C	15 s	Either 15 or 20 cycles
Annealing	60°C	30 s	
Extension	72°C	30 s	
Final extension	72°C	60 s	1
Hold	4°C	Hold	

   d. Add 2 μL DNA Gel Loading Dye to each PCR reaction.
   e. Load PCR reactions on a 3% (w/v) NuSieve agarose gel made up in 0.5x TAE buffer, alongside 7 μL of Quick-Load Purple Low Molecular Weight DNA Ladder.
   f. Submerged in 0.5x TAE buffer (0.0024% ethidium bromide), run the gel in an electrophoresis system for ~90 min at 50 V.
   g. Image the gel using a UV Gel Imaging System, ensuring that the DNA ladder bands are clearly visible. Assess the gel and determine the optimal number of cycles (between 15 and 20) for large-scale PCR amplification of the library.
15. Set up large-scale PCR amplification in a MAXYMum Recovery 0.2 mL PCR tube.
   a. Mix by flicking the tube.
   b. Briefly centrifuge to bring content to bottom of tube.
   c. Run on a 96-well thermal cycler using the same cycling conditions in step 14.

Large-scale PCR amplification

Reagent	Volume
KAPA HiFi HotStart Uracil+ ReadyMix	25 μL
TruSeq PCR Primer Cocktail	6 μL

(*Continued on next page*)

Continued	
Reagent	Volume
Bisulfite-converted DNA (from step 13)	6 μL
Nuclease-free Water	13 μL
Total	50 μL

Note: following the semi-quantitative and large-scale amplification PCR steps above, ~11 μL of bisulfite-converted DNA from step 13 will remain. If any issues arise with remaining steps, this remaining DNA can be used for large-scale amplification.

16. Bead purification and size-selection:
    a. Bring an aliquot of TruSeq Sample Purification beads to room temperature from 4°C storage and vortex to make a homogenous solution.
    b. Add 30 μL of the beads (0.6× volume ratio) to PCR product from step 15.
        i. Pipette up and down to thoroughly mix.
        ii. Incubate at room temperature for 15 min.
        iii. Place the PCR tube on a magnetic rack for 5 min until the beads form a small dark pellet on the side of the tube and the solution is clear.
        iv. Carefully pipette the clear supernatant, which contains the library, into a new 0.2 mL PCR tube without disturbing the beads. Discard the original PCR tube with the beads.
    c. To the new tube with the library, add 20 μL fresh TruSeq Sample Purification beads (0.4× of original volume).
        i. Pipette up and down to thoroughly mix.
        ii. Incubate at room temperature for 15 min.
        iii. Place the PCR tube on a magnetic rack for 5 min until the bead pellet is visible on the side of the tube and the solution is clear.
        iv. Carefully remove the clear supernatant by pipette, which contains the small molecular weight contaminants, and discard.
    d. To wash the beads, add 200 μL of freshly prepared 70% ethanol (v/v).
        i. Carefully remove the ethanol by pipette and discard.
        ii. Repeat the wash once more and ensure all ethanol is removed.
    e. Remove the tube from the magnetic rack and allow the bead pellet to dry by leaving the tube open at room temperature for ~1 min.

Note: This requires watching closely and the pellet will go from shiny to matte. Try not to let bead pellet crack as this can result in a lower yield of purified DNA.

    f. Add 20 μL of TruSeq Resuspension Buffer (pre-warmed to 30°C) to the pellet.
        i. Mix well with the pipette to resuspend the beads.
        ii. Incubate at room temperature for a minimum of 15 min.
        iii. Return the PCR tube to the magnetic rack for 5 min until the bead pellet is visible on the side of the tube and the solution is clear.
        iv. Carefully pipette the clear supernatant, which contains the purified library, into a 1.5 mL LoBind tube without disturbing the beads (~2 μL of supernatant will be left behind regardless of the elution volume).

Pause point: bisulfite-converted DNA methylome libraries can be stored for 2–3 days at 4°C and long-term at –20°C.

17. Library quantification and quality assessment:
    a. Using a Qubit Fluorometer and dsDNA BR Assay kit according to the manufacturer's instructions, quantify the DNA library.

b. Prepare a ~2 ng/μL dilution of the DNA library and run on a 2100 Bioanalyzer instrument using a high-sensitivity DNA kit.
c. Assess the Bioanalyzer electropherogram, which is a data plot of DNA fragment size (bp) versus fluorescence intensity.

Note: The profile should display a broad peak at ~260 bp, corresponding to 150–325 bp size-selected fragments. Troubleshooting

d. Use the following formula to determine library molarity:

$$nM = \text{Qubit concentration (ng / μL)} / (\text{average fragment length (bp)} \times 0.00065)$$

18. Sequencing submission:
    a. In consultation with a sequencing provider, prepare a 30 μL equimolar pool (10 nM) of multiple DNA libraries in a LoBind tube, ensuring the TruSeq Nano Adapter indexes are unique for each.

Note: We typically pool up to 12 libraries for NextSeq 2000 P2 sequencing with 100 bp single-end reads (~30 million reads/sample). The sequencing provider denatures the pooled library with NaOH and then dilutes it to a final concentration of 8 pM, which is used for cluster generation on the Illumina cBot system and then preliminary QC analysis is performed on a MiSeq platform before proceeding to full sequencing on a high throughput system such as HiSeq or NextSeq.

⚠ CRITICAL: The preparation of the EZ DNA Methylation-Direct CT conversion reagent provides enough for 10 reactions. Any remaining reagent can be stored for up to 1 week at 4°C or up to 4 weeks at –20°C. Following storage, the CT reagent needs to be mixed thoroughly – incubate on Thermomixer for 1 h at room temperature with high shaking.

⚠ CRITICAL: Assessing the gel image from semi-quantitative PCR is necessary to determine the optimal number of cycles for large-scale amplification in the next step. By comparing the products of the PCR with either 15 or 20 cycles: a) there should be a sufficient amount of 150–330 bp product without non-specific amplification; b) the number of cycles should be kept to a minimum to reduce amplification bias – we do not recommend any higher than 20 cycles. A 125 bp adaptor-dimer band is typically present, but this is removed following large-scale PCR amplification of the library using bead purification.

Data processing and quality control

Timing: 2 days

Next generation sequencing generates large amounts of data, which needs to be processed and quality-checked with specialist software before proceeding with downstream analyses. Here we present a simple workflow to: obtain the raw sequencing data; perform quality control analysis; remove adapter sequences and low-quality reads; map the sequencing reads to a reference genome; extract the methylation information for each MspI fragment/CpG and annotate them with overlapping genomic features.

19. Data acquisition:
    a. Sequencing providers often have multiple options to give access to the demultiplexed sequencing data, for example:
        i. Use the BaseSpace platform (https://basespace.illumina.com/) to access Illumina run metrics and data.

ii. Securely transfer large amounts of data using a file-sharing service such as Globus (https://www.globus.org/data-transfer).
iii. For rapid copying and synchronization of sequencing data between servers, use the following command in bash:

```
$ rsync -avP source_destination
```

Note: We recommend retaining a backup of the raw, unprocessed sequencing data in a secure location. FASTQ is the standard file format (.fq or .fastq extensions) for high throughput sequencing data and will often be gzip compressed (.gz extension). Although FASTQ is a text-based format that includes both read sequence and quality information, specialist programs are required to process and analyze the millions of sequencing reads (e.g., we typically use NextSeq 2000 P2 sequencing, which yields up to 400M single-end 100 bp reads).

20. FastQC is an efficient program that can be used to perform quality control analysis of high throughput sequencing data. Use the following command to generate a FastQC output:

```
$ fastqc --outdir qc *.fastq.gz
```

Note: The summarized output provides useful information on per base sequence quality, per base sequence content, overrepresented sequences, and adapter content. The FastQC analysis of raw sequencing reads will show the presence of adapter sequence and low-quality sequence reads (Phred score <20), which will need to be removed to obtain optimal genome alignment.

21. Adaptor sequence clean-up from datasets:
    a. Use the cleanadaptors function of DMAP to remove adapter sequences provided in the contam.fa file, trims 3 bases back from the adapter to delete the C incorporated during end repair (-t 3) and filters out any reads trimmed to <20 bp (-x 20). Assess the trimming performance by repeating the FastQC analysis on the trimmed FASTQ file.

```
$ cleanadaptors -I contam.fa -t 3 -x 20 \
-z -F methylome_01.fastq.gz -o methylome_01_trimmed.fastq.gz
```

Note: In this example, cleanadaptors removes adapter sequences provided in the contam.fa file, trims 3 bases back from the adapter to delete the C incorporated during end repair (-t 3) and filters out any reads trimmed to <20 bp (-x 20). Assess the trimming performance by repeating the FastQC analysis on the trimmed FASTQ file.

    b. Alternatively, use Trim Galore for quality and adapter trimming (which also produces a trimming report and FastQC of the trimmed FASTQ file):

```
$ trim_galore --rrbs –fastqc methylome_01.fastq.gz
```

22. Genome alignment:
    a. Following trimming and filtering, align the high-quality sequencing reads to the human reference genome (in FASTA format, available from http://ftp.ensembl.org/pub/) using Bismark, a program specifically designed for mapping bisulfite-treated sequencing reads. In the same directory as the reference genome files, build the Bismark libraries with a one-off command:

```
$ bismark_genome_preparation ./
```

b. Align the sequencing reads to the Bismark-prepared reference genome, allowing for only one mismatch in the seed (i.e., in the first 28 bp of each read):

```
$ bismark -n 1 -genome /path_to_reference_genome/ \ methylome_01_trimmed.fastq.gz
```

Note: The standard output file for Bismark is a binary alignment map (BAM), a machine-readable file that contains information on where the sequencing reads map in the human genome. An alignment report file is also produced, which gives an overview of mapping efficiency (number of unique alignments/sequences analyzed in total) and proportions of context-specific cytosine methylation: percentage of methylation (context) = 100 × methylated Cs (context) / (methylated Cs (context) + unmethylated Cs (context)). For libraries generated from fresh-frozen tissues and cultured cells, we routinely obtain 60%–70% alignment efficiency. For FFPE tissues, the alignment efficiency is sometimes decreased, but is dependent on how the tissue was fixed, how it was stored, and the age of the block – all of which affect the quality of the DNA extracted. Troubleshooting

23. Extraction of DNA methylation data:
    a. Using the genome alignment information in a BAM file, run the diffmeth program in DMAP to generate a readable coverage-filtered methylome text file using MspI-digested fragments as the unit of analysis:

```
$ diffmeth -F 2 -t 10 -z -G /path_to_reference_genome/ \
-L 40,220 -R methylome_01_trimmed_bismark_bt2.bam \
> methylome_01_fragments.txt
```

Note: In this example, the output will show 40–220 bp fragments (-L 40,220) with at least two CpG sites (-F 2) covered by 10 or more sequenced reads (-t 10). Full documentation and user guide for DMAP are available as a GitHub repository (https://github.com/peterstockwell/DMAP).

    b. Also using a BAM as an input, run Bismark methylation extractor to generate a genome-wide cytosine report and a text file in bedgraph format that contains the methylation call for each single C analyzed:

```
$ bismark_methylation_extractor --bedGraph --comprehensive \
--merge_non_CpG --cutoff 10 \ methylome_01_trimmed_bismark_bt2.bam
```

Note: The output of this example will show each CpG site covered by 10 or more sequenced reads (–cutoff 10) and strand-specific methylation combined into a single context-dependent file (–comprehensive).

    c. Annotate each MspI fragment from the DMAP output file or each CpG from the Bismark methylation extractor with genomic features using the identgeneloc program DMAP:

```
$ identgeneloc -i -Q -U -R -B "protein_coding" \

-G /path_to_feature_tables/ \

-r methylome_01_trimmed_bismark_bt2.txt \

> methylome_01_trimmed_bismark_bt2_geneloc.txt
```

Example of identgeneloc annotated output:

Chr	Start	End	Meth%	Enddist	Overlap	CpGIs_range	CpGIs_dist	CpGI_relation	Strand	GeneID	GeneRange
1	897154	897155	100	-1187	on_intron	894314-902654	2841	CpGI_core	5'	KLHL17	895967-901095
1	897164	897165	100	-1197	on_intron	894314-902654	2851	CpGI_core	5'	KLHL17	895967-901095
1	897192	897193	100	-1225	on_intron	894314-902654	2879	CpGI_core	5'	KLHL17	895967-901095
1	897200	897201	90.90909091	-1233	on_intron	894314-902654	2887	CpGI_core	5'	KLHL17	895967-901095
1	897219	897220	100	-1252	on_exon	894314-902654	2906	CpGI_core	5'	KLHL17	895967-901095
1	897220	897221	100	-1253	on_exon	894314-902654	2907	CpGI_core	5'	KLHL17	895967-901095
1	897243	897244	100	-1276	on_exon	894314-902654	2930	CpGI_core	5'	KLHL17	895967-901095
1	897244	897245	100	-1277	on_exon	894314-902654	2931	CpGI_core	5'	KLHL17	895967-901095
1	897247	897248	100	-1280	on_exon	894314-902654	2934	CpGI_core	5'	KLHL17	895967-901095
1	897248	897249	100	-1281	on_exon	894314-902654	2935	CpGI_core	5'	KLHL17	895967-901095

d. Use BEDTools for further bespoke annotation with genomic features:

```
$ bedtools intersect -wa -wb -a methylome_01_fragments.bed \

-b repeatmasker.bed > methylome_01_repeats.txt
```

Note: This example identifies where each MspI fragment overlaps a repeat element (obtained from the RepeatMasker track in UCSC genome browser: https://genome.ucsc.edu/).

e. With the unionbedg function, use BEDTools to combine multiple sorted methylome outputs from DMAP or Bismark methylation extractor (this approach can also be used to compare the same regions from other datasets):

```
$ bedtools unionbedg -i methylome_01_fragments.bed \ methylome_02_fragments.bed methylome_
03_fragments.bed \

-header -filler NA > methylomes_combined.txt
```

f. Import the annotated coverage-filtered methylomes into R Studio for further downstream analysis:

```
> methylomes_combined.txt <- read.delim("~/methylomes_combined.txt")

> View(methylomes_combined.txt)
```

Troubleshooting.

⚠ CRITICAL: Methylation in mammalian somatic tissues is exclusively in a CpG context; however, the Bismark alignment report gives a proportion of ''methylated'' cytosines in a non-CpG context, and this gives an indirect measure of bisulfite conversion efficiency. If the non-CpG methylation is >2%, this indicates insufficient bisulfite conversion and the methylation data are less reliable.

Analysis of DNA methylome data

Timing: >7 days

Following processing and quality control analysis, the extracted methylome data can now be applied to a wide range of downstream analyses that are dependent on the research questions being asked and the overall design of the project. Here we provide some common approaches to analyze and present cancer methylome data, including differential methylation analysis, imputation of missing data and hierarchical clustering/dimensionality reduction analyses.

24. Differential methylation analysis:
    a. To identify interindividual variability in DNA methylation across multiple samples, run the diffmeth program in DMAP using a Chi-squared (-X) test:

```
$ diffmeth -F 2 -t 10 -G /path_to_reference_genome/ -X 40,220 \
-z -R methylome_01_trimmed_bismark_bt2.bam \
-R methylome_02_trimmed_bismark_bt2.bam \
-R methylome_03_trimmed_bismark_bt2.bam \
-R methylome_04_trimmed_bismark_bt2.bam \
> methylome_chi_squared.txt
```

    b. To identify differentially methylated fragments between two samples (-R and -S), run the diffmeth program in DMAP using a Fisher's exact (-P) test:

```
$ diffmeth -F 2 -t 10 -G /path_to_reference_genome/ \
-P 40,220 -z -R methylome_01_trimmed_bismark_bt2.bam \
-S methylome_02_trimmed_bismark_bt2.bam \
> methylome_FE_test.txt
```

    c. To identify differentially methylated fragments between two groups of samples (-R and -S), run the diffmeth program in DMAP using an ANOVA F ratio (-B) test:

```
$ diffmeth -F 2 -t 10 -G /path_to_reference_genome/ \
-B 40,220 -z -R methylome_01_trimmed_bismark_bt2.bam \
-R methylome_02_trimmed_bismark_bt2.bam \
-S methylome_03_trimmed_bismark_bt2.bam \
-S methylome_04_trimmed_bismark_bt2.bam \
> methylome_anova.txt
```

    d. As the above examples involve large numbers of simultaneous statistical tests, there is an increased likelihood of false positives. To adjust P-values for multiple comparisons, run the R program p.adjust (the Bonferroni correction method is used in the following example):

```
>methylome_anova$P_adjust <- p.adjust(methylome_anova$Pvalue, method = ''Bonferroni'', n =
length(methylome_anova$Pvalue))
```

25. Consideration of groupwise analyses with low coverage or missing data:
    a. Run the diffmeth program in DMAP using an ANOVA F ratio test, setting the minimum group size threshold (>15 in the following example, -I 15):

```
$ diffmeth -F 2 -t 10 -G /path_to_reference_genome/ \
-B 40,220 -I 15 -z -R methylome_01_trimmed_bismark_bt2.bam \
-R methylome_02_trimmed_bismark_bt2.bam \
-S methylome_03_trimmed_bismark_bt2.bam \
-S methylome_04_trimmed_bismark_bt2.bam \
> methylome_anova.txt
```

    b. As an alternative, filter out fragments or CpGs with a maximum threshold of missing data points (>5 in the following example) using bash commands:

```
$ cut -f 4-23 group1vsgroup2_methylation.txt | awk -F 'NA' '{print NF-1}' > \ group1_NA_count.txt
#methylation data for group 1 in columns 4-23
$ cut -f 24-43 group1vsgroup2_methylation.txt | awk -F 'NA' '{print NF-1}' > \ group1_NA_count.txt
#methylation data for group 2 in columns 24-43
$ paste group1vsgroup2_methylation.txt group1_NA_count.txt \ group2_NA_count.txt | awk ' $44 <=
5 && $45 <= 5 ' > \ group1vsgroup2_methylation_filtered.txt
```

Note: some samples may not have adequate coverage for specific CpGs or regions. Depending on the size of the groups and whether the groups are of equal size, these CpGs or regions can still be included by setting a minimum threshold to be included in analyses (typically >70% of the group size).

    c. Following statistical analysis, missing data can be imputed but should be interpreted with caution. To impute missing methylation data, run the R package methyLImp:

```
> library(methyLImp)
> methylome_anova.imp <- methyLImp(t(methylome_anova),min=0,max=1)
```

Troubleshooting.

Downstream analysis and visualization

26. Graphical summarization of methylation data:
    a. Hierarchical clustering is a method that can be used to group samples with similar methylation patterns into clusters (Figure 1A). To generate a basic hierarchical clustering dendrogram of methylome data, run the following in R:

```
> distance.matrix <- dist(scale(methylome_anova), method = "euclidean")
> hierarchical.clustering <- hclust(distance.matrix, method = "ward.D2")
> plot(as.dendrogram(hierarchical.clustering))
> library(dendextend)
> distance.matrix <- dist(scale(methylome_anova), method = "euclidean")
```

```
> hierarchical.clustering <- as.dendrogram(hclust(distance.matrix, method = "ward.D2"))
> labels_colors(dend) <- c(rep("blue", 20),rep("red", 20),
rep("green", 20))
> plot(dend)
```

b. When analyzing methylome data, a volcano plot is a type of scatter plot used to show an overview of methylation changes between two groups or individual samples (Figure 1B). To generate a volcano plot of methylome data, run the following in R:

```
> library(ggplot2)
> methylome_anova $diffmethylated <- "NO"
> methylome_anova$diffmethylated[methylome_anova$abs_diffmeth > 0.2 & methylome_anova
$FDR < 0.05] <- "YES" # if absolute methylation difference > 0.2 and FDR-adjusted P-value < 0.05,
set as "YES"
> p <- ggplot(data= methylome_anova, aes(x=methdiff, y=-log10(FDR), colour=diffmethylated)) +
theme_classic() + geom_point(alpha=0.67, size=1) + theme(legend.position = "none") + xlim
(c(-0.5,0.5)) + ylim(c(0, 30)) + xlab("Methylation difference") + ylab(expression(-log[10]~"FDR-
adjusted P-value")) + scale_colour_manual(values = c("darkgray","firebrick"))
```

c. Principal component analysis (PCA) is a method that can transform large amounts of high dimension methylome data into lower dimensions, which are easier to interpret but still retain large amounts of information (Figure 1C). To perform PCA of methylome data, run the following in R:

```
> library(tidyverse)
> library(factoextra)
> pca_prcomp <- prcomp(methylome_anova, center = TRUE, scale = TRUE)
> summary(pca_prcomp)
> pca_plot <- fviz_pca_ind(pca_prcomp, col.ind = "name", palette = c(rep("red", 20), re-
p("blue", 20), rep("green", 20)),
repel = TRUE))
```

d. Heatmaps are commonly used graphical representations of methylome data, where low to high methylation values are shown on a color gradient and methylation patterns are clustered based on similarity (Figure 1D). To generate a heatmap of differentially methylated regions, run the following in R:

```
> library(gplots)
> library(viridis)
> my_palette <- magma(n = 299)
> png("~/methylome_heatmap.png", width = 5*800, height = 5*1200, res = 600, pointsize = 6.5)
> row_distance = dist(methylome.matrix, method = "euclidean")
> row_cluster = hclust(row_distance, method = "ward.D2")
> col_distance = dist(t(methylome.matrix), method = "euclidean")
```

```
> col_cluster = hclust(col_distance, method = "ward.D2")

> par(font.axis=4)

> heatmap.2(methylome.matrix, ColSideColors = c(rep("red", 20), rep("blue", 20), rep("green", 20)), main = "Methylome heatmap", density.info = "none", trace = "none", key.xlab = "Methylation", colsep=c(20,30), col = my_palette, Rowv = as.dendrogram(row_cluster), Colv = as.dendrogram(col_cluster))

> dev.off()
```

e. A powerful approach to understanding the functional consequences of DNA methylation changes is to integrate the DNA methylome data with other genome-scale data from the same samples, such as RNA-seq (transcriptome analysis) and ATAC-seq (chromatin accessibility). In the following example in R, combine the methylome data with transcripts per million (TPM) data derived from RNA-seq; calculate the correlation of methylation at a specific locus and expression of the associated gene:

```
> methylome_transcriptome <- merge(methylome_data, transcriptome_TPM_data, by = "GeneID")

> CD151_meth <- methylome_transcriptome[methylome_transcriptome$GeneID == "CD151",5:7]

> CD151_tpm <- methylome_transcriptome[methylome_transcriptome$GeneID == "CD151",12:14]

> cor.test(CD151_meth,CD151_tpm)
```

Example of integrated methylome and transcriptome data:

Chr	Start	End	CpGs	S1_meth	S2_meth	S3_meth	Enddist	Overlap	GeneID	ENSEMBL_ID	S1_tpm	S2_tpm	S3_tpm	Correlation
1	153959704	153959847	9	0.6184	0.5081	0.56325	3389	on_promoter	RPS27	ENSG00000177954	907.2	878.8	934.3	0.511
5	96855341	96855544	11	0.9635	0.9615	0.9625	336378	intergenic	-	-	-	-	-	
6	34202939	34203113	13	0.2363	0.1784	0.20735	1465	on_promoter	HMGA1	ENSG00000137309	510.9	535	398.6	-0.165
6	34206113	34206191	5	0	0.0347	0.01735	-1613	on_intron	HMGA1	ENSG00000137309	510.9	535	398.6	0.165
9	33750402	33750551	5	0.4924	0.3504	0.4214	-35	on_promoter	PRSS3	ENSG00000010438	244.1	288.5	210.4	-0.567
9	33750975	33751028	6	0.1641	0.2222	0.19315	-512	on_promoter	PRSS3	ENSG00000010438	244.1	288.5	210.4	0.567
11	832495	832579	5	0.3361	0.5294	0.43275	265	on_promoter	CD151	ENSG00000177697	100.9	118.3	88.3	0.577
11	8705317	8705384	6	0.1299	0.375	0.25245	-1425	on_intron	RPL27A	ENSG00000166441	161.3	162.6	179.5	0.068
16	68772278	68772329	6	0.7197	0.2603	0.49	-1200	exon_intron_boundary	CDH1	ENSG00000039068	23.4	16.8	35.1	0.357
17	42403535	42403618	5	0.7111	0.5429	0.627	1298	on_promoter	SLC25A39	ENSG00000013306	132.6	152.5	125.1	-0.701

EXPECTED OUTCOMES

The DNA extraction protocol we present here for FFPE-derived tissue, fresh-frozen tissue, or cell lines yields an ample quantity of high-quality genomic DNA for the generation of reliable DNA methylome data. The DNA quality can be checked on a NanoPhotometer, which should indicate a A260/A280 ratio of 1.7–1.9. The DNA concentration can be determined using a Qubit Fluorometer and dsDNA BR Assay Kit; 25 mg tissue or 5 × 10^6 cells should yield ~400 ng/μL in 50 μL i.e., ~20 μg of DNA).

The DNA methylome library preparation protocol generates a large number of MspI-digested DNA fragments that are adapter-ligated, bisulfite-converted, and size-selected for comprehensive profiling of genome-scale DNA methylation patterns. Using a Qubit Fluorometer and dsDNA BR Assay kit, the final RRBS library should be ~10–60 ng/L. By assessing the library on a Bioanalyzer, the electropherogram profile should display a broad peak at ~260 bp, corresponding to 150–325 bp size-selected fragments.

Following processing and quality control analysis, RRBS data should have high-quality DNA methylation information for ~650,000 MspI fragments, ranging in size from 40 to 220 bp, which collectively

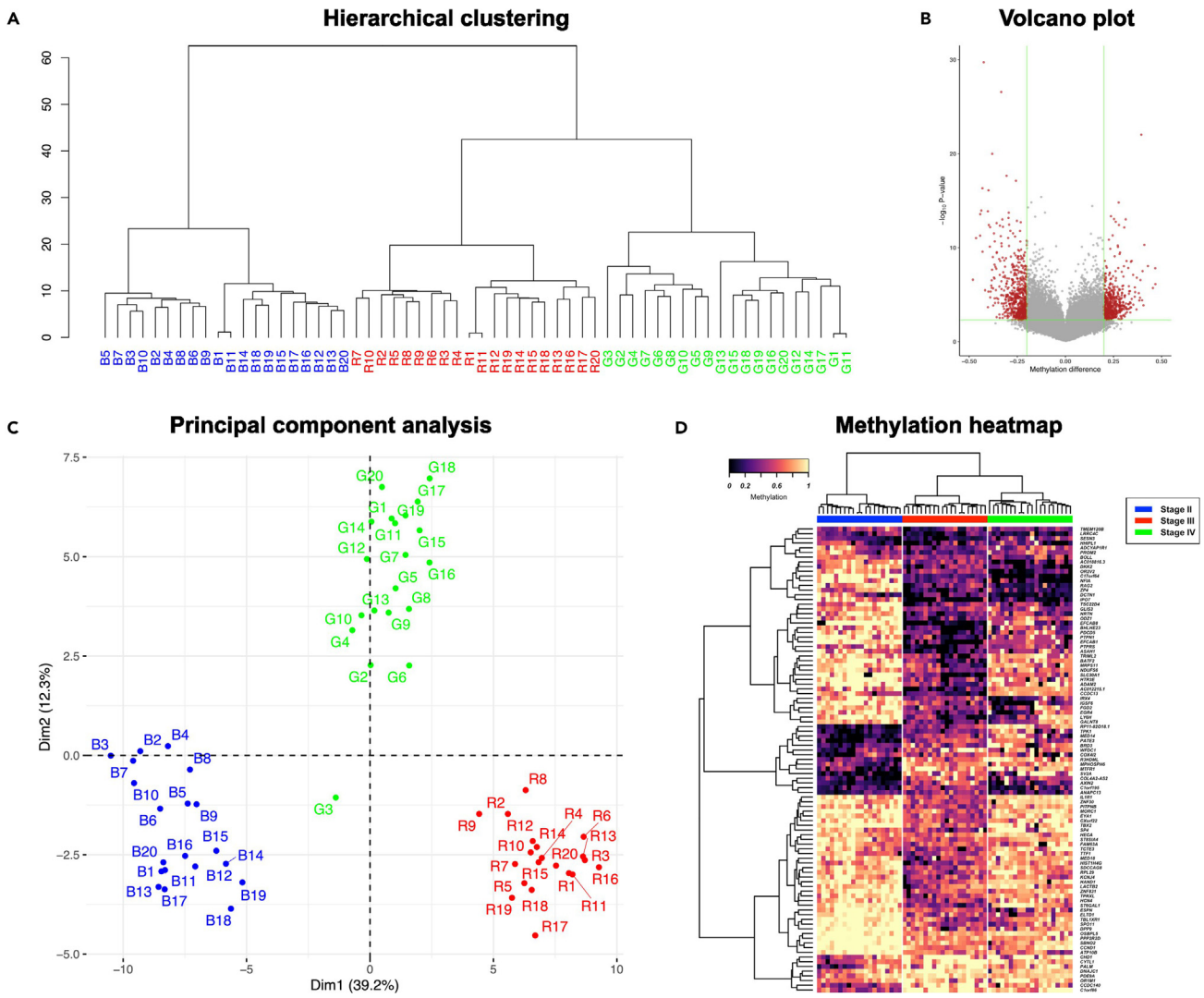


Figure 1. Examples of downstream analysis and visualization of methylome data

(A) Hierarchical clustering groups samples with similar methylation patterns into clusters.

(B) Principal component analysis (PCA) reduces multidimensional methylome data into a format that enables interpretation by identifying clusters of closely related data points.

(C) A volcano plot provides an overview of all methylation differences between two groups or individual samples.

(D) A Heatmap shows low to high methylation values on a color gradient and the overall methylation patterns of samples are clustered based on similarity.

have ~4 million CpG sites at single nucleotide resolution. These fragments are dispersed across the genome, particularly in CpG-dense regions, and cover gene promoters, introns, exons and intergenic regions.

LIMITATIONS

There are two major genome-scale methylation profiling methods, which are both viable alternatives to RRBS – Infinium MethylationEPIC arrays and whole genome bisulfite sequencing (WGBS). MethylationEPIC arrays ensure consistent interrogation of <940 k CpG sites across the genome and are particularly popular for the analysis of large cohorts at low cost. In contrast, although RRBS covers ~4 million CpG sites, the coverage of each site can be variable between samples as this depends on sequence coverage and library preparation variation (for example, PCR bias).

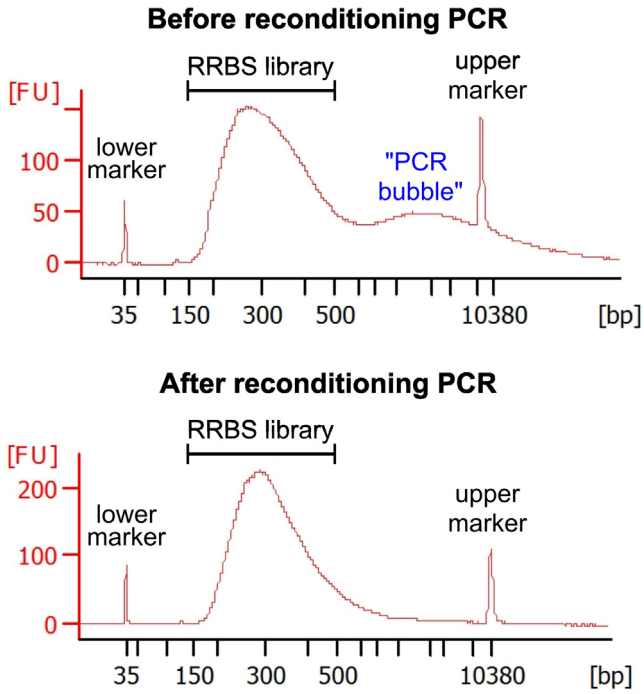


Figure 2. Representative Bioanalyzer electropherogram of a DNA methylome library before and after reconditioning PCR to remove the "PCR bubble" artifact

Although RRBS does not cover the ~28 million CpG sites in the human genome that WGBS does, many of these sites are almost always methylated and therefore analysis of them is not useful for most applications. Furthermore, WGBS requires significantly more sequencing than RRBS, which makes it more resource-intensive. As they are sequencing-based, an advantage of both RRBS and WGBS is that allele-specific methylation or the methylation contribution of cell subpopulations can be determined. We have provided the comparative description and limitations and advantages of these methods previously.[24]

All these methods require sodium bisulfite treatment, which involves denaturing the DNA in a high concentration of sodium bisulfite salt at low pH and high temperature, which often leads to high levels of DNA degradation. The sodium bisulfite treatment results in the conversion of an unmethylated cytosine to uracil (and then thymine following PCR amplification) but a methylated cytosine remains unaffected, allowing for methylated cytosines to be distinguished from unmethylated cytosines. However, it should be noted that bisulfite sequencing does not distinguish between methylated cytosine and another distinct epigenetic mark, hydroxymethylated cytosine.

The standard RRBS protocol presented here can reliably generate cancer methylome data with a DNA input of 100–500 ng. Although we have used this method to generate libraries with as little as 10 ng of DNA, the genomic complexity was reduced, and the overall coverage was lower than normal. However, one strategy to address this issue would be to increase the sequencing depth. Several RRBS methods have been developed for lower DNA input, such as single-cell RRBS (sc-RRBS),[26] cell-free RRBS (cf-RRBS)[27] and cfMethyl-Seq.[28]

TROUBLESHOOTING

Problem 1

High levels of DNA degradation (related to Step 7).

Potential solution

To generate reliable methylome data, the input DNA needs to have minimal degradation. For some FFPE tissues, degradation may not be avoidable as it depends on how the tissue was fixed, how it was stored, and the age of the block. Various methods can be used to assess DNA integrity in addition to the A260/A280 ratio described in the protocol. For example, the DNA can be run on a standard agarose gel or Bioanalyzer - intact DNA should appear as a compact and high molecular weight band, whereas degraded DNA should be visible as a smear. A multiplex PCR assay of different target sites in the *GAPDH* gene has been developed where intact DNA should yield 100, 200, 300 and 400 bp products, but degraded DNA would only amplify the smaller products (or none at all).[29] If there is any question about the DNA quality, we recommend a thorough assessment before proceeding with methylome library preparation.

Problem 2

Bioanalyzer results indicate artifacts in the methylome library (related to Step 24).

Potential solution

The Bioanalyzer electropherogram of the RRBS library should display a broad peak at ~260 bp, corresponding to 150–325 bp size-selected fragments. A sharp peak at ~125 bp is indicative of adapter dimers, which negatively impacts library sequencing. Excess adapter dimers should be removed using a one-step purification with TruSeq Sample Purification beads (1.8× volume ratio), similar to step 23. The Bioanalyzer profile may also show a PCR artifact referred to as a "PCR bubble", where the size distribution of the library is significantly more than expected. This artifact is a result of PCR products annealing to each other after depletion of available primers, which cannot be eliminated by bead-based or Blue Pippin size selection. However, the PCR bubble can be removed by using a "reconditioning PCR", which involves reamplifying a tenfold-diluted library with a low number (~3) of cycles[30] (Figure 2).

Problem 3

Low genome alignment efficiency (related to Step 29).

Potential solution

We routinely obtain 60%–70% alignment efficiency for libraries generated from fresh-frozen tissues and cultured cells, and this is sometimes considerably lower for FFPE tissues, depending on the quality of the DNA. Our Bismark alignment is typically quite stringent, only allowing for one mismatch in the seed; relaxing the number of mismatches to two (the default) could improve the mapping efficiency to a small degree. Alternatively, the alignment could be improved by iteratively optimizing the trimming and alignment parameters (e.g., the alignment score stringency can be altered using the –score_min <func> option).

Problem 4

Low coverage of reads per MspI fragment or CpG (related to Step 30).

Potential solution

For RRBS, the coverage of aligned reads per CpG across the genome is highly variable. Depending on the total number of sequenced reads and the alignment efficiency, the coverage can range from 0 to several thousand reads per CpG site. The higher the coverage, the more certainty there is that the methylation of a CpG or a fragment has been accurately called. We typically use a minimum threshold of 10 reads per MspI fragment (diffmeth -t 10) or per CpG site (bismark_methylation_extractor –cutoff 10). If the total number of methylation values per sample is significantly less than expected (i.e., ~650,000 MspI fragments or ~4 million CpG sites), this is possibly because of: 1. An insufficient number of sequencing reads per sample; 2. Decreased alignment efficiency due to DNA degradation; or 3. Incomplete MspI digestion. To address the issue of low read coverage, the minimum threshold of reads per MspI fragment or CpG could be lowered to 5. Furthermore,

to increase coverage the RRBS library can be resequenced and the data combined to increase the total number of reads per sample.

Problem 5

Small differences in methylation between groups (related to Step 31).

Potential solution

In cancer, there are often large differences in methylation at specific loci and for this reason we typically identify differentially methylated regions using an adjusted P-value < 0.05 and a methylation difference ≥ 20%. However, depending on the research question or disease context, it is possible that the methylation differences are more subtle. For example, in a groupwise comparison the mean methylation difference between the groups could be small, possibly due to heterogeneity within the groups. Likewise, in the study of chronic diseases[6,31,32] or for analysis of an environmental factor or phenotype,[12,33] methylation differences may initially be small at early time points because changes in DNA methylation are replication-dependent and therefore take time to be altered and effect size of methylation will vary depending on the phenotype in question. To capture regions where the changes in methylation are not as pronounced, the minimum difference threshold can be lowered (e.g., ≥ 10%). The visualization tools described previously are helpful in assessing the nature and extent of methylation differences (e.g., volcano plots show the range of all methylation differences, heatmaps can be used to observe heterogeneity and subpopulations).

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources and reagents should be directed to and will be fulfilled by the lead contact, Aniruddha Chatterjee (aniruddha.chatterjee@otago.ac.nz).

Materials availability

This study did not generate new materials.

Data and code availability

This study did not generate new datasets or code.

ACKNOWLEDGMENTS

The authors appreciate funding from the Royal Society of New Zealand Te Apārangi (for a Rutherford Discovery Fellowship to A.C. and Marsden Fund to E.J.R. and A.C.) and the Health Research Council of New Zealand (to E.J.R. and A.C.). We are also grateful for additional support from the TD Scott Chair in Urology and Department of Surgical Sciences (to A.C. and E.J.R.). We would like to thank Professor Ian Morison (Department of Pathology) for his input in the development of this protocol and acknowledge the Otago Genomics Facility (University of Otago) for providing sequencing service and assistance.

AUTHOR CONTRIBUTIONS

Conceptualization, E.J.R. and A.C.; methodology, E.J.R., S.A., P.A.S., and A.C.; biological resources, E.J.R. and A.C.; data analysis, E.J.R., S.A., and P.A.S.; writing – original draft, E.J.R.; writing – review and editing, all authors; funding acquisition, E.J.R. and A.C. All authors read and approved the final version of the manuscript.

DECLARATION OF INTERESTS

The authors declare no competing interests.

REFERENCES

1. Rodger, E.J., Gimenez, G., Ajithkumar, P., Stockwell, P.A., Almomani, S., Bowden, S.A., Leichter, A.L., Ahn, A., Pattison, S., McCall, J.L., et al. (2023). An epigenetic signature of advanced colorectal cancer metastasis. iScience *26*, 106986. https://doi.org/10.1016/j.isci.2023.106986.

2. Rodger, E.J., Almomani, S.N., Ludgate, J.L., Stockwell, P.A., Baguley, B.C., Eccles, M.R., and Chatterjee, A. (2021). Comparison of Global DNA Methylation Patterns in Human Melanoma Tissues and Their Derivative Cell Lines. Cancers *13*, 2123. https://doi.org/10.3390/cancers13092123.

3. Tiffen, J., Gallagher, S.J., Filipp, F., Gunatilake, D., Emran, A.A., Cullinane, C., Dutton-Register, K., Aoude, L., Hayward, N., Chatterjee, A., et al. (2020). EZH2 Cooperates with DNA Methylation to Downregulate Key Tumor Suppressors and IFN Gene Signatures in Melanoma. J. Invest. Dermatol. *140*, 2442–2454.e5. https://doi.org/10.1016/j.jid.2020.02.042.

4. Chatterjee, A., Rodger, E.J., Ahn, A., Stockwell, P.A., Parry, M., Motwani, J., Gallagher, S.J., Shklovskaya, E., Tiffen, J., Eccles, M.R., and Hersey, P. (2018). Marked Global DNA Hypomethylation Is Associated with Constitutive PD-L1 Expression in Melanoma. iScience *4*, 312–325. https://doi.org/10.1016/j.isci.2018.05.021.

5. Chatterjee, A., Stockwell, P.A., Ahn, A., Rodger, E.J., Leichter, A.L., and Eccles, M.R. (2017). Genome-wide methylation sequencing of paired primary and metastatic cell lines identifies common DNA methylation changes and a role for EBF3 as a candidate epigenetic driver of melanoma metastasis. Oncotarget *8*, 6085–6101. https://doi.org/10.18632/oncotarget.14042.

6. Helliwell, A.M., Sweetman, E.C., Stockwell, P.A., Edgar, C.D., Chatterjee, A., and Tate, W.P. (2020). Changes in DNA methylation profiles of myalgic encephalomyelitis/chronic fatigue syndrome patients reflect systemic dysfunctions. Clin. Epigenetics *12*, 167. https://doi.org/10.1186/s13148-020-00960-z.

7. Thunders, M., Holley, A., Harding, S., Stockwell, P., and Larsen, P. (2019). Using NGS-methylation profiling to understand the molecular pathogenesis of young MI patients who have subsequent cardiac events. Epigenetics *14*, 536–544. https://doi.org/10.1080/15592294.2019.1605815.

8. Bowden, S.A., Rodger, E.J., Bates, M., Chatterjee, A., Eccles, M.R., and Stayner, C. (2018). Genome-Scale Single Nucleotide Resolution Analysis of DNA Methylation in Human Autosomal Dominant Polycystic Kidney Disease. Am. J. Nephrol. *48*, 415–424. https://doi.org/10.1159/000494739.

9. Chatterjee, A., Macaulay, E.C., Rodger, E.J., Stockwell, P.A., Parry, M.F., Roberts, H.E., Slatter, T.L., Hung, N.A., Devenish, C.J., and Morison, I.M. (2016). Placental Hypomethylation Is More Pronounced in Genomic Loci Devoid of Retroelements. G3 (Bethesda) *6*, 1911–1921. https://doi.org/10.1534/g3.116.030379.

10. Chatterjee, A., Stockwell, P.A., Rodger, E.J., Duncan, E.J., Parry, M.F., Weeks, R.J., and Morison, I.M. (2015). Genome-wide DNA methylation map of human neutrophils reveals widespread inter-individual epigenetic variation. Sci. Rep. *5*, 17328. https://doi.org/10.1038/srep17328.

11. Tschirley, A.M., Stockwell, P.A., Rodger, E.J., Eltherington, O., Morison, I.M., Christensen, N., Chatterjee, A., and Hibma, M. (2021). The Mouse Papillomavirus Epigenetic Signature Is Characterised by DNA Hypermethylation after Lesion Regression. Viruses *13*, 2045. https://doi.org/10.3390/v13102045.

12. Falisse, E., Ducos, B., Stockwell, P.A., Morison, I.M., Chatterjee, A., and Silvestre, F. (2018). DNA methylation and gene expression alterations in zebrafish early-life stages exposed to the antibacterial agent triclosan. Environ. Pollut. *243*, 1867–1877. https://doi.org/10.1016/j.envpol.2018.10.004.

13. Chatterjee, A., Lagisz, M., Rodger, E.J., Zhen, L., Stockwell, P.A., Duncan, E.J., Horsfield, J.A., Jeyakani, J., Mathavan, S., Ozaki, Y., and Nakagawa, S. (2016). Sex differences in DNA methylation and expression in zebrafish brain: a test of an extended 'male sex drive' hypothesis. Gene *590*, 307–316. https://doi.org/10.1016/j.gene.2016.05.042.

14. Chatterjee, A., Ozaki, Y., Stockwell, P.A., Horsfield, J.A., Morison, I.M., and Nakagawa, S. (2013). Mapping the zebrafish brain methylome using reduced representation bisulfite sequencing. Epigenetics *8*, 979–989. https://doi.org/10.4161/epi.25797.

15. Chatterjee, A., Stockwell, P.A., Horsfield, J.A., Morison, I.M., and Nakagawa, S. (2014). Base-resolution DNA methylation landscape of zebrafish brain and liver. Genom. Data *2*, 342–344. https://doi.org/10.1016/j.gdata.2014.10.008.

16. Voisin, A.S., Suarez Ulloa, V., Stockwell, P., Chatterjee, A., and Silvestre, F. (2022). Genome-wide DNA methylation of the liver reveals delayed effects of early-life exposure to 17-alpha-ethinylestradiol in the self-fertilizing mangrove rivulus. Epigenetics *17*, 473–497. https://doi.org/10.1080/15592294.2021.1921337.

17. Langmead, B., and Salzberg, S.L. (2012). Fast gapped-read alignment with Bowtie 2. Nat. Methods *9*, 357–359. https://doi.org/10.1038/nmeth.1923.

18. Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., and Durbin, R.; 1000 Genome Project Data Processing Subgroup (2009). The Sequence Alignment/Map format and SAMtools. Bioinformatics *25*, 2078–2079. https://doi.org/10.1093/bioinformatics/btp352.

19. Krueger, F., and Andrews, S.R. (2011). Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications. Bioinformatics *27*, 1571–1572. https://doi.org/10.1093/bioinformatics/btr167.

20. Stockwell, P.A., Chatterjee, A., Rodger, E.J., and Morison, I.M. (2014). DMAP: differential methylation analysis package for RRBS and WGBS data. Bioinformatics *30*, 1814–1822. https://doi.org/10.1093/bioinformatics/btu126.

21. Quinlan, A.R., and Hall, I.M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics *26*, 841–842. https://doi.org/10.1093/bioinformatics/btq033.

22. Ludgate, J.L., Wright, J., Stockwell, P.A., Morison, I.M., Eccles, M.R., and Chatterjee, A. (2017). A streamlined method for analysing genome-wide DNA methylation patterns from low amounts of FFPE DNA. BMC Med. Genomics *10*, 54. https://doi.org/10.1186/s12920-017-0290-1.

23. Meissner, A., Mikkelsen, T.S., Gu, H., Wernig, M., Hanna, J., Sivachenko, A., Zhang, X., Bernstein, B.E., Nusbaum, C., Jaffe, D.B., et al. (2008). Genome-scale DNA methylation maps of pluripotent and differentiated cells. Nature *454*, 766–770. https://doi.org/10.1038/nature07107.

24. Chatterjee, A., Rodger, E.J., Morison, I.M., Eccles, M.R., and Stockwell, P.A. (2017). Tools and Strategies for Analysis of Genome-Wide and Gene-Specific DNA Methylation Patterns. Methods Mol. Biol. *1537*, 249–277. https://doi.org/10.1007/978-1-4939-6685-1_15.

25. Al Momani, S., Rodger, E.J., Stockwell, P.A., Eccles, M.R., and Chatterjee, A. (2022). Generating Sequencing-Based DNA Methylation Maps from Low DNA Input Samples. Methods Mol. Biol. *2458*, 3–21. https://doi.org/10.1007/978-1-0716-2140-0_1.

26. Guo, H., Zhu, P., Wu, X., Li, X., Wen, L., and Tang, F. (2013). Single-cell methylome landscapes of mouse embryonic stem cells and early embryos analyzed using reduced representation bisulfite sequencing. Genome Res. *23*, 2126–2135. https://doi.org/10.1101/gr.161679.113.

27. Van Paemel, R., De Koker, A., Vandeputte, C., van Zogchel, L., Lammens, T., Laureys, G., Vandesompele, J., Schleiermacher, G., Chicard, M., Van Roy, N., et al. (2021). Minimally invasive classification of paediatric solid tumours using reduced representation bisulphite sequencing of cell-free DNA: a proof-of-principle study. Epigenetics *16*, 196–208. https://doi.org/10.1080/15592294.2020.1790950.

28. Stackpole, M.L., Zeng, W., Li, S., Liu, C.C., Zhou, Y., He, S., Yeh, A., Wang, Z., Sun, F., Li, Q., et al. (2022). Cost-effective methylome sequencing of cell-free DNA for accurately detecting and locating cancer. Nat. Commun. *13*, 5566. https://doi.org/10.1038/s41467-022-32995-6.

29. van Beers, E.H., Joosse, S.A., Ligtenberg, M.J., Fles, R., Hogervorst, F.B.L., Verhoef, S., and Nederlof, P.M. (2006). A multiplex PCR predictor for aCGH success of FFPE samples. Br. J. Cancer *94*, 333–337. https://doi.org/10.1038/sj.bjc.6602889.

30. Thompson, J.R., Marcelino, L.A., and Polz, M.F. (2002). Heteroduplexes in mixed-template amplifications: formation, consequence and elimination by 'reconditioning PCR'. Nucleic Acids Res. *30*, 2083–2088. https://doi.org/10.1093/nar/30.9.2083.

31. Helliwell, A.M., Stockwell, P.A., Edgar, C.D., Chatterjee, A., and Tate, W.P. (2022). Dynamic Epigenetic Changes during a Relapse and Recovery Cycle in Myalgic Encephalomyelitis/Chronic Fatigue Syndrome. Int. J. Mol. Sci. *23*, 11852. https://doi.org/10.3390/ijms231911852.

32. Bowden, S.A., Stockwell, P.A., Rodger, E.J., Parry, M.F., Eccles, M.R., Stayner, C., and Chatterjee, A. (2020). Extensive Inter-Cyst DNA Methylation Variation in Autosomal Dominant Polycystic Kidney Disease Revealed by Genome Scale Sequencing. Front. Genet. *11*, 348. https://doi.org/10.3389/fgene.2020.00348.

33. Voisin, A.S., Suarez Ulloa, V., Stockwell, P., Chatterjee, A., and Silvestre, F. (2022). Genome-wide DNA methylation of the liver reveals delayed effects of early-life exposure to 17-alpha-ethinylestradiol in the self-fertilizing mangrove rivulus. Epigenetics *17*, 473–497. https://doi.org/10.1080/15592294.2021.1921337.