Title

Targeted bisulfite sequencing for biomarker discovery.

Authors
Morselli, Marco
Farrell, Colin
Rubbi, Liudmilla
et al.

Publication Date
2021-03-01

DOI
10.1016/j.ymeth.2020.07.006

Peer reviewed

Published in final edited form as:
Methods. 2021 March ; 187: 13–27. doi:10.1016/j.ymeth.2020.07.006.

Targeted Bisulfite Sequencing for Biomarker Discovery

Marco Morselli[a,b,c,*], **Colin Farrell**[a], **Liudmilla Rubbi**[a], **Heather L. Fehling**[d], **Rebecca Henkhaus**[d], **Matteo Pellegrini**[a,b,c,*]

[a]Department of Molecular, Cell and Developmental Biology, University of California Los Angeles, Los Angeles, California, 90095, United States

[b]UCLA-DOE Institute for Genomics and Proteomics, University of California Los Angeles, Los Angeles, California, 90095, United States

[c]Institute for Quantitative and Computational Biosciences – The Collaboratory, University of California Los Angeles, Los Angeles, California, 90095, United States

[d]Clinical Reference Laboratory, Inc., Lenexa, Kansas, 66215, United States

Abstract

[*] Corresponding Authors: mmorselli@ucla.edu, marco.murslegn@gmail.com, matteop@mcdb.ucla.edu.

CRediT Author Statement
Marco Morselli: Conceptualization, Methodology, Investigation, Writing - Original Draft, Writing - Review/Editing, Visualization.
Colin Farrell: Conceptualization, Methodology, Software, Formal Analysis, Data Curation, Writing - Original Draft, Writing - Review/Editing, Visualization.
Liudmilla Rubbi: Methodology, Investigation, Writing - Review/Editing, Supervision.
Rebecca Henkhaus: Methodology, Investigation, Resources.
Heather Fehling: Methodology, Investigation, Resources.
Matteo Pellegrini: Conceptualization, Methodology, Formal Analysis, Writing - Review/Editing, Supervision.

10. Author's Contributions
Marco Morselli: Conceptualization, Methodology, Investigation, Writing - Original Draft, Writing - Review/Editing, Visualization. **Colin Farrell:** Conceptualization, Methodology, Software, Formal Analysis, Data Curation, Writing - Original Draft, Writing - Review/Editing, Visualization. **Liudmilla Rubbi:** Methodology, Investigation, Writing - Review/Editing, Supervision. **Rebecca Henkhaus:** Methodology, Investigation, Resources. **Heather Fehling:** Methodology, Investigation, Resources. **Matteo Pellegrini:** Conceptualization, Methodology, Formal Analysis, Writing - Review/Editing, Supervision.

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

11. Informed Consent
The retrospective human blood samples used in this study were de-identified samples. Informed Consent was obtained.

13. Supplementary Material

- **13.1.** Supplementary Data Analysis
- **13.2.** Supplementary Tables
  - **13.2.1.** IDT Pre-methylated Unique-Dual Indices (UDI) Adapters
  - **13.2.2.** Probes coordinates (hg38 UCSC)
  - **13.2.3.** Methylation Value Imputation Parameters
  - **13.2.4.** Mapping Stats
- **13.3.** Code for Data Processing: https://github.com/NuttyLogic/MethodsTBS

14. Declaration of interest:
'Declarations of interest: none'.

Cytosine methylation is one of the best studied epigenetic modifications. In mammals, DNA methylation patterns vary among cells and is mainly found in the CpG context. DNA methylation is involved in important processes during development and differentiation and its dysregulation can lead to or is associated with diseases, such as cancer, loss-of-imprinting syndromes and neurological disorders. It has been also shown that DNA methylation at the cellular, tissue and organism level varies with age. To overcome the costs of Whole-Genome Bisulfite Sequencing, the gold standard method to detect methylation at a single base resolution, DNA methylation arrays have been developed and extensively used. This method allows one to assess the status of a fraction of the CpG sites present in the genome of an organism. In order to combine the relatively low cost of Methylation Arrays and digital signals of bisulfite sequencing, we developed a Targeted Bisulfite Sequencing method that can be applied to biomarker discovery for virtually any phenotype. Here we describe a comprehensive step-by-step protocol to build a DNA methylation-based epigenetic clock.

Keywords

Next-Generation Sequencing; Target Bisulfite-seq; DNA methylation; Biomarker Discovery; Epigenetic Clock

1. Introduction

The methylation of the 5th carbon of cytosines is a covalent modification found in many organisms, such as prokaryotes, fungi, algae, plants and animals [1–4]. In mammals the reaction is catalyzed by the activity of three enzymes called DNA Methyltransferases (DNMT3a, DNMT3b and DNMT1) and it is found predominantly in the CpG dinucleotide context [5,6]. For each cell type, DNA methylation patterns are established during development and differentiation and faithfully maintained through cell division [7]. DNA methylation is involved in numerous processes such as genomic imprinting, X-chromosome inactivation, genome stability and transcriptional regulation [8]. Aberrant DNA methylation patterns have been observed in loss-of-imprinting syndromes, many cancer types, autoimmune diseases, and metabolic, neurological and psychological disorders [9]. DNA methylation changes have not been observed only during differentiation or in diseases, but also during the aging process, at a cellular, tissue, and organism level [10–14].

Recently, DNA methylation-based biomarkers have been developed to estimate the age of an individual, known as epigenetic clocks [10,15–17], and other traits such as BMI [18], smoking [18,19] and type two diabetes [20]. The difference between the epigenetic and the chronological age can inform on the biological or physiological age of an individual [21]. Other biomarkers have been utilized to predict biological age, but DNA methylation is generally more accurate than other approaches [22].

Several methods have been described to detect DNA methylation and they can be classified into four major groups:

1. methylation-specific restriction endonucleases;

2. Immunoprecipitation using anti-5^{me}C antibodies or affinity purification by methyl-DNA binding proteins;
3. Sodium bisulfite treatment;
4. Sequencing of the native DNA molecule using third-generation sequencing technologies (Pacific Biosciences and Oxford Nanopore).

The first approach is limited to the enzyme recognized sequences, and the second group doesn't yield single-nucleotide resolution. Third-generation sequencing technologies allow for the detection of modified bases, but since they are reading native molecules, they usually have high-input requirements. Bisulfite treatment converts unmethylated cytosines into uracil, but it doesn't affect 5meC under the conditions used [23]. After PCR amplification, the information regarding the status of each individual cytosine methylation is then calculated by assessing the ratio of C (originally 5meC) and T (originally unmethylated C, then converted to U, and amplified as T). This chemical treatment has a very high conversion rate (>99%) resulting in very accurate single-nucleotide information. However, bisulfite treatment causes degradation of the treated DNA and it can lead to biases in the amplification efficiency of fragments with different methylation status. Nonetheless, bisulfite treatment is still typically considered the gold- standard method for 5meC detection.

Four main technologies are used to read the signal after Bisulfite conversion: PCR, Sanger Sequencing, microarrays and Next-Generation Sequencing. PCR-based readout methods are utilized when the methylation status of a limited number of cytosines has to be assessed. Sanger Sequencing can be utilized when a limited number of regions are analyzed. The methylation status of several CpGs can be assessed, but it requires the laborious step of cloning and colony isolation, making it a low-throughput technique. By contrast, both methylation arrays and NGS can detect the status of thousands or millions of CpG sites. Both have advantages and disadvantages. The advantages of array-based methods are cost, single-nucleotide resolution, accuracy, and a good balance between coverage and throughput, making them one of the top choices for epigenome-wide studies of large cohorts of samples. However, array-based methods can suffer from batch-effects if not handled properly and the presence of cross-hybridization to probes because of the low-complexity of the bisulfite-treated DNA input. Moreover, commercially available arrays are available for a limited number of organisms. Whole-genome bisulfite-sequencing (WGBS) or MethylC-seq [24–26] couples the bisulfite treatment with high-throughput sequencing and it is considered the gold standard method for genome-wide studies. It provides the methylation status of most of the cytosines in a given genome. Despite the decrease in sequencing costs, the high number of reads required by this approach makes it expensive, especially for big genomes and/or large-scale studies.

Several methodologies have been developed to reduce the cost of the assay and focus the sequencing only on a small subset of sites. One approach, Reduced- Representation Bisulfite Sequencing (RRBS) [27–31], utilizes a methylation-independent enzyme (or enzymes) to select CG-rich regions, that are usually enriched in regulatory elements (CpG islands, promoters, and enhancers). RRBS allows the interrogation of a fraction of the CpG sites in the genome (1.5–2 M) for a fraction of the cost of WGBS, but it suffers from non-even coverage and it selects for regions with no or low methylation variability, effectively decreasing the amount of valuable information. Moreover, the analyzed regions are limited to the presence of the restriction site(s). In order to circumvent these problems, enrichment strategies based on hybridization probes have been developed. Several commercial products have been developed to enrich for a specific set of genes (such as disease-specific panels) or to target most or all regulatory regions, thus having the ability to enrich for thousands to millions of CpG sites [32,33]. The capture can either occur before or after bisulfite conversion. The first approach guarantees a better enrichment since it hybridizes probes to an unconverted genome, but it requires a larger input amount of DNA to reduce the loss of complexity during the library preparation, target capture, and bisulfite-conversion steps. The second approach results in higher complexity of the captured molecules, but the hybridization of converted DNA produces more off-target reads and it requires the design of probes encompassing all the possible methylation combinations to reduce enrichment biases.

Here we describe a method that utilizes biotinylated RNA probes to capture a small fraction of the genome (around 4 Mb) before bisulfite conversion and sequencing. We present an optimized sequencing processing pipeline built to handle large quantities of targeted bisulfite sequencing data efficiently. Finally, we highlight the utility of a targeted approach by fitting a penalized regression model of epigenetic aging of several individuals. While the approach highlighted uses epigenetic age as an example, the method can be utilized to model other phenotypes of interest, given that the probes are targeted to informative regions.

2. Materials

2.1. Reagents

Reagent	Supplier	Catalog #	Notes
Blood collection, DNA Extraction and Quantification			
Lavender Top EDTA tube (6 mL)	BD	367863	or equivalent
QIAamp DNA Blood Midi kit (100)	QIAGEN	51185	
Content: • QIAamp Midi Spin Columns • Collection Tubes (15 ml) • Buffer AL • Buffer AW1 (see 2.2.1) • Buffer AW2 (see 2.2.2) • Buffer AE • QIAGEN Protease			
PBS Buffer (1x) pH 7.4	ThermoFisher Scientific	10010023	or equivalent
Qubit™ dsDNA HS Assay	ThermoFisher Scientific	Q32854	
Content:			

Reagent	Supplier	Catalog #	Notes
• Qubit® dsDNA HS Reagent (200X concentrate in DMSO) • Qubit® dsDNA HS Buffer • Qubit® dsDNA HS Standard #1 (0 ng/μL in TE buffer) • Qubit® dsDNA HS Standard #2 (10 ng/μL in TE buffer)			
Qubit™ dsDNA BR Assay	ThermoFisher Scientific	Q32853	
Content: • Qubit® dsDNA BR Reagent (200X concentrate in DMSO) • Qubit® dsDNA BR Buffer • Qubit® dsDNA BR Standard #1 (0 ng/μL in TE buffer) • Qubit® dsDNA BR Standard #2 (100 ng/μL in TE buffer)			
Unmethylated lambda phage genomic DNA (250 μg*)	Promega	D1521	*Measure the concentration of the DNA using the Qubit BR dsDNA Assay. Genome sequence: GenBank #J02459
Library Preparation			
UltraPure™ 1M Tris-HCI, pH 8.0	ThermoFisher Scientific	15568025	or equivalent
High Sensitivity D1000 Assay	Agilent Technologies		
Content: • High Sensitivity D1000 ScreenTape • High Sensitivity D1000 Reagents (Sample Buffer + Ladder)		5067– 5584 5067– 5585	
NEBNext® Ultra™ II DNA Library Prep with Sample Purification Beads	New England Biolabs	E7103S/L (24/96 samples)	Alternative formats: NEBNext® Ultra™ II DNA Library Prep Kit for Illumina (E7645S/L) and separate Purification Beads (see Note 9.3)
Content: • NEBNext Ligation Enhancer • NEBNext Ultra II End Prep Enzyme Mix • NEBNext Ultra II End Prep Reaction Buffer • NEBNext Ultra II Ligation Master Mix • NEBNext Ultra II Q5® Master Mix (2x) • NEBNext Sample Purification Beads (store at room temperature)			
Ethanol, Absolute (200 Proof), Molecular Biology Grade	Fisher Scientific	64-17-5	or equivalent
Hybridization Capture			
UltraPure™ DNase/RNase-Free Distilled Water	ThermoFisher Scientific	10977015	or equivalent
myBaits Custom 1–20K DNA-seq kit	Arbor Biosciences	300116 (16 captures)	Other formats are available. Each

Reagent	Supplier	Catalog #	Notes
			reaction corresponds to a capture.
Content: • Hyb N (red cap) • Hyb S (teal cap) • Beads (streptavidin beads) • Binding Buffer • Wash Buffer • Hyb D (yellow cap) • Hyb R (purple cap) • Block C (green cap) • Block O (blue cap) • Block A (orange cap) • Baits (white cap)			Order TruSeq-style double-index blockers.
Tween-20 (10% solution)	Sigma-Aldrich	11332465001	or equivalent
EDTA 0.5 M, pH 8	Sigma-Aldrich	03690–100ML	or equivalent
Bisulfite Conversion and Library Amplification			
EZ DNA methylation-Lightning® kit	Zymo Research	D5030	Other formats are available
Content: • Lightning Conversion Reagent • M-Binding Buffer • M-Wash Buffer • L-Desulphonation Buffer • M-Elution Buffer • Zymo-Spin IC Columns • Collection Tubes			
KAPA HiFi HotStart Uracil+ ReadyMix Kit	Roche Sequencing	7959052001	KAPA cat #: KK2801

2.2. Solutions, Master Mixes, and Buffers

ID	Buffer	Ingredients	Supplier	cat #
2.2.1	Buffer AW1	Buffer AW1 (concentrate) Ethanol (add as indicated on the bottle)	QIAGEN Fisher Scientific	part of 51185 64-17-5
	Notes: Store at room temperature			
2.2.2	Buffer AW2	Buffer AW2 (concentrate) Ethanol (add as indicated on the bottle)	QIAGEN Fisher Scientific	part of 51185 64-17-5
	Notes: Store at room temperature			
2.2.3	EB Buffer	10 mM Tris-HCl pH 8	ThermoFisher Scientific	15568025

ID	Buffer	Ingredients	Supplier	cat #
	Notes: Dilute with Nuclease-free water. Store at Room Temperature			
2.2.4	EndPrep Master Mix	N × 3 µl of NEBNext Ultra II End Prep Enzyme Mix	New England Biolabs	part of E7103S/L
		N × 7 µl of End Prep Reaction Buffer	New England Biolabs	part of E7103S/L
	Notes: Prepare just before use. Mix thoroughly and store in ice until ready to use. N = number of samples			
2.2.5	Ligation Master Mix	N × 30 µl of NEBNext Ligation Master Mix	New England Biolabs	part of E7103S/L
		N × 1 µl of Ligation Enhancer	New England Biolabs	part of E7103S/L
	Notes: prepare just before use, stable up to 8 h at 4°C. Because of the viscosity of the solution, make sure it's mixed thoroughly. Store in ice until ready to use. N = number of samples			
2.2.6	Blockers Mix	H × 0.5 µl of Block A	Arbor Biosciences	part of 300116
		H × 2.5 µl of Block C	Arbor Biosciences	part of 300116
		H × 2.5 µl of Block O	Arbor Biosciences	part of 300116
	Notes: Prepare just before use in a low-bind tube. For each capture reaction, aliquot 5 µL of Blockers Mix to a low-bind 0.2 ml tube. The tubes are now called LIB. H = number of capture reactions.			
2.2.7	Hybridization Mix (HYB)	Hyb NDR mix:		
		H × 9.25 µl Hyb N	Arbor Biosciences	part of 300116
		H × 3.5 µl of Hyb D	Arbor Biosciences	part of 300116
		H × 1.25 µl of Hyb R	Arbor Biosciences	part of 300116
		H × 0.5 µl of Hyb S	Arbor Biosciences	part of 300116
		H × 5.5 µl of custom RNA baits	Arbor Biosciences	part of 300116
	Notes: Prepare just before use in a 1.7 ml tube. Warm the Hyb NDR mix tube and Hyb S tube at 60°C for 5 min. Add the appropriate amount of Hyb S and RNA baits to the Hyb NDR tube. This tube is now called HYB. H = number of capture reactions.			
2.2.8	Wash Buffer X	1–9 captures:		
		H × 12.0 µl of Hyb S	Arbor Biosciences	part of 300116
		H × 300 µl of Wash Buffer	Arbor Biosciences	part of 300116
		H × 1188 µl of Nuclease-free H_2O	ThermoFisher Scientific	10977015
		10 captures:		
		100 µl of Hyb S	Arbor Biosciences	part of 300116
		2.5 ml of Wash Buffer	Arbor Biosciences	part of 300116
		9.9 ml of Nuclease-free H_2O	ThermoFisher Scientific	10977015
	Notes: stable at 4°C for up to 1 month. H = number of capture reactions.			
2.2.9	SBE Buffer (Streptavidin-Beads Elution Buffer)	10 mM Tris-HCl, pH 8	ThermoFisher Scientific	15568025
		0.05% Tween-20	Sigma-Aldrich	11332465001
	Notes: Store at room temperature			

ID	Buffer	Ingredients	Supplier	cat #
2.2.10	M-Wash Buffer	M-Wash Buffer (concentrate)	Zymo Research	part of D5030
		Ethanol (add as indicated on the bottle)	Fisher Scientific	64-17-5
	Notes: Store at room temperature			
2.2.11	PCR Master Mix	H × 25 µl of KAPA HiFi HotStart Uracil+ ReadyMix	Roche Sequencing	7959052001 (KAPA cat #: KK2801)
		H × 1.5 µl of IDT xGen Primers (20 µM)	Integrated DNA Technologies	1077675
	Notes: Prepare just before use. Mix thoroughly and store in ice until ready to use. H = number of capture reactions. See 2.5.1 for the primer sequences.			
2.2.12	Sequencing Buffer	10 mM Tris-HCl, pH 8	ThermoFisher Scientific	15568025
		0.1% Tween-20	Sigma-Aldrich	11332465001
	Notes: Store at room temperature			

2.3. Equipment and Consumables

ID	Item	Name	Supplier	cat #	Notes
2.3.1	Multichannel pipettes	Pipet-Lite Multi Pipette L8–10XLS+; L8–20XLS +; L8–200XLS+	Rainin		or equivalent
2.3.2	Water bath incubator	Precision 260 Circulating Water Bath	Thermo Fisher Scientific		or equivalent
2.3.3	Thermomixer	Eppendorf ThermoMixer	Eppendorf		or equivalent
2.3.4	Fluorometer	Qubit Fluorometer	ThermoFisher Scientific		any version
2.3.5	Tubes	Qubit Assay Tubes	ThermoFisher Scientific	Q32856	
2.3.6	Sonicator	Bioruptor Pico	Diagenode	B01060010	See Note 9.1 for alternatives
2.3.7	Tubes	Bioruptor Pico 0.2 ml microtubes	Diagenode	C30010020	
2.3.8	Thermocycler with heated lid	T100 Thermal Cycler	BioRad	1861096EDU	or equivalent
2.3.9	PCR tubes	MAXYMum Recovery™ PCR Tubes 0.2 ml	Corning	PCR-02D-L-C	or equivalent
2.3.10	Low-bind 1.7 ml tubes	DNA LoBind Tubes 1.5 ml	Eppendorf	22431021	or equivalent
2.3.11	Minicentrifuge tubes	Posi-Click tube, 1.7ml natural color	Denville Scientific	C2170	or equivalent
2.3.12	Mini centrifuge	MyFuge™ 12 mini centrifuge	Benchmark Scientific	C1012	or equivalent
2.3.13	Benchtop Centrifuge	Centrifuge 5425	Eppendorf	5405000042	or equivalent
2.3.14	Vacuum concentrator	Vacufuge Concentrator 5301	Eppendorf	022820001	or equivalent. See Note 9.5 for alternative methods to concentrate the samples.

ID	Item	Name	Supplier	cat #	Notes
2.3.15	Magnetic rack for PCR tubes	DynaMag™-96 Side Magnet PCR Strip MagStand	ThermoFisher Scientific Zymo Research	12331D 3DP-1002	or equivalent or equivalent
2.3.16	Magnetic rack for 1.5 ml tubes	DynaMag™-2 Magnet	ThermoFisher Scientific	12321D	or equivalent
2.3.17	Vortexer	Fisherbrand™ Analog Vortex Mixer	Fisher Scientific	02-215-414	or equivalent

2.4. Software Packages

All listed software packages for data processing are freely available under non-restrictive licensing. All data processing commands are carried out in a Unix like environment (Linux, MacOS, Windows Linux Subsystem).

2.4.1. Unix-like shell (Linux, MacOS Terminal or Windows Linux Subsystem)

2.4.2. Python 3 (>=3.6) (available at python.org)

2.4.3. Python packages

2.4.3.1. BSBolt (1.3.0) (available at https://bsbolt.readthedocs.io/en/latest/)

2.4.3.2. SciKit-Learn (>=0.22.2)

2.4.3.3. Numpy (>=1.18.1)

2.4.3.4. Matplotlib (>=3.1.2)

2.4.3.5. Scipy (>=1.3.3)

2.4.3.6. Pandas (>=1.0.0)

2.4.3.7. Jupyter (>=2.1.0)

2.4.4. samtools (>= 1.2) (available at htslib.org)

2.4.5. Cutadapt (>=2.4) (available at cutadapt.readthedocs.io)

2.4.6. FastQC (>=11.9) (available at https://www.bioinformatics.babraham.ac.uk/projects/fastqc/)

2.5. Primers/Oligonucleotides/Adapters

2.5.1. PCR Primers: xGen Library Amplification Primer Mix (20 μM), Integrated DNA Technologies, cat # 1077675.

Alternatively order the following oligos:

Name	Sequence	Notes
Fw	5’-AATGATACGGCGACCACCGAGAT-3’	HPLC purification
Rev	5’-CAAGCAGAAGACGGCATACGA-3’	HPLC purification

Resuspend each oligo to 100 μM with 10 mM Tris-HCl (pH 8) + 0.1 mM EDTA. Store the stock solution at −20°C. In a new tube add 20 μl of each oligo and 60 μl of 10 mM Tris-HCl (pH 8) + 0.1 mM EDTA. Store the working solution at −20°C.

2.5.2. Adapters: pre-methylated unique dual-indexed DNA Adapters synthesized by Integrated DNA Technologies. Order through the "Custom NGS Adapter Configuration Tool" (https://www.idtdna.com/pages/products/next-generation-sequencing/adapters/custom-ngs-adapters; https://www.idtdna.com/site/order/adapter). The sequences are available in Supplementary Table 1. See Note 9.2

2.5.3. Probes: Biotinylated RNA baits were synthesized by Arbor Biosciences (see Section 3. Experimental Design). The position of the targeted regions is available in Supplementary Table 2.

3. Experimental Design

Before starting a Targeted Bisulfite Sequencing Experiment, biotinylated probes capturing the fragments of interest need to be designed and synthesized. Genomic regions were selected for inclusion in the probe panel if the region has been previously identified as relevant to a health outcome, indicative of a specific cell type, or the region has high inter/intra-tissue variance in the proportion of methylated reads to total reads across individuals. CpG sites used in published DNA methylation-based predictive models for traits such as age [10,15], type 2 diabetes risk [18], smoking status [19], and more [34–36] were included in the panel.

Cell types have specific epigenomic arrangements, including genome wide DNA methylation patterns, that reflect their developmental lineage [37]. As a result, population based studies that rely on DNA methylation sequencing data have the potential to be confounded by the cell type composition of individual samples [38]. Cell type composition can be estimated using regression techniques [39] that rely on cell type specific methylation patterns. Cell type specific methylation patterns can be identified using publicly available methylation profiles of purified cell types [39,40]. Cell type specific regions for blood cell types, including macrophages, neutrophils, B cells, CD4þT cells, CD8þT cells, NK cells were included in the panel.

To enrich for dynamic CpG sites, sites that are variable within tissue types and/or variable across tissue types were also included in the panel. To select CpG sites that show within tissue variability we use data from a 2013 study by Ziller et al. [41] that examined the dynamic nature of CpG sites for 30 human tissues. We rank regions by the mean methylation difference observed between individuals in the same cell type. To select variable sites across tissue we calculated the variability in methylation values taken from WGBS across 37 human tissue types [40,41], and ranked sites according to the calculated variance.

Probe panels can be designed using various criteria and for different species. The number of designed probes can be variable, and it can be modified depending on specific needs. There are, however, limitations based on the number of probes:

- Few probes (< 5’000): the amount of the recovered material after hybridization is usually very little and it will require more PCR cycles to obtain a sufficient concentration for NGS-sequencing. In our experience panels targeting less than 250 Kb will result in low complexity libraries and a higher rate of PCR duplicates that need to be filtered out before further analysis.
- Large probe set (>40’000): the cost of probe synthesis increases, and it will require a deeper sequencing coverage, making it less attractive for large-scale projects. In addition, due to the variability of capture efficiency for the different sites, the number of regions that have sufficient coverage might be a few percent.

We then suggest synthesizing panels targeting between 5’000 and 20’000 sites.

We discourage designing probes to capture repetitive regions or regions with high similarity to other parts of the genome since it will increase the number of off-target reads and decrease the coverage of other regions of interest. For the same reason, we suggest filtering out regions with high similarity to the mitochondrial DNA. We noticed an anti-correlation between the GC content and the sequencing coverage of the targeted regions likely due to the combination of both hybridization efficiency and bisulfite treatment. We therefore suggest designing tiling probes targeting regions of interest with a high GC content to counterbalance this phenomenon.

The number of probes is also related to the amount of DNA used as input. We recommend inputs between 250–500 ng for each sample. Decreasing the amount of DNA will reduce the complexity of the final library leading to a higher rate of PCR duplicates. Should the designed panel target less than 5’000 sites, we suggest increasing the amount of starting DNA to 500–1’000 ng per sample. If the panel targets more than 20’000 sites, we do not recommend decreasing the amount of starting material (250–500 ng) but reducing the number of samples pooled for the hybridization capture.

One drawback of this methodology is that the probe design relies on existing DNA methylation data. Unfortunately, only a few organisms, such as Homo sapiens presented here, have extensive and available DNA methylation data to support probe design. It is still possible to apply this method to non-model organisms or organisms that have very little DNA methylation data by using a probe panel capturing the ultra-conserved regions, such as myBaits Expert UCE panel (Arbor Biosciences).

4. DNA extraction and Quantification

Whole blood samples are collected from individuals via venipuncture into a standard lavender top EDTA tube. DNA extraction is performed using the Qiagen QIAmp Midi-Prep kit with 1–2mL of blood according to the manufacturer’s instructions. The extraction protocol will take approximately 60 minutes to complete. Other tissues can be used as input material for genomic DNA extraction. Virtually all the DNA extraction kits or protocols appropriate for the tissue of interest can be used. We also recommend performing a column purification if the DNA extraction method involves the use of phenol since we saw variability in the purity of the DNA depending on the user that executed the extraction.

4.1. Add 200 μl of QIAGEN Protease at the bottom of a 15 ml centrifuge tube;

4.2. Add 1–2 ml of blood and mix briefly. Bring the volume to 2 ml using PBS buffer;

4.3. Add 2.4 ml of Buffer AL and mix thoroughly by inversion and vortexing for 1 minute;

4.4. Incubate for 10 min at 70°C in a water bath;

4.5. Add 2 ml of ethanol (96–100%) to the sample and mix thoroughly by inversion and vortexing for 1 minute;

4.6. Load half of the solution into a QIAamp Midi column placed in a 15 ml tube;

4.7. Close the cap and centrifuge at 1850 g for 3 minutes. Discard the filtrate;

4.8. Repeat steps 4.6 and 4.7 with the rest of the sample;

4.9. Add 2 ml of Buffer AW1 (see 2.2.1) to the column and centrifuge for 1 minute at 4500 g;

4.10. Without discarding the flow-through, add 2 ml of Buffer AW2 (see 2.2.2) to the column and centrifuge for 15 minutes at 4500 g;

4.11. Place the column into a new 15 ml tube and pipet 330 μl of AE Buffer and incubate at RT for 5 minutes;

4.12. Centrifuge for 2 minutes at 4500 g;

4.13. Reload the eluate onto the column and incubate at RT for 5 minutes;

4.14. Centrifuge for 2 minutes at 4500 g;

4.15. Purified DNA is measured with NanoDrop to assess the presence of contaminants. Acceptable values are $1.7 < A_{260/280} < 2$ and $1.8 < A_{260/230} < 2.3$. If values are outside the range, please proceed with further purification to remove the presence of RNA (if $A_{260/280} > 2$), proteins ($A_{260/280} < 1.7$) or other organic molecules ($A_{260/230} < 1.8$).

4.16. 2 μl of purified DNA are further quantified using Qubit dsDNA BR according to manufacturer’s recommendations to accurately measure the concentration.

5. Library Preparation

Purified DNA is fragmented to an average of 250–300 bp using a BioRuptor Pico sonication device (2.3.6). We suggest a starting amount of 500 ng of purified genomic DNA for each sample, however as low as 250 ng have been successfully processed with this protocol (quantification using a fluorimetric method). Fragmented genomic DNA is then End Repaired (blunting, 5’-phosphorylation and 3’-dephosphorylation), A-tailed (3’- dA overhang) and ligated to Y-shaped 3'-dT-overhanged pre-methylated adapters (Figure 2). The entire procedure will take approximately 2–2:30 hours. The addition of unmethylated lambda genomic DNA is used to check for bisulfite conversion efficiency.

5.1. Turn on the Bioruptor main switch and the Water Cooler at least 30 minutes before the sonication step;

5.2. Transfer 500 ng of purified DNA supplemented with 1 ng of unmethylated lambda genomic DNA into a Bioruptor 0.2 ml tube in a final volume of 50 μl. Add EB Buffer (see 2.2.3) if needed;

5.3. Sonicate for 15 cycles of 30 seconds ON, 30 seconds OFF;

5.4. Optional: run 2 μl of fragmented DNA on a High-Sensitivity D1000 ScreenTape. Fill to 50 μl with EB Buffer;

5.5. Transfer the sonicated DNA into 0.2 ml PCR tubes;

5.6. Add 10 μl of EndPrep Master Mix (see 2.2.4) to each sample, cap the tubes, briefly vortex, spin-down and incubate in a thermocycler as follows: 20°C for 30 minutes; 65°C for 30 minutes; hold at 4°C. Set the lid to 70°C. The total volume is 60 μl;

5.7. Add 2.5 μl of 15 μM pre-methylated Adapters (see Note 9.2) and mix. Each sample will have a different barcoded Adapter.

5.8. Add 31 μl of Ligation Master Mix (see 2.2.5) to each sample.

5.9. Incubate the sample in a thermocycler for 20 minutes at 20°C (lid off), hold at 4°C. The total volume is 93.5 μl.

5.10. Proceed immediately to SPRI-beads Purification

5.10.1. Make sure the Purification Beads are fully resuspended and at room temperature before starting the procedure (Note 9.3). Beads can be resuspended by vortexing until the color is homogeneous and there is no visible bead pellet.

5.10.2. Add 80 μl of the resuspended Purification Beads (0.85 volumes) to each sample;

5.10.3. Mix by vortexing and incubate for 10 minutes at Room Temperature;

5.10.4. Spin-down to collect the solution at the bottom of the tube and place the tubes on a magnet until separation occurs (approximately 5–10 minutes). The tubes remain on the magnet until step 5.10.10 included;

5.10.5. Keeping the tubes in the magnet, remove most of the supernatant (160–165 μl) and discard it;

5.10.6. Add 200–250 μl of freshly prepared 80% Ethanol solution to each sample and incubate at Room Temperature for 30 seconds;

5.10.7. Remove the supernatant with a pipette or a vacuum aspiration system;

5.10.8. Repeat steps 5.10.6 and 5.10.7 an additional time;

5.10.9. Incubate the open tubes for 2 minutes at Room Temperature, then remove any additional Ethanol;

5.10.10. Air-dry the tubes for additional 3–5 minutes (or until the beads are dry);

5.10.11. Remove the tubes from the magnet and add 17 μl of EB buffer to each tube. Resuspend the beads either by pipetting or by vortexing;

5.10.12. Incubate 3 minutes at Room Temperature;

5.10.13. Spin-down briefly and place it back on the magnet until separation occurs (1–2 minutes);

5.10.14. Transfer the supernatant (15 μl) of the samples to be pooled into a new 1.5 ml tube (see Note 9.4).

5.10.15. The purified DNA Library can be stored at −20°C until ready to proceed.

6. Target Enrichment through Hybridization Capture with RNA Probes

Pooled libraries are then vacuum concentrated (alternatively, see Note 9.5) and incubated with blockers against genome repetitive elements and the adapter portion of the libraries to prevent spurious hybridization. Blocked DNA Libraries and RNA biotinylated probes are then hybridized in solution and the complexes are captured with streptavidin magnetic beads (Figure 2). The entire procedure takes approximately 20–24 hours, including the overnight hybridization.

6.1. Dry each pool of samples using a Vacuum Concentrator (see 2.3.14 and Note 9.5). Dried DNA can be stored at Room Temperature overnight or at −20°C for longer periods.

6.2. Add 7 μl of nuclease-free H_2O to the tube of dried Library and vortex for 30 seconds. Add 5 μl of Blockers Mix (see 2.2.6) and incubate at 60°C for at least 15 minutes in a ThermoMixer at 1400 rpm to fully resuspend the Library.

6.3. During the incubation, prepare the Hybridization Mix (see 2.2.7);

6.4. Incubate the tube containing the Hybridization Mix at 60°C for 10 minutes in a ThermoMixer;

6.5. Cool down the Hybridization Mix tube at Room Temperature for 5 minutes;

6.6. Transfer the Blockers-Library mix (12 μl) into a new 0.2 ml PCR tube for each capture reaction. Label the tube(s) as LIB.

6.7. Incubate the LIB tube(s) in a thermocycler at 95°C for 5 minutes; hold at the Hybridization Temperature (65°C). Set the lid to 105°C and the volume to 18 µl.

6.8. While the LIB tube(s) are incubating, transfer 18.5 µl of the Hybridization Mix into a new 0.2 ml tube for each capture reaction. Label the tube(s) as HYB.

6.9. Once the temperature reaches the Hybridization Temperature (65°C), place the HYB tube(s) in the Thermocycler and incubate both LIB and HYB for at least 5 minutes at Hybridization Temperature;

6.10. With the tubes still in the thermocycler, transfer 18 µl from the HYB tube(s) into the correspondent LIB tube(s) using a multichannel pipette.

6.11. Discard the HYB tube(s) and incubate the LIB+HYB tube(s) at the hybridization temperature (65°C) for 16–20 hours. Make sure the lid temperature is above 80°C. To avoid lower on-target read proportion it is important that the temperature of the capture reaction (LIB+HYB) doesn't drop below the hybridization temperature.

6.12. Before the end of the hybridization incubation, warm the Wash Buffer X (see 2.2.8) at the Hybridization Temperature (65°C) for 15–30 minutes;

6.13. While the Wash Buffer X is incubating, bring the Bead Binding Buffer at Room Temperature and prepare the Streptavidin Beads;

6.14. Aliquot 30 µl for each capture reaction of the resuspended beads to a 1.7 ml DNA LoBind tube (*no more than 7 reactions*);

6.15. Put the beads-containing tube on the magnet (compatible with 1.5/2 ml tubes) and incubate for 2 minutes;

6.16. Discard the supernatant and resuspend the beads in 200 µl x #reactions of Bead Binding Buffer (*for 7 reactions, wash with 200 µl x7=1400 µl*);

6.17. Repeat steps 6.15 and 6.16 two additional times, for a total of three washes;

6.18. Once the final supernatant has been removed, resuspend the beads in 70 µl x #reactions of Bead Binding Buffer (*for 7 reactions, resuspend with 70 µl x7= 490 µl*);

6.19. Aliquot 70 µl the Beads/Binding Buffer mix into new LoBind tubes (one per capture);

6.20. Warm the beads-containing tubes at the hybridization temperature (65°C) for 2 to 5 minutes;

6.21. Transfer each capture reaction (LIB+HYB) into a beads-containing tube;

6.22. Incubate the tubes in a Thermomixer at 65°C for 30 min (600 rpm). Flick and briefly spin-down every 10 minutes;

6.23. Transfer the tubes to the magnet and incubate at RT until the supernatant is clear. Remove the supernatant;

6.24. Resuspend the beads in 375 µl of 65°C-warm Wash Buffer X;

6.25. Incubate in the Thermomixer at 65°C for 10 minutes (600 rpm);

6.26. Transfer the tube on a magnet and remove the supernatant when clear;

6.27. Repeat steps 6.24, 6.25 and 6.26 two more times, for a total of three washes;

6.28. Remove carefully all the residual Wash Buffer X from the tube;

6.29. Resuspend the beads in 23 μl of SBE Buffer (see 2.2.9)

6.30. Transfer the mixture into a 0.2 ml PCR tube and incubate the suspension at 95°C for 5 minutes;

6.31. Immediately spin-down and place the tube(s) on the magnet (compatible with 0.2 ml tubes);

6.32. Transfer the supernatant (20 μl) into a new 0.2 ml tube.

6.33. Store the captured DNA at −20°C or proceed directly to the Bisulfite Conversion.

7. Bisulfite Conversion, Library Amplification and Sequencing

Perform Bisulfite Conversion using the Zymo Lightning Methylation kit. De-sulphonated libraries are then amplified and subject to Quality Control before Sequencing Submission (Figure 2). The entire procedure will take approximately 3–4 hours.

7.1. Add 130 μl of Lightning Conversion Reagent (RT) to 20 μl of the captured DNA;

7.2. Mix and incubate the tubes in a Thermocycler with the following settings: 98°C for 8 minutes; 54°C for 30 minutes; 98°C for 3 minutes; 54°C for 30 minutes; hold at 4°C.

7.3. Once the incubation is over, add 600 μl of M-Binding Buffer to a Zymo Spin-IC column/Collection Tube;

7.4. Load the samples in the Binding Buffer-containing Column, cap the tubes and invert several times;

7.5. Centrifuge 30 seconds at 11’000 g. Discard the flow-through;

7.6. Add 100 μl of M-Wash Buffer (see 2.2.10) to the column and centrifuge as in Step 7.5;

7.7. Add 200 μl of L-Desulphonation Buffer and incubate at RT for 15–18 minutes;

7.8. Centrifuge 30 seconds at 11’000 g;

7.9. Add 200 μl of M-Wash Buffer and centrifuge as in Step 7.8;

7.10. Repeat the wash step in 7.9;

7.11. Discard the flow-through and centrifuge for 1 minute at 11’000 g;

7.12. Transfer the column to a new 1.5 ml tube and discard the Collection Tube;

7.13. Add 25 µl of warm (60–70°C) M-Elution Buffer to the center of the column;

7.14. Incubate at RT for 2 minutes and centrifuge 1 minute at 11’000 g.

7.15. Reload the eluate at the center of the column and centrifuge for 1 minute at 11’000 g;

7.16. Proceed to the PCR Step or store the samples at −20°C.

7.17. Transfer 23.5 µl of the Bisulfite-treated captured DNA Library into a new 0.2 ml PCR tube;

7.18. Add 26.5 µl of the PCR Master Mix (see 2.2.11) to each tube and incubate the samples in the Thermocycler as follows: 98°C for 2 minutes; 14 cycles of {98°C for 20 seconds, 60°C for 30 seconds, 72°C for 30 seconds}; Final Extension of 5 minutes at 72°C; hold at 6°C.

7.19. Proceed immediately to SPRI-beads Purification

7.19.1. Make sure the Purification Beads are fully resuspended and at room temperature before starting the procedure (Note 9.3). Beads can be resuspended by vortexing until the color is homogeneous and there is no visible bead pellet.

7.19.2. Add 45 µl of the resuspended Purification Beads (0.9 volumes) to each sample;

7.19.3. Mix by vortexing and incubate for 5 minutes at Room Temperature;

7.19.4. Spin-down to collect the solution at the bottom of the tube and place the tubes on a magnet until separation occurs (usually less than 5 minutes). The tubes remain on the magnet until step 7.19.10 included;

7.19.5. Keeping the tubes in the magnet, remove most of the supernatant (85–90 µl) and discard it;

7.19.6. Add 200–250 µl of freshly prepared 80% Ethanol solution to each sample and incubate at Room Temperature for 30 seconds;

7.19.7. Remove the supernatant with a pipette or a vacuum aspiration system;

7.19.8. Repeat steps 7.19.6 and 7.19.7 an additional time;

7.19.9. Incubate the open tubes for 2 minutes at Room Temperature, then remove any additional Ethanol;

7.19.10. Air-dry the tubes for additional 5 minutes (or until the beads are dry);

7.19.11. Remove the tubes from the magnet and add 17 µl of EB buffer to each tube. Resuspend the beads either by pipetting or by vortexing;

7.19.12. Incubate 3 minutes at Room Temperature;

7.19.13. Spin-down briefly and place it back on the magnet until separation occurs (1–2 minutes);

7.19.14. Transfer the supernatant (15 μl) of the Bisulfite-converted and Enriched Final Library Pool into a new tube (either 0.2 ml or 1.5 ml).

7.20. 2 μl of the Final Library Pool are further quantified using Qubit dsDNA HS according to manufacturer's recommendations to accurately measure the concentration. Concentrations are usually around 4 ng/μl;

7.21. The average size of the Final Library Pool is measured using a D1000 HS dsDNA Tape on an Agilent 2200 TapeStation (see Note 9.6). The usual Average size is between 320 and 350 bp (Figure 3). If bands smaller than 150 bp are visible (Primer/Adapter Dimers), please bring the volume of the Final Library Pool to 50 μl with EB Buffer and repeat the last SPRI beads Purification (see 7.19).

7.22. The Final Library is then diluted to the desired concentration for sequencing with Sequencing Buffer (see 2.2.12). To calculate the molarity, use the formula:

$$\left(\text{concentration}\,(\text{ng}/\mu\text{l}) \times 10^{6}\right) / \left(660\text{g}/\text{mol} \times \text{average size}\,(\text{bp})\right) = \text{molarity}\,(\text{nM})$$

7.23. See Note 9.7 for a discussion about sequencing options.

8. Data Analysis

Processing Bisulfite Sequencing data generally occurs in two distinct stages (Figure 4). First individual samples are mapped to a reference sequence using a Bisulfite-aware alignment tool, duplicate alignments are marked, and methylation values are called from the resulting alignments. Reads that are discarded during this step can be classified into two categories (Supplementary Figure 1, Supplementary Figure 2 and Supplementary Table 4):

- Reads that are off-target: regions of the genome that are captured even if not part of the original design. This phenomenon can be due to the high similarity between two sequences in the genome and the partial specificity of the hybridization capture step. We expect to have a few regions with high coverage (first category) and many regions with very low coverage (second category).
- Reads derived from PCR amplification and identified as "duplicated reads" according to samtools markdup.

Second, methylation values from several samples are aggregated together into a combined methylation matrix and missing values are imputed, producing an analysis ready methylation matrix. In addition, we highlight one downstream application of the methylation matrix by fitting a cross validated epigenetic clock. All code for the full processing pipeline can be found at GitHub (https://github.com/NuttyLogic/MethodsTBS). The Supplementary Data Analysis document is a copy of Section 8 - Data Analysis with the corresponding code.

8.1. Pipeline Setup

8.1.1. The bisulfite sequencing processing pipeline begins with sequence reads stored as demultiplexed FASTQ files [42–44], text-based files that store read base calls along with their corresponding quality scores. Paired-end and single-end sequencing experiments have two and one FASTQ files per sample, respectively.

8.1.2. The generation of an alignment index requires a reference sequence in FASTA format. The example alignment index was generated from the UCSC hg38 reference genome (https://genome.ucsc.edu/cgi-bin/hgGateway).

8.1.3. All software should be installed prior to pulling data through the pipeline. Python3 should be installed before installing other required tools. The installation process for all tools can be found by referencing the link listed in the required software (see Section 2.4). All third-party python packages can be installed using the python package manager (pip) after installing python.

8.1.4. All example code was run on the UCLA Hoffman2 cluster, running UGE (v8.6.4) with 8 requested cores and 32GB ram. The pipeline is scalable for sequential runs on one machine with lower resource requirements. All example commands are written for the listed target specification so commands should be adjusted accordingly.

8.2. Reads Pre-Processing, Genome Indexing, Reads Alignment and Methylation Calling

8.2.1. Processing of bisulfite sequencing data is carried out using BSBolt, a bisulfite sequencing processing platform, and samtools, a tool for manipulation of alignment files.

8.2.2. The alignment of targeted bisulfite sequence data first requires the creation of a bisulfite alignment reference. The alignment reference contains alignment information for both the Watson (sense) and Crick (anti-sense) strands, as after bisulfite conversion the strands are no longer complementary. For best performance of mapping targeted bisulfite sequencing data, we will create a restricted alignment index using the BSBolt Index command. A restricted alignment index masks region outside the targeted areas, forcing reads to align to within the target regions. While this approach will increase the number of reads mapping outside their region of origin it will also increase the depth for methylation calling which can improve performance of downstream analysis tools.

8.2.3. Adapter sequences may reduce alignment quality and negatively affect downstream analysis, so it is good practice to remove adapter sequences and low-quality base calls before alignment. We recommend cutadapt [42,45] to quickly, and efficiently remove adapter sequences.

8.2.4. Trimmed reads are aligned to the reference genome using the BSBolt Align function. Aligning bisulfite sequencing works by aligning reads to both bisulfite reference strands and picking the best possible alignment. If a read has equally good alignments to both reference strands the read is reported as bisulfite ambiguous with low mapping quality. Alignment of the FASTQ files produces a compressed alignment file (BAM) in the SAM format [43,46,47].

8.2.4.1. An important consideration when mapping bisulfite sequencing data is directionality of the library. In directional bisulfite sequencing libraries only the original bisulfite converted DNA is sequenced; in un-directional libraries the PCR product of the bisulfite converted DNA is also sequenced. Due to the asymmetric nature of bisulfite sequencing alignment, aligning an un-directional library can take roughly double the time for the same number of reads. Un-directional alignment will work for directional libraries, but not vice versa.

8.2.4.2. The alignment file contains all the reads aligned to the Watson and Crick bisulfite reference strands on the positive and negative strands respectively. Each alignment has a SAM tag indicating the alignment reference strand and a SAM tag indicating if the alignment is bisulfite unique.

8.2.5. Following alignment paired reads are fixed, the alignment file is coordinate-sorted, and duplicate reads are marked using samtools. Duplicate reads are PCR duplicates that have the same alignment position for both reads in a pair. Removing duplicates helps improve methylation calling and downstream analysis. See Supplementary Figure 1 and 2 for the visualization of mapping stats and coverage of the target sites before and after duplicate reads removal.

8.2.6. Methylation values can be called for all cytosines using BSBolt CallMethylation. Methylation values for reference cytosines and guanine are called using read alignments on the Watson and Crick strands respectively. Methylation calling is also context-specific. Often, only methylation values from CpG sites are used for downstream analysis in mammalian genomes. Methylation calling can thus omit non-CpG sites, saving computation time. The resulting file is output as a compressed, unsorted CGmap file. CGmap format is described in Table 4.

8.3. Methylation Matrix Assembly

8.3.1. After samples are processed individually, that data is aggregated together into a combined methylation matrix (CGmatrix file, Table 6). The matrix is constructed using methylation sites that are consistently covered at the desired depth in a set proportion of the processed samples. The matrix is assembled using the BSBolt AggregateMatrix command by passing a list of files, sample names, the minimum coverage threshold, and the proportion of samples a site must be observed in at the coverage threshold. Additionally, there are options to include only CpG sites and increase the number of processing threads to decrease processing time.

8.3.2. OPTIONAL: During matrix assembly, sites that were not covered at the required threshold in individual samples are reported as null. If desired, methylation values for these null sites can be imputed using BSBolt Impute. This function uses a kNN sliding window to leverage the local methylation structure to select the nearest neighbors. The average value of k nearest neighbors is used to impute the missing values. With targeted bisulfite sequencing data, the window size can be set proportional to the size of the capture regions using the BSBolt Impute -*W* option.

8.4. Fitting an Epigenetic Clock

8.4.1. With the appropriate phenotype data, DNA methylation-based prediction models can be fit to numerous traits. The most well-known of these models predict age and are termed "epigenetic clocks" [10,15]. Here, we highlight the process for fitting a leave-one-out cross-validated epigenetic clock using a penalized regression-based approach. When fitting a leave- one-out model a separate model is fit for each sample with that sample held out from model training. This leverages the maximal amount of data for model training while providing the least biased estimate of epigenetic age for the sample held out.

8.4.2. All analysis is conducted inside a Jupyter notebook (https://jupyter.org/) [48], an interactive programming environment that streamlines data analysis and offers support for the Python and R programming languages. Jupyter notebooks are launched from the command line by running *jupyter lab* in a Unix terminal. This will launch a web-based widow browser where a user can launch a jupyter notebook. Inside a jupyter notebook code is executed in distinct cells. All cells are run independently, so if code fails in one cell and throws an error it doesn't affect other cells. Variables created in other cells are accessible in all other cells after they are created.

8.4.3. The analysis begins by importing the necessary libraries to conduct interactive data analysis. All packages can be installed through the python package index.

8.4.3.1. SciKit-Learn (>=0.22.2)

8.4.3.2. Matplotlib (>=3.1.2)

8.4.3.3. Scipy (>=1.3.3)

8.4.3.4. Pandas (>=1.0.0)

8.4.4. The methylation matrix can be imported using the Pandas package by calling the pandas.read_csv function. After importing the methylation matrix any sites with null values should be dropped from further analysis, as null sites will interfere with downstream tools. Additionally, if individual samples have an excess of missing data, the samples should be removed from analysis at this step. Phenotype data stored as .csv, .txt, or .tsv files, or gzipped files of these types, can be loaded using the same process.

8.4.4.1. Phenotype data can generally be imported using the same process. In the example notebook ages are provided with the jupyter notebook itself.

8.4.5. Before proceeding to model fitting, we perform a quick quality control check by principal component analysis on the methylation matrix and plotting the first two principal components (PCs). Samples with technical problems like incomplete bisulfite conversion or inefficient capture will be apparent as outliers in the PCA plot due to wide-scale methylation changes. Note, when interpreting the PCA plot it is important to consider the proportion of variation explained by the individual PCs. A sample may appear to be an outlier, but if the proportion of variation is low the sample can be used for downstream analysis. Additionally, if there is a systemic problem that affects all samples in an experiment this will not be apparent using PCA as PCA captures relative difference within the methylation matrix.

8.4.5.1. In the example analysis, the first and second PCs explain little-observed variation and there are no clear outliers. All samples were kept for fitting an epigenetic clock.

8.4.6. Fitting an epigenetic clock is accomplished using penalized regression, and more specifically elastic net regression. In a penalized regression coefficient are shrinked towards zero, this results in site selection of relevant CpG sites while fitting a model to predict a trait of interest.

8.4.6.1. In the example, we utilized scikit-learn ElasticNet regression [49]. When fitting the model in the sample we favored the lasso penalty relative to the ridge penalty, the ratio of lasso to ridge penalties can be adjusted using the l1_ratio option.

8.4.6.2. The example dataset is limited by the number of samples (n=48), in this context it is best to fit a leave-one-out cross-validated model. In other words, 48 separate regression models will be fit for each sample. Forty-seven separate samples will be used for model training with one sample removed for model testing. The collection of models will be evaluated to determine the overall fit for the collection.

8.4.6.3. Leave-one-out regression is easily accomplished using a for loop. In the example each sample selected within the loop is held as the testing sample, the other samples are used for model training. After the model is trained an age prediction is made for the sample held out of model training.

8.4.7. Following model training, we evaluate the model by fitting a trendline of the predicted epigenetic ages against chronological age. The trendline is fit in the example using scipy.optimize [50] and results plotted using matplotlib [51] (Figure 5).

9. Notes

9.1. The DNA can be fragmented with other types of sonicator. Please adjust the conditions to obtain fragments in the 200–400 bp range. However, enzymatic DNA fragmentation using several commercially available kits resulted in a global loss of DNA methylation using bisulfite-sequencing when compared to the sonication method, which is considered the gold standard.

9.2. We use unique dual-index pre-methylated adapters synthesized from Integrated DNA Technologies (IDT - Coralville, Iowa, USA). The complete list for 96 adapters can be found in Supplementary Table 1. Alternatively, it's also possible to use pre-methylated adapters commercially available, such as:

9.2.1. Illumina TruSeq single indexes (Set A: 20015960 or Set B: 20015961);

9.2.2. Roche SeqCap Adapters (Set A: 07141530001 or Set B: 07141548001);

9.2.3. Diagenode Premium WGBS Indexes (cat # C05010032, C05010033 or C05010034);

9.2.4. Perkin-Elmer NEXTFLEX Bisulfite-seq Barcodes (cat # NOVA-511911, NOVA-511912 or NOVA-511913).

We strongly recommend the use of unique-dual indices to mitigate the index hopping effect. Moreover, the relatively small number of reads required for each sample allows for high multiplexing, much more than the 24 pre-methylated adapters available commercially. Our pipeline removes PCR duplicates in silico, but another option would be to add UMIs (or Unique Molecular Identifiers) in the adapter design.

9.3. Purification Beads: there are numerous variants for the Purification Beads. We use the NEBNext Purification Beads that come with the kit (NEB, cat # E7103S/L). Please note that the beads are not present in the NEB kits E7645S/L. Alternatives are:

9.3.1. SPRIselect Reagent, cat # B23317, B23318 or B23319 (Beckman Coulter);

9.3.2. AMPure XP Beads, cat # A63880, A63881 or A63882 (Beckman Coulter) - stored at 4°C. Warm-up at Room Temperature Before use;

9.3.3. KAPA HyperPure Beads, cat # KK8007, KK8008, KK8009, KK8010 or KK8011 (Roche - Sequencing) - stored at 4°C. Warm-up at Room Temperature Before use;

9.3.4. Sera-Mag Select, cat # 29343045, 29343052 or 29343057 (GE Healthcare - Life Sciences) - stored at 4°C. Warm-up at Room Temperature Before use;

9.3.5. Home-made Purification Beads: Sera-Mag SpeedBead Carboxylate-Modified Magnetic Particles - Hydrophobic, cat # 65152105050250 (GE Healthcare - Life Sciences) [52].

9.4. We usually pool between 12 to 16 samples with different barcodes before the Target Hybridization (Enrichment) step. Other pooling strategies may be possible, but they need to be tested.

9.5. Alternatives to Vacuum concentration:

9.5.1. Column Purification using Zymo DNA Clean & Concentrator-5 columns (Zymo Research, cat # D4013).

9.5.1.1. Add 7 volumes of DNA Binding Buffer and mix by pipetting;

9.5.1.2. Load 750 µl of the mixture into a Zymo Spin column;

9.5.1.3. Incubate for 1 minute at Room Temperature, then centrifuge for 30 seconds at 11’000g. Discard the flow-through;

9.5.1.4. Repeat steps 9.5.1.2 and 9.5.1.3 for the remaining solution;

9.5.1.5. Add 200 µl of DNA Wash Buffer to the column;

9.5.1.6. Centrifuge for 30 seconds at 11’000g. Discard the flow-through;

9.5.1.7. Repeat steps 9.5.1.5 and 9.5.1.6 an additional time;

9.5.1.8. Centrifuge for 1 minute at 11’000g to completely remove traces of ethanol.

9.5.1.9. Discard the 2 ml collection tube and transfer the column into a new 1.5 ml tube;

9.5.1.10. Add 10 µl of warm (60–70°C) EB Buffer in the center of the column and incubate for 1 minute;

9.5.1.11. Centrifuge for 30 seconds at 11’000 g;

9.5.1.12. Reload the eluate at the center of the column and centrifuge for 1 minute at 11’000 g.

9.5.2. SPRI Beads Purification. The Purification is performed as described in 5.10 with a few modifications. Use 1.2x volumes of beads for step 5.10.2 (120 µl of beads every 100 µl of DNA solution). Resuspend the beads in 10 µl of EB buffer (Step 5.10.11) and recover 8 µl of the supernatant (Step 5.10.14).

9.6. Alternatives for QC:

9.6.1. Average Size: 2100 Agilent Bioanalyzer, Agilent 4200 TapeStation, Agilent 4150 TapeStation, PerkinElmer LabChip GX, Bio-Rad Experion or similar.

9.6.2. Concentration: qPCR-based Quantification kits are commercially available from Roche Sequencing, NEB, Takara, ThermoFisher Scientific, and more or the home-made protocol available from the Illumina website [53].

9.7. Each individual sample requires about 5–8 M reads for about 15’000 targeted regions. Additional testing is necessary for different panels. We usually sequence around 48 samples (3 pools of 16 samples) per NovaSeq 6000 SP lane in PE150 mode. It’s important not to have overlapping barcodes to avoid reads misassignment. Due to the lower complexity of Bisulfite Libraries, we spike-in Illumina PhiX Control at 10%. Usually we recover around 300–350 M reads from our bisulfite-treated libraries from an SP lane. Other sequencing options are valid as long as each library is sequenced deeply enough.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Funding

This work was supported by the National Institutes of Health (T32CA201160 to C.F.). M.M. is partially supported by the UCLA QCBio Collaboratory Fellowship (https://qcb.ucla.edu/collaboratory-2/qcb-postdoctoral-fellows/).

17. References

[1]. Blow MJ, Clark TA, Daum CG, Deutschbauer AM, Fomenkov A, Fries R, Froula J, Kang DD, Malmstrom RR, Morgan RD, Posfai J, Singh K, Visel A, Wetmore K, Zhao Z, Rubin EM, Korlach J, Pennacchio LA, Roberts RJ, The Epigenomic Landscape of Prokaryotes, PLoS Genet. 12 (2016) e1005854. [PubMed: 26870957]

[2]. Schmitz RJ, Lewis ZA, Goll MG, DNA methylation: Shared and Divergent Features across Eukaryotes, Trends Genet. 35 (2019) 818–827. [PubMed: 31399242]

[3]. de Mendoza A, Lister R, Bogdanovic O, Evolution of DNA Methylome Diversity in Eukaryotes, J. Mol. Biol (2019). 10.1016/j.jmb.2019.11.003.

[4]. Zhong X, Comparative epigenomics: a powerful tool to understand the evolution of DNA methylation, New Phytol. 210 (2016) 76–80. [PubMed: 26137858]

[5]. Gowher H, Jeltsch A, Mammalian DNA methyltransferases: new discoveries and open questions, Biochem. Soc. Trans 46 (2018) 1191–1202. [PubMed: 30154093]

[6]. Lyko F, The DNA methyltransferase family: a versatile toolkit for epigenetic regulation, Nat. Rev. Genet 19 (2018) 81–92. [PubMed: 29033456]

[7]. Greenberg MVC, Bourc’his D, The diverse roles of DNA methylation in mammalian development and disease, Nat. Rev. Mol. Cell Biol 20 (2019) 590–607. [PubMed: 31399642]

[8]. Luo C, Hajkova P, Ecker JR, Dynamic DNA methylation: In the right place at the right time, Science. 361 (2018) 1336–1340. [PubMed: 30262495]

[9]. Jin Z, Liu Y, DNA methylation in human diseases, Genes Dis. 5 (2018) 1–8. [PubMed: 30258928]

[10]. Horvath S, DNA methylation age of human tissues and cell types, Genome Biology. 14 (2013) R115. 10.1186/gb-2013-14-10-r115. [PubMed: 24138928]

[11]. Horvath S, Ritz BR, Increased epigenetic age and granulocyte counts in the blood of Parkinson’s disease patients, Aging. 7 (2015) 1130–1142. [PubMed: 26655927]

[12]. Horvath S, Langfelder P, Kwak S, Aaronson J, Rosinski J, Vogt TF, Eszes M, Faull RLM, Curtis MA, Waldvogel HJ, Choi O-W, Tung S, Vinters HV, Coppola G, Yang XW, Huntington's disease accelerates epigenetic aging of human brain and disrupts DNA methylation levels, Aging. 8 (2016) 1485–1512. [PubMed: 27479945]

[13]. Levine ME, Lu AT, Bennett DA, Horvath S, Epigenetic age of the pre-frontal cortex is associated with neuritic plaques, amyloid load, and Alzheimer's disease related cognitive functioning, Aging. 7 (2015) 1198–1211. 10.18632/aging.100864. [PubMed: 26684672]

[14]. Horvath S, Levine AJ, HIV-1 Infection Accelerates Age According to the Epigenetic Clock, J. Infect. Dis 212 (2015) 1563–1573. [PubMed: 25969563]

[15]. Hannum G, Guinney J, Zhao L, Zhang L, Hughes G, Sadda S, Klotzle B, Bibikova M, Fan J-B, Gao Y, Deconde R, Chen M, Rajapakse I, Friend S, Ideker T, Zhang K, Genome-wide methylation profiles reveal quantitative views of human aging rates, Mol. Cell 49 (2013) 359–367. [PubMed: 23177740]

[16]. Weidner CI, Lin Q, Koch CM, Eisele L, Beier F, Ziegler P, Bauerschlag DO, Jöckel K-H, Erbel R, Mühleisen TW, Zenke M, Brümmendorf TH, Wagner W, Aging of blood can be tracked by DNA methylation changes at just three CpG sites, Genome Biol. 15 (2014) R24. [PubMed: 24490752]

[17]. Bell CG, Lowe R, Adams PD, Baccarelli AA, Beck S, Bell JT, Christensen BC, Gladyshev VN, Heijmans BT, Horvath S, Ideker T, Issa J-PJ, Kelsey KT, Marioni RE, Reik W, Relton CL, Schalkwyk LC, Teschendorff AE, Wagner W, Zhang K, Rakyan VK, DNA methylation aging clocks: challenges and recommendations, Genome Biol. 20 (2019) 249. [PubMed: 31767039]

[18]. Wahl S, Drong A, Lehne B, Loh M, Scott WR, Kunze S, Tsai P-C, Ried JS, Zhang W, Yang Y, Tan S, Fiorito G, Franke L, Guarrera S, Kasela S, Kriebel J, Richmond RC, Adamo M, Afzal U, Ala-Korpela M, Albetti B, Ammerpohl O, Apperley JF, Beekman M, Bertazzi PA, Black SL, Blancher C, Bonder M-J, Brosch M, Carstensen-Kirberg M, de Craen AJM, de Lusignan S, Dehghan A, Elkalaawy M, Fischer K, Franco OH, Gaunt TR, Hampe J, Hashemi M, Isaacs A, Jenkinson A, Jha S, Kato N, Krogh V, Laffan M, Meisinger C, Meitinger T, Mok ZY, Motta V, Ng HK, Nikolakopoulou Z, Nteliopoulos G, Panico S, Pervjakova N, Prokisch H, Rathmann W, Roden M, Rota F, Rozario MA, Sandling JK, Schafmayer C, Schramm K, Siebert R, Slagboom PE, Soininen P, Stolk L, Strauch K, Tai E-S, Tarantini L, Thorand B, Tigchelaar EF, Tumino R, Uitterlinden AG, van Duijn C, van Meurs JBJ, Vineis P, Wickremasinghe AR, Wijmenga C, Yang T-P, Yuan W, Zhernakova A, Batterham RL, Smith GD, Deloukas P, Heijmans BT, Herder C, Hofman A, Lindgren CM, Milani L, van der Harst P, Peters A, Illig T, Relton CL, Waldenberger M, Järvelin M-R, Bollati V, Soong R, Spector TD, Scott J, McCarthy MI, Elliott P, Bell JT, Matullo G, Gieger C, Kooner JS, Grallert H, Chambers JC, Epigenome-wide association study of body mass index, and the adverse outcomes of adiposity, Nature. 541 (2017) 81–86. [PubMed: 28002404]

[19]. Zhang Y, Schöttker B, Florath I, Stock C, Butterbach K, Holleczek B, Mons U, Brenner H, Smoking-Associated DNA methylation Biomarkers and Their Predictive Value for All-Cause and Cardiovascular Mortality, Environ. Health Perspect 124 (2016) 67–74. [PubMed: 26017925]

[20]. Orozco LD, Farrell C, Hale C, Rubbi L, Rinaldi A, Civelek M, Pan C, Lam L, Montoya D, Edillor C, Seldin M, Boehnke M, Mohlke KL, Jacobsen S, Kuusisto J, Laakso M, Lusis AJ, Pellegrini M, Epigenome-wide association in adipose tissue from the METSIM cohort, Hum. Mol. Genet 27 (2018) 2586. [PubMed: 29893869]

[21]. Horvath S, Raj K, DNA methylation-based biomarkers and the epigenetic clock theory of ageing, Nat. Rev. Genet 19 (2018) 371–384. [PubMed: 29643443]

[22]. Jylhävä J, Pedersen NL, Hägg S, Biological Age Predictors, EBioMedicine. 21 (2017) 29–36. [PubMed: 28396265]

[23]. Clark SJ, Statham A, Stirzaker C, Molloy PL, Frommer M, DNA methylation: bisulphite modification and analysis, Nat. Protoc 1 (2006) 2353–2364. [PubMed: 17406479]

[24]. Lister R, Pelizzola M, Dowen RH, Hawkins RD, Hon G, Tonti-Filippini J, Nery JR, Lee L, Ye Z, Ngo Q-M, Edsall L, Antosiewicz-Bourget J, Stewart R, Ruotti V, Millar AH, Thomson JA, Ren B, Ecker JR, Human DNA methylomes at base resolution show widespread epigenomic differences, Nature. 462 (2009) 315–322. [PubMed: 19829295]

[25]. Cokus SJ, Feng S, Zhang X, Chen Z, Merriman B, Haudenschild CD, Pradhan S, Nelson SF, Pellegrini M, Jacobsen SE, Shotgun bisulphite sequencing of the Arabidopsis genome reveals DNA methylation patterning, Nature. 452 (2008) 215–219. 10.1038/nature06745. [PubMed: 18278030]

[26]. Urich MA, Nery JR, Lister R, Schmitz RJ, Ecker JR, MethylC-seq library preparation for base-resolution whole-genome bisulfite sequencing, Nat. Protoc 10 (2015) 475–483. [PubMed: 25692984]

[27]. Meissner A, Mikkelsen TS, Gu H, Wernig M, Hanna J, Sivachenko A, Zhang X, Bernstein BE, Nusbaum C, Jaffe DB, Gnirke A, Jaenisch R, Lander ES, Genome-scale DNA methylation maps of pluripotent and differentiated cells, Nature. 454 (2008) 766–770. [PubMed: 18600261]

[28]. Boyle P, Clement K, Gu H, Smith ZD, Ziller M, Fostel JL, Holmes L, Meldrim J, Kelley F, Gnirke A, Meissner A, Gel-free multiplexed reduced representation bisulfite sequencing for large-scale DNA methylation profiling, Genome Biol. 13 (2012) R92. [PubMed: 23034176]

[29]. Akalin A, Garrett-Bakelman FE, Kormaksson M, Busuttil J, Zhang L, Khrebtukova I, Milne TA, Huang Y, Biswas D, Hess JL, Allis CD, Roeder RG, Valk PJM, Löwenberg B, Delwel R, Fernandez HF, Paietta E, Tallman MS, Schroth GP, Mason CE, Melnick A, Figueroa ME, Base-pair resolution DNA methylation sequencing reveals profoundly divergent epigenetic landscapes in acute myeloid leukemia, PLoS Genet. 8 (2012) e1002781. [PubMed: 22737091]

[30]. Meissner A, Reduced representation bisulfite sequencing for comparative high-resolution DNA methylation analysis, Nucleic Acids Research. 33 (2005) 5868–5877. 10.1093/nar/gki901. [PubMed: 16224102]

[31]. Lee YK, Jin S, Duan S, Lim YC, Ng DP, Lin XM, Yeo GS, Ding C, Improved reduced representation bisulfite sequencing for epigenomic profiling of clinical samples, Biol. Proced. Online 16 (2014) 1. [PubMed: 24406024]

[32]. Allum F, Shao X, Guénard F, Simon M-M, Busche S, Caron M, Lambourne J, Lessard J, Tandre K, Hedman ÅK, Kwan T, Ge B, Multiple Tissue Human Expression Resource Consortium, Rönnblom L, McCarthy MI, Deloukas P, Richmond T, Burgess D, Spector TD, Tchernof A, Marceau S, Lathrop M, Vohl M-C, Pastinen T, Grundberg E, Characterization of functional methylomes by next-generation capture sequencing identifies novel disease-associated variants, Nat. Commun 6 (2015) 7211. [PubMed: 26021296]

[33]. Li Q, Suzuki M, Wendt J, Patterson N, Eichten SR, Hermanson PJ, Green D, Jeddeloh J, Richmond T, Rosenbaum H, Burgess D, Springer NM, Greally JM, Post-conversion targeted capture of modified cytosines in mammalian and plant genomes, Nucleic Acids Res. 43 (2015) e81. [PubMed: 25813045]

[34]. Levine ME, Lu AT, Quach A, Chen BH, Assimes TL, Bandinelli S, Hou L, Baccarelli AA, Stewart JD, Li Y, Whitsel EA, Wilson JG, Reiner AP, Aviv A, Lohman K, Liu Y, Ferrucci L, Horvath S, An epigenetic biomarker of aging for lifespan and healthspan, Aging. 10 (2018) 573–591. [PubMed: 29676998]

[35]. Ligthart S, Marzi C, Aslibekyan S, Mendelson MM, Conneely KN, Tanaka T, Colicino E, Waite LL, Joehanes R, Guan W, Brody JA, Elks C, Marioni R, Jhun MA, Agha G, Bressler J, Ward-Caviness CK, Chen BH, Huan T, Bakulski K, Salfati EL, WHI-EMPC Investigators, Fiorito G, CHARGE epigenetics of Coronary Heart Disease, Wahl S, Schramm K, Sha J, Hernandez DG, Just AC, Smith JA, Sotoodehnia N, Pilling LC, Pankow JS, Tsao PS, Liu C, Zhao W, Guarrera S, Michopoulos VJ, Smith AK, Peters MJ, Melzer D, Vokonas P, Fornage M, Prokisch H, Bis JC, Chu AY, Herder C, Grallert H, Yao C, Shah S, McRae AF, Lin H, Horvath S, Fallin D, Hofman A, Wareham NJ, Wiggins KL, Feinberg AP, Starr JM, Visscher PM, Murabito JM, Kardia SLR, Absher DM, Binder EB, Singleton AB, Bandinelli S, Peters A, Waldenberger M, Matullo G, Schwartz JD, Demerath EW, Uitterlinden AG, van Meurs JBJ, Franco OH, Chen Y-DI, Levy D, Turner ST, Deary IJ, Ressler KJ, Dupuis J, Ferrucci L, Ong KK, Assimes TL, Boerwinkle E, Koenig W, Arnett DK, Baccarelli AA, Benjamin EJ, Dehghan A, DNA methylation signatures of chronic low-grade inflammation are associated with complex diseases, Genome Biol. 17 (2016) 255. [PubMed: 27955697]

[36]. Liu C, Marioni RE, Hedman ÅK, Pfeiffer L, Tsai P-C, Reynolds LM, Just AC, Duan Q, Boer CG, Tanaka T, Elks CE, Aslibekyan S, Brody JA, Kühnel B, Herder C, Almli LM, Zhi D, Wang Y, Huan T, Yao C, Mendelson MM, Joehanes R, Liang L, Love S-A, Guan W, Shah S, McRae

AF, Kretschmer A, Prokisch H, Strauch K, Peters A, Visscher PM, Wray NR, Guo X, Wiggins KL, Smith AK, Binder EB, Ressler KJ, Irvin MR, Absher DM, Hernandez D, Ferrucci L, Bandinelli S, Lohman K, Ding J, Trevisi L, Gustafsson S, Sandling JH, Stolk L, Uitterlinden AG, Yet I, Castillo-Fernandez JE, Spector TD, Schwartz JD, Vokonas P, Lind L, Li Y, Fornage M, Arnett DK, Wareham NJ, Sotoodehnia N, Ong KK, van Meurs JBJ, Conneely KN, Baccarelli AA, Deary IJ, Bell JT, North KE, Liu Y, Waldenberger M, London SJ, Ingelsson E, Levy D, A DNA methylation biomarker of alcohol consumption, Mol. Psychiatry 23 (2018) 422–433. [PubMed: 27843151]

[37]. Natoli G, Maintaining cell identity through global control of genomic organization, Immunity. 33 (2010) 12–24. [PubMed: 20643336]

[38]. Teschendorff AE, Zheng SC, Cell-type deconvolution in epigenome-wide association studies: a review and recommendations, Epigenomics. 9 (2017) 757–768. [PubMed: 28517979]

[39]. Houseman EA, Accomando WP, Koestler DC, Christensen BC, Marsit CJ, Nelson HH, Wiencke JK, Kelsey KT, DNA methylation arrays as surrogate measures of cell mixture distribution, BMC Bioinformatics. 13 (2012) 86. [PubMed: 22568884]

[40]. Martens JHA, Stunnenberg HG, BLUEPRINT: mapping human blood cell epigenomes, Haematologica. 98 (2013) 1487–1489. [PubMed: 24091925]

[41]. Ziller MJ, Gu H, Müller F, Donaghey J, Tsai LT-Y, Kohlbacher O, De Jager PL, Rosen ED, Bennett DA, Bernstein BE, Gnirke A, Meissner A, Charting a dynamic DNA methylation landscape of the human genome, Nature. 500 (2013) 477–481. [PubMed: 23925113]

[42]. Cock PJA, Fields CJ, Goto N, Heuer ML, Rice PM, The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants, Nucleic Acids Res. 38 (2010) 1767–1771. [PubMed: 20015970]

[43]. Zhang H, Overview of Sequence Data Formats, Methods Mol. Biol 1418 (2016) 3–17. [PubMed: 27008007]

[44]. Illumina FastQ, Illumina - FastQ File Format. (n.d.). https://support.illumina.com/bulletins/2016/04/fastq-files-explained.html.

[45]. Martin M, Cutadapt removes adapter sequences from high-throughput sequencing reads, EMBnet.journal 17 (2011) 10. 10.14806/ej.17.1.200.

[46]. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, 1000 Genome Project Data Processing Subgroup, The Sequence Alignment/Map format and SAMtools, Bioinformatics. 25 (2009) 2078–2079. [PubMed: 19505943]

[47]. HTS format specifications, (n.d.). https://samtools.github.io/hts-specs/ (accessed April 28, 2020).

[48]. Basu A, Reproducible research with jupyter notebooks, Authorea. (n.d.). 10.22541/au.151460905.57485984.

[49]. Garreta R, Moncecchi G, Learning scikit-learn: Machine Learning in Python, Packt Publishing Ltd, 2013.

[50]. Virtanen P, Gommers R, Oliphant TE, Haberland M, Reddy T, Cournapeau D, Burovski E, Peterson P, Weckesser W, Bright J, van der Walt SJ, Brett M, Wilson J, Millman KJ, Mayorov N, Nelson ARJ, Jones E, Kern R, Larson E, Carey CJ, Polat İ, Feng Y, Moore EW, VanderPlas J, Laxalde D, Perktold J, Cimrman R, Henriksen I, Quintero EA, Harris CR, Archibald AM, Ribeiro AH, Pedregosa F, van Mulbregt P, Vijaykumar A, Bardelli AP, Rothberg A, Hilboll A, Kloeckner A, Scopatz A, Lee A, Rokem A, Woods CN, Fulton C, Masson C, Häggström C, Fitzgerald C, Nicholson DA, Hagen DR, Pasechnik DV, Olivetti E, Martin E, Wieser E, Silva F, Lenders F, Wilhelm F, Young G, Price GA, Ingold G-L, Allen GE, Lee GR, Audren H, Probst I, Dietrich JP, Silterra J, Webber JT, Slavič J, Nothman J, Buchner J, Kulick J, Schönberger JL, de Miranda Cardoso JV, Reimer J, Harrington J, Rodríguez JLC, Nunez-Iglesias J, Kuczynski J, Tritz K, Thoma M, Newville M, Kümmerer M, Bolingbroke M, Tartre M, Pak M, Smith NJ, Nowaczyk N, Shebanov N, Pavlyk O, Brodtkorb PA, Lee P, McGibbon RT, Feldbauer R, Lewis S, Tygier S, Sievert S, Vigna S, Peterson S, More S, Pudlik T, Oshima T, Pingel TJ, Robitaille TP, Spura T, Jones TR, Cera T, Leslie T, Zito T, Krauss T, Upadhyay U, Halchenko YO, Vázquez-Baeza Y, SciPy 1.0 Contributors, SciPy 1.0: fundamental algorithms for scientific computing in Python, Nat. Methods (2020). 10.1038/s41592-019-0686-2.

[51]. Hunter JD, Matplotlib: A 2D Graphics Environment, Computing in Science & Engineering. 9 (2007) 90–95. 10.1109/mcse.2007.55.

[52]. Miranda M, Preparation of 1.5 mg/mL Sera-mag carboxylate modified magnetic particles v1, (n.d.). 10.17504/protocols.io.g2abyae.

[53]. qPCR Quantification Guide, Illumina. (2011). https://support.illumina.com/content/dam/illumina-support/documents/documentation/chemistry_documentation/qpcr/sequencing-library-qpcr-quantification-guide-11322363-c.pdf.

Highlights

- This article presents a step-by-step protocol to perform Targeted Bisulfite Sequencing for Biomarker Discovery
- We also present the computational steps to fit a 5^{me}C-based epigenetic clock, using age information
- The approach can be applied to develop cost effective epigenetic biomarkers

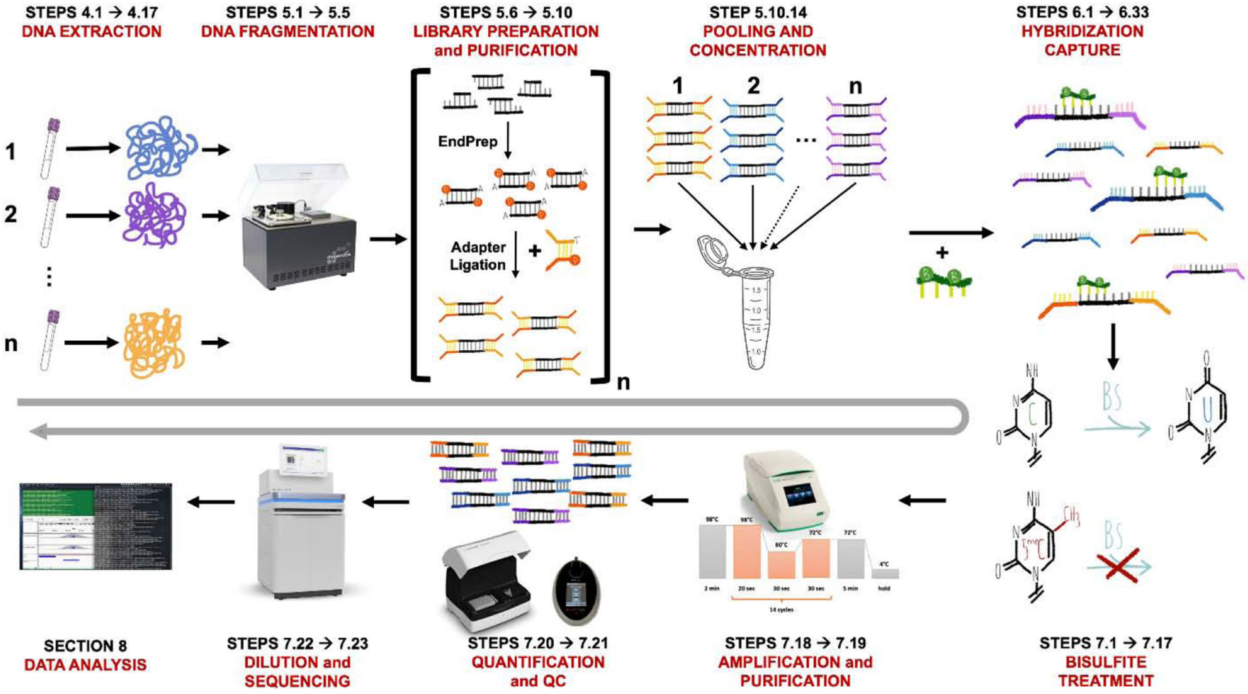


Figure 1: Overview of the Targeted Bisulfite Sequencing Protocol.

Genomic DNA is extracted from collected blood samples (section 4), fragmented and subject to NGS library preparation (5.1 5.10). Adapter ligated libraries are then pooled (5.10.14), concentrated (6.1) and incubated with RNA biotinylated probes to enrich for target regions (6.2 6.33 and Figure 2). Captured DNA fragments are then bisulfite-treated (7.1 7.17), PCR amplified, Quality Controlled (7.18 7.22) and sequenced (7.23). Data is then analyzed (see Section 8 and Figure 4 for more details).

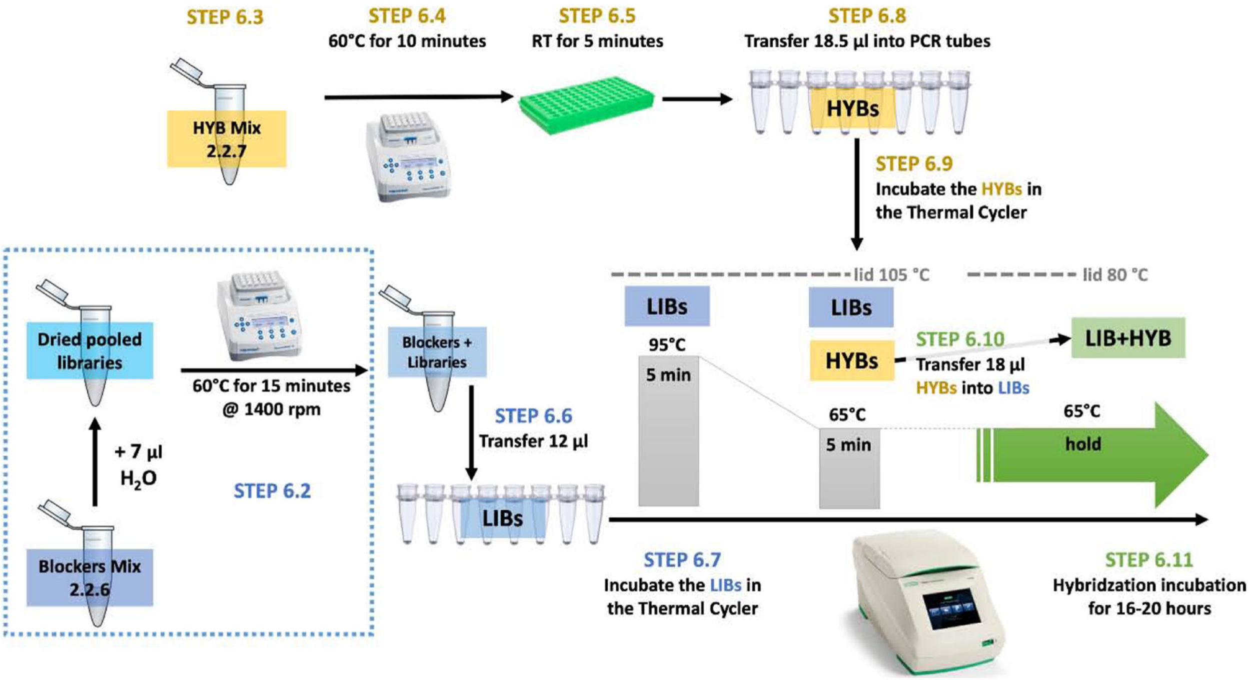
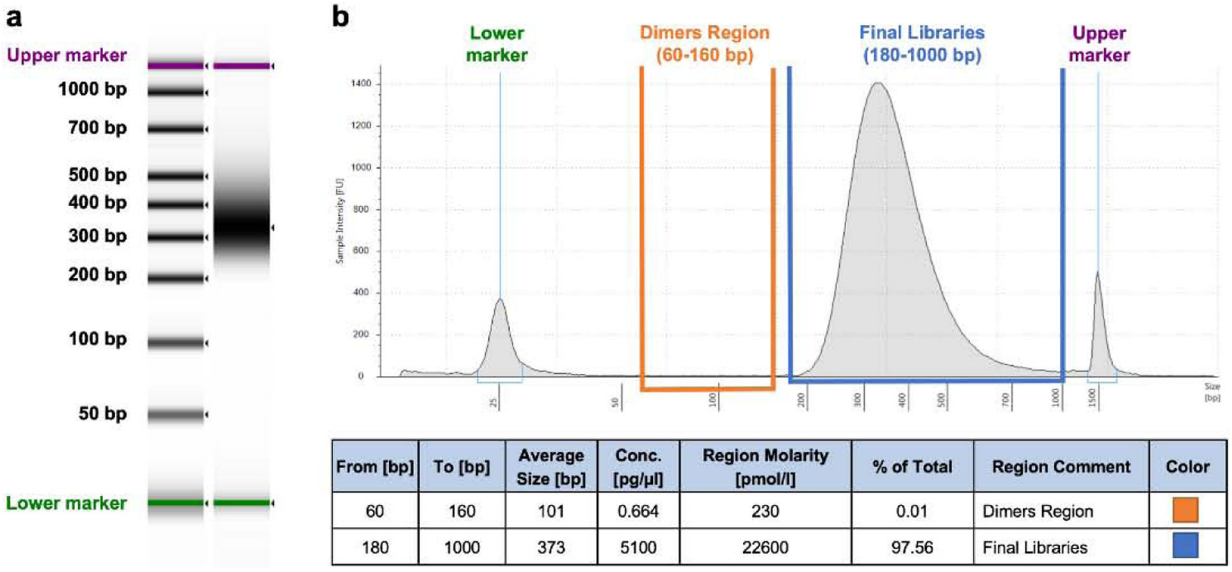


Figure 2: Hybridization Capture setup.

The Hybridization mix (top, yellow boxes – 2.2.7) is prepared and incubated in a Thermomixer for 10 min at 60°C (6.4). After 5 min of Room Temperature (RT) incubation (6.5), 18.5 µl are transferred into PCR tubes (6.8). The HYB tubes are stored at RT until step 6.9. Dried library pools from step 6.1 are resuspended for 15 min at 60°C in a Thermomixer after the addition of 7 µl of H2O and 5 µl of Blockers Mix (2.2.6) (Step 6.2). The blockers + libraries mixture (12 µl) is then transferred into PCR tubes that are now called LIBs (6.6). LIB tubes are then transferred into a thermocycler for denaturation and hybridization temperature equilibration (step 6.7). After the thermocycler reaches the hybridization temperature (65°C), HYB tubes are added into the machine (6.9). After 5 minutes at 65°C, 18 µl of the HYB mix are transferred into each LIB tube (still in the thermocycler) (6.10). HYB tubes are discarded, and the LIB+HYB tubes are incubated for 16–20 hours at the hybridization temperature (6.11).



From [bp]	To [bp]	Average Size [bp]	Conc. [pg/µl]	Region Molarity [pmol/l]	% of Total	Region Comment	Color
60	160	101	0.664	230	0.01	Dimers Region	🟧
180	1000	373	5100	22600	97.56	Final Libraries	🟦

Figure 3: Final Libraries Quality Control.

Agilent TapeStation 2200 High Sensitivity D1000 ScreenTape Assay (7.21) of the Final Libraries obtained with the Targeted Bisulfite Sequencing Approach (7.19.14). (a) Gel View of the Ladder and Final Libraries. (b) Electropherogram View with information about the Dimers Region (orange, 60–160 bp) and the Final Libraries Region (blue, 180–1000 bp).

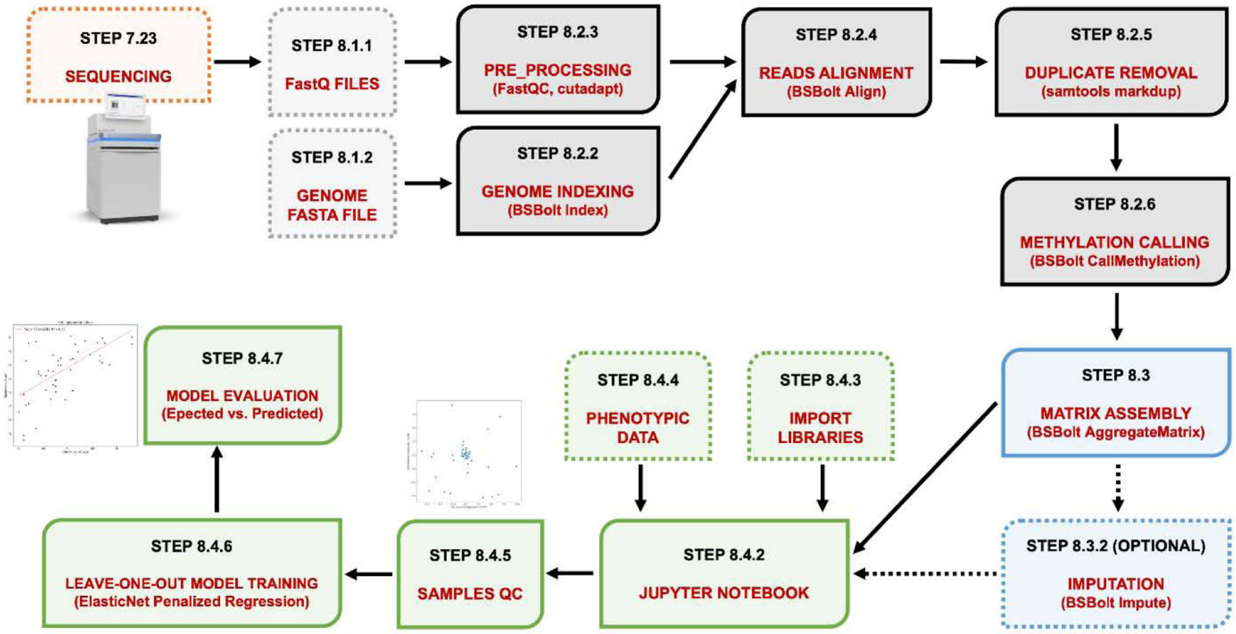


Figure 4: Data Analysis Pipeline described in Section 8.

FastQ files (8.1.1) are obtained from NGS-sequencing (7.23) and subject to quality control (FastQC) before and after Adapter Trimming (cutadapt) (8.2.3). Genome fasta files can be obtained from sequence databases (8.1.2) and then subject to Genome Indexing (BSBolt Index – 8.2.2). Both Genome Index and trimmed FastQ files are used as input for Reads alignment (BSBolt Align – 8.2.4). After Alignment, duplicated reads are removed (samtools markdup – 8.2.5) and methylation is called for every cytosine (BSBolt CallMethylation – 8.2.6), creating CGmap files. Several CGmap files can be combined into a single matrix using BSBolt AggregateMatrix (8.3). Methylation values for missing sites can be imputed using BSBolt Impute (OPTIONAL – 8.3.2). Model fitting is then performed in a Jupyter Notebook (8.4.2), where external libraries (8.4.3) and phenotypic data (8.4.4) are imported. After Samples QC to detect outliers (8.4.5), the model is trained using a leave-one-out elastic-net penalized regression (8.4.6). The model is then evaluated by fitting a trendline between the known and the predicted phenotypic value (8.4.7).

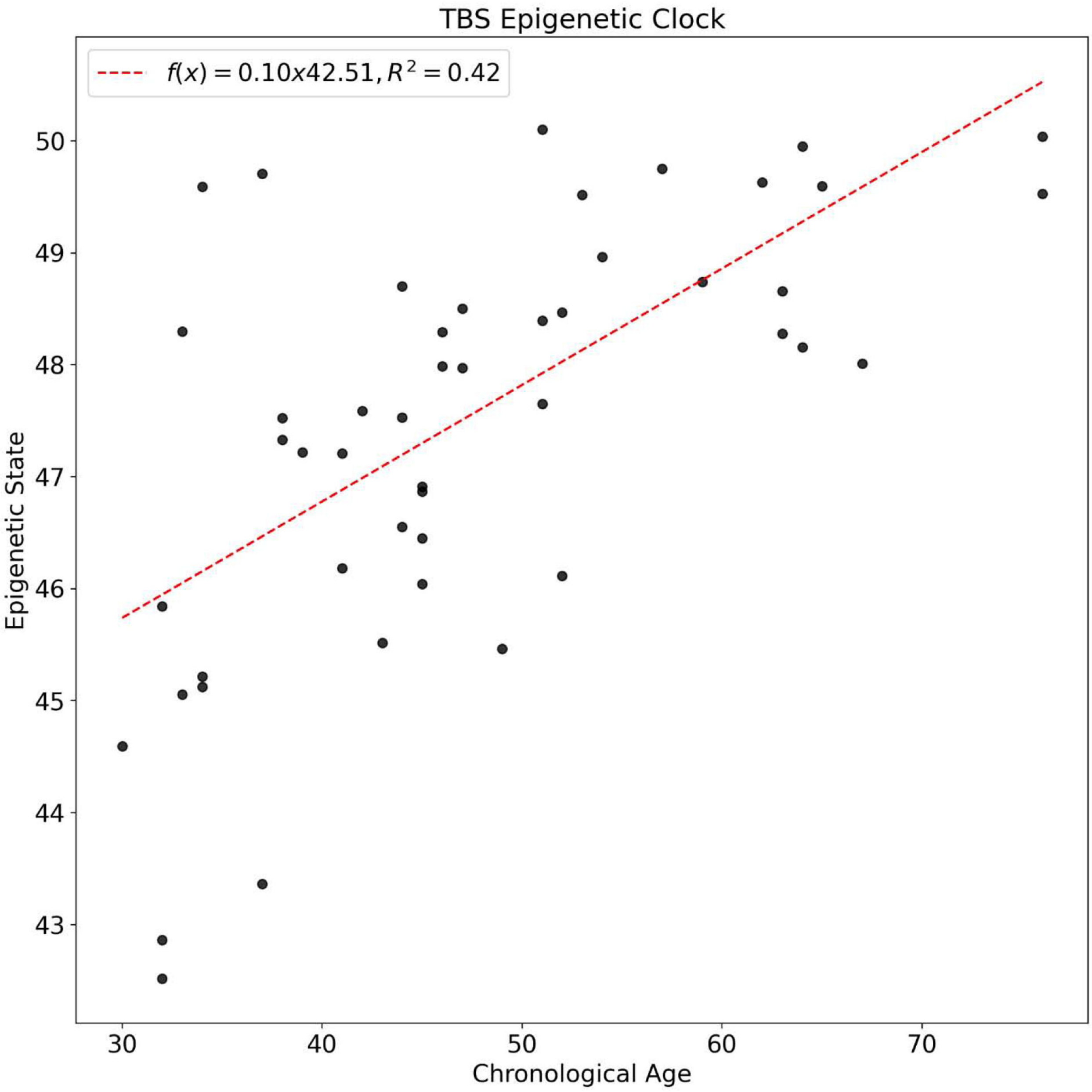


Figure 5: Targeted Bisulfite Sequencing Epigenetic Clock.

Epigenetic age predictions for (n=48) samples made using penalized regression models compared to the chronological age of each sample with a line of best fit. The chronological (observed) age is represented on the x-axis, while the predicted epigenetic age is on the y-axis.

Table 1:

BSBolt Index Parameters.

Option	Description
-G	Path to reference genome fasta file, fasta file should contain all contigs
-DB	Path to index directory, will create directory if folder does not exist
-MR	Path to bed file of mappable regions. Index will be built using a masked contig sequence. *We suggest to align to the entire genome to reduce false positives, hence, do not restrict the Genome Indexing to the target regions.

Table 2:

BSBolt Align Parameters.

Option	Description
-F1	Path to fastq 1
-F2	Path to fastq 2
-UN	Library Undirectional, Consider PCR products of bisulfite converted DNA
-O	Path to Output Prefix
-G	Path to BSBolt Database
-t	Number of bwa threads [1]
-k	Minimum seed length [19]
-w	Band width for banded alignment [100]
-d	off-diagonal X-dropoff [100]
-r	look for internal seeds inside a seed longer than {-k} * FLOAT [1.5]
-y	seed occurrence for the 3rd round seeding [20]
-c	skip seeds with more than INT occurrences [500]
-D	drop chains shorter than FLOAT fraction of the longest overlapping chain [0.50]
-W	discard a chain if seeded bases shorter than INT [0]
-m	perform at most INT rounds of mate rescues for each read [50]
-S	skip mate rescue
-P	skip pairing; mate rescue performed unless -S also in use
-A	score for a sequence match, which scales options -TdBOELU unless overridden [1]
-B	penalty for a mismatch [4]
-INDEL	gap open penalties for deletions and insertions [6,6]
-E	gap extension penalty; a gap of size k cost '{-O} + {-E}*k' [1,1]
-L	penalty for 5'- and 3'-end clipping [30,30]
-U	penalty for an unpaired read pair [17]
-p	smart pairing (ignoring in2.fq)
-R	read group header line such as '@RG ID:foo SM:bar' [null]
-H	insert STR to header if it starts with @; or insert lines in FILE [null]
-j	treat ALT contigs as part of the primary assembly (i.e. ignore <idxbase>.alt file)
-T	minimum score to output [80], set based on read length
-XA	if there are <INT hits with score >80 percent of the max score, output all in XA [5,200]
-M	mark shorter split hits as secondary
-I	specify the mean, standard deviation (10 percent of the mean if absent), max (4 sigma from the mean if absent) and min of the insert size distribution. FR orientation only. [inferred], Float,Float,Int,Int

Table 3:

BSBolt CallMethylation Parameters.

Option	Description
-I	Input BAM, input file must be in BAM format with index file
-DB	Path to index directory
-O	Output prefix
-remove-ccgg	Remove methylation calls in ccgg sites,default=False
-verbose	Verbose Output, default=False
-text	Output plain text files, default=False
-remove-sx	deprecated
-ignore-overlap	Only consider higher quality base when paired end reads overlap, default=False
-max	Max read depth to call methylation
-min	Minimum read depth required to report methylation site
-t	Number of threads to use when calling methylation values
-BQ	Minimum base quality for a base to considered for methylation calling, default=0
-MQ	Minimum alignment quality for an alignment to be considered for methylation calling, default=20
-CG	Only output CpG sites in CGmap file
-ATCG	Output ATCGmap file
-IO	Ignore orphans during methylation call

Table 4:

CGmap file structure.

Column	Value
1	Chromosome
2	Nucleotide, C for reads mapped to the Watson (sense) strand and G for reads mapped to the Crick (anti-sense) strand
3	Position, base-pairs from start
4	Context, three base pair methylation context
5	Sub-Context, two base pair methylation context
6	Methylation Value, proportion of methylation reads to total reads
7	Methylation Bases, methylated nucleotides observed
8	All Bases, total number of nucleotides observed at the mapping position

Table 5:

BSBolt AggregateMatrix Parameters.

Option	**Description**
-F	Comma separated list of CGmap file paths, or path to text file with list of line separated CGmap file paths
-S	Comma separated list of samples labels. If sample labels are not provided sample labels are extracted from CGmap file paths. Can also pass path to txt for line separated sample labels.
-min-coverage	Minimum site read depth coverage for a site to be included in the aggregate matrix
-min-sample	Proportion of samples that must have a valid site (above minimum coverage threshold), for a site to be included in the aggregate matrix.
-O	Aggregate matrix output path
-CG	Only output CG sites
-verbose	Verbose aggregation
-t	Number of threads to use when assembling matrix

Table 6:

CGmatrix file.

Site	S1	S2	S3	S4
chr1:807868	0.3	0	0.125	0
chr1:808073	0.724138	0.7	0.772727	0.666667
chr1:808133	1	1	0.83333	0.69412
chr1:821016	0.727273	0.673077	0.724638	0.689655
chr1:821074	0.814815	0.810811	0.815789	0.741935
chr1:821139	1	0.91414	0.666667	1
chr1:821154	0.289474	0.268817	0.27551	0.268293
chr1:821179	0.77273	1	1	0.65079
chr1:821240	0.888889	0.86647	0.769231	0.76607
chr1:821261	0.741935	0.785714	0.833333	0.666667
chr1:821401	0.758621	0.73913	0.975	0.909091
chr1:821459	0.928571	0.875	1	1
chr1:821506	0.916667	0.93468	0.866667	1
chr1:821576	0.714286	0.375	0.8125	0.5
chr1:1138336	0.5	0.5	0.272727	0.444444
chr1:1138426	0.166667	0.178571	0.391304	0.076923
chr1:1138427	0.272727	0.238095	0.235294	0.2
chr1:1138462	0.40625	0.296296	0.259259	0.333333
chr1:1138463	0.409091	0.173913	0.208333	0.411765
chr1:1138548	0.466667	0.366667	0.392857	0.315789

Example of a CGmatrix File, output of BSBolt AggregateMatrix command. Only 20 cytosines for each of four samples are shown. The complete CGmatrix can be found at https://github.com/NuttyLogic/MethodsTBS.