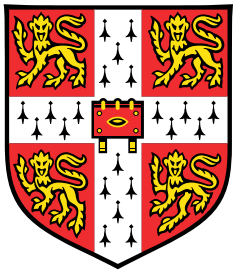
On the epigenetic ageing clock in humans



Daniel Elías Martín Herranz

European Bioinformatics Institute (EMBL-EBI)
University of Cambridge

This dissertation is submitted for the degree of
Doctor of Philosophy

Churchill College

April 2019

I would like to dedicate this thesis to my loving parents ...

Declaration

I hereby declare that except where specific reference is made to the work of others, the contents of this dissertation are original and have not been submitted in whole or in part for consideration for any other degree or qualification in this, or any other university. This dissertation is my own work and contains nothing which is the outcome of work done in collaboration with others, except as specified in the text and Acknowledgements. This dissertation contains fewer than 65,000 words including appendices, bibliography, footnotes, tables and equations and has fewer than 150 figures.

Daniel Elías Martín Herranz
April 2019

Acknowledgements

And I would like to acknowledge ...

Abstract

This is where you write your abstract ...

Table of contents

List of figures

List of tables

Abbreviations and acronyms

27K	Illumina Infinium HumanMethylation27 array
450K	Illumina Infinium HumanMethylation450 array
5mC	5-methylcytosine
BMIQ	Beta-mixture quantile normalisation
bp	Base pairs
CG	5′-cytosine-phosphate-guanine-3′
CGI	CpG island
CHG	5′-cytosine-phosphate-H-phosphate-guanine-3′, where H corresponds to adenine, thymine or cytosine
CHH	5′-cytosine-phosphate-H-phosphate-H-3′, where H corresponds to adenine, thymine or cytosine
ChIP-seq	Chromatin immunoprecipitation and sequencing
CpG	5′-cytosine-phosphate-guanine-3′
CPU	Central processing unit
CRF	Cost Reduction Factor in cuRRBS
CTCF	CCCTC-binding factor
cuRRBS	customised Reduced Representation Bisulfite Sequencing
DMRs	Differentially methylated regions
DNA	Deoxyribonucleic acid
EPIC	Illumina Infinium MethylationEPIC array
EV	Enrichment Value in cuRRBS

FN	False negatives
FP	False positives
GB	Gigabytes
Gbp	Giga base pairs
GC content	Guanine + cytosine content
GEO	Gene Expression Omnibus repository
hg38	Reference human genome assembly 38
iPSCs	Induced pluripotent stem cells
kb	Kilo base pairs
MBD	Methyl-CpG-binding domain
MEFs	Mouse embryonic fibroblasts
NF	Theoretical number of fragments sequenced in cuRRBS
NRF1	Nuclear respiratory factor 1
PCA	Principal component analysis
PCR	Polymerase chain reaction
QC	Quality control
R	It can have two meanings: robustness variable in cuRRBS or the R programming language
RAM	Random-access memory
RNA	Ribonucleic acid
RRBS	Reduced Representation Bisulfite Sequencing
SD	Standard deviation

Sex_p	Sex predicted for a sample using DNA methylation data
TKO	Triple knockout
TN	True negatives
TP	True positives
TSS	Transcription start site
WGBS	Whole Genome Bisulfite Sequencing

Chapter 1

Getting started

1.1 What is loren ipsum?

Chapter 2

Statistical aspects of the epigenetic clock

2.1 Analysing the blood methylome to study human ageing

2.1.1 Building a DNA methylation dataset from public data

During the last years large amounts of DNA methylation data have been generated to study complex diseases and ageing [1, 2]. Many of these datasets can be obtained from public repositories, such as the NCBI-hosted Gene Expression Omnibus (GEO) [3]. Given its clinical accessibility and ease of collection, blood is one of the most commonly profiled tissues in human DNA methylation studies [2], including published studies on developmental disorders [4] (see Chapter 3). Therefore, I decided to use blood as my surrogate tissue to broaden our understanding of the human epigenetic ageing clock.

Furthermore, most of these human datasets have been generated using different versions of the Illumina Infinium array technology, with the Illumina Infinium HumanMethylation450 array (450K) being the most frequently used platform [2]. Additionally, given that the different array versions have different chemistries and biases [5–7], I decided to focus on 450K data for my analyses. Using the *GEOquery* R package [8], I programmatically downloaded from GEO all the DNA methylation data that I could find or human blood that satisfied the following criteria:

- Raw DNA methylation data was available (i.e. IDAT files). This was required so the pre-processing pipeline and the batch effect correction (which requires access to control probes intensities, see 'xxxxxxx' section) could be consistently applied across all the samples in the study.

- Metadata for the samples was available, with the chronological age as an absolute requirement.
- In order to study physiological ageing, the blood samples corresponded to humans without any major disease. However, it is important to mention that I could never be completely certain of this, since there could be a lack of diagnosis and/or lack of reporting of the disease in the metadata.

This allowed me to assembly a human blood DNA methylation dataset for healthy individuals (after QC, total $N = 2218$) with the characteristics shown in Table 2.1, which spans the entire human lifespan (0.5 to 101 years). Fig. 2.1 shows that the chronological age distribution is bimodal, with peaks around 10.69 and 58.81 years respectively. This reflects a sampling bias in human population studies, with more data being generated for the periods of postnatal development and during the appearance of age-related disease. However, in order to understand the development of complex diseases as a consequence of the ageing process, efforts should be made to sample people also in their middle ages, before the diseases are normally diagnosed.

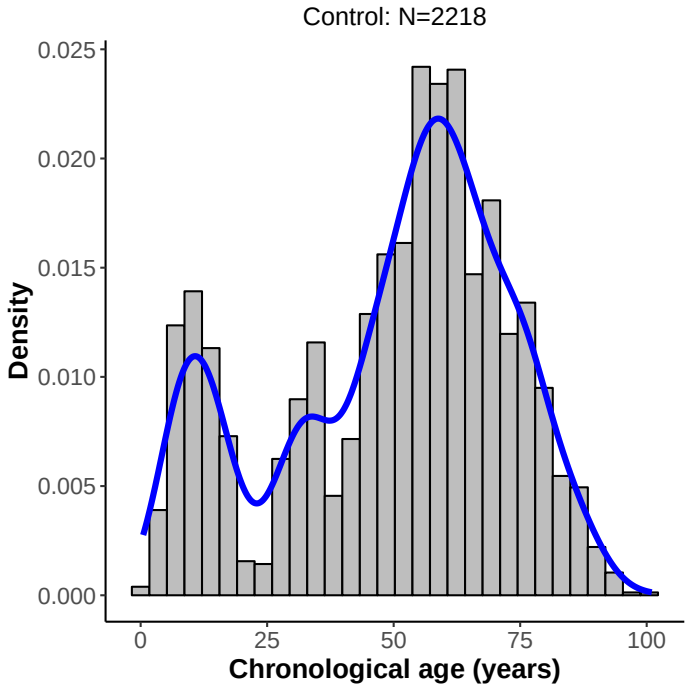


Fig. 2.1 Chronological age distribution in our healthy individuals. Histogram showing the chronological age distribution for all the healthy individuals included in our DNA methylation dataset. The blue line represents the 1D kernel density estimate, as calculate by the *stat_density* function in R with default parameters.

Batch name	**N♀**	**N♂**	**N**	**Median age (years)**	**Other comments**
Europe	0	121	121	10.96	-
Feb_2016	0	1	1	0.50	-
GSE104812	19	29	48	9.00	-
GSE111629	111	124	235	71.00	-
GSE40279	336	314	650	65.00	-
GSE41273	0	51	51	10.25	-
GSE42861	239	96	335	55.00	-
GSE51032	253	78	331	54.57	Only people that remained cancer-free in the follow-up after sample collection were included
GSE55491	1	5	6	29.50	-
GSE59065	49	46	95	34.00	-
GSE61496	72	78	150	57.00	Only one member of each twins pair was included
GSE74432	29	22	51	12.00	-
GSE81961	25	0	25	30.05	-
GSE97362	39	80	119	13.00	-
Total	1173	1045	2218	55.00	-

Table 2.1 Overview of the blood DNA methylation dataset from healthy individuals. All the batches were downloaded from GEO [3], with the exception of 'Europe' and 'Feb_2016', which were generated in-house by our collaborators in Canada (see Chapter 3). $N_{♀}$: number of samples from females. $N_{♂}$: number of samples from males. N: total number of samples. These numbers correspond to the samples left after applying quality control (QC, see 'Overview of the DNA methylation pre-processing pipeline')

2.1.2 Main DNA methylation data pre-processing pipeline

The analysis of DNA methylation data generated in Illumina arrays has been a topic of huge discussion and statistical innovation in the epigenetic community. There are plenty of reviews in the literature that discuss the different steps that should be involved in the pre-processing of this data type [9–11]. More specifically, a recent study by Je Liu and Kimberly D. Siegmund systematically benchmarked the pre-processing methods available for the 450K array in order to reduce variation among technical replicates and improve the detection of biological differences [11]. Inspired by their results, I implemented, using the *minfi* R package [12], a pre-processing pipeline for the 450K data with the following steps (Fig. 2.2):

1. **Background correction**. I used the *noob* method [13], as implemented in the *preprocessNoob* function from the *minfi* R package [12]. *noob* allows accounting for technical variation in the background (i.e. non-specific) fluorescence signal, which can lead to a reduced dynamic range for the β-values obtained (Fig. 2.2b, Fig. S1.1) [13]. Briefly, when measuring fluorescence intensities in the Illumina array platforms, the observed intensity (also known as foreground, X_f) is composed of:

$$X_f = X_s + X_b \tag{2.1}$$

   where X_s is the true signal and X_b is the background signal. Making use of a normal-exponential convolution (which assumes $X_b \sim N(\mu, \sigma^2)$ and $X_s \sim Exp(\gamma)$) and the 'out-of-band' (OOB) intensities (fluorescence signals in the opposite colour channel in Infinium I probes) to model X_b , *noob* is capable of estimating X_s given X_f. Furthermore, I also applied the default dye-bias correction strategy, which controls for the different average intensities in the two colour channels [13].

2. **Quality control** (QC). Following guidelines from the *minfi* R package [12], I kept only those samples that satisfied the following criteria:

   (a) The sex predicted from the DNA methylation data (Sex_p) was the same as the reported sex in the metadata. The sex was predicted using the *getSex* function from the *minfi* R package [12], which employs intensity information from the sex chromosomes, such that:

$$\text{Sex}_p = \begin{cases} \text{female}, & \text{if: } (\text{median}\left\{\log_2(M_y + U_y)\right\} - \text{median}\left\{\log_2(M_x + U_x)\right\}) < c \\ \text{male}, & \text{if: } (\text{median}\left\{\log_2(M_y + U_y)\right\} - \text{median}\left\{\log_2(M_x + U_x)\right\}) \geq c \end{cases} \tag{2.2}$$

where M_y and U_y represent the methylated and unmethylated intensity measurements for the array probes in the Y chromosome, M_x and U_x represent the methylated and unmethylated intensity measurements for the array probes in the X chromosome and c is a predefined cutoff (default in *minfi*: $c = -2$).

(b) They were not outliers according to their global intensity values after background correction, such that:

$$\frac{\text{median}\left\{\log_2(M_i)\right\} + \text{median}\left\{\log_2(U_i)\right\}}{2} \geq 10.5 \tag{2.3}$$

where M_i and U_i represent the background-corrected methylated and unmethylated intensity measurements for all the 450K array probes (Fig. S1.2).

3. **Probe filtering**. I filtered out the following types of probes:

   - Probes that contain SNPs at the single base extension site (position 0) or at the proximal CpG on the probe (positions 1-2), using the *dropLociWithSnps* function in the *minfi* package [12].
   - Cross-reactive probes, as defined by Chen *et al.* [14]. These are probes that can co-hybridise to alternative genomic sequences that are highly homologous to the target sequences [14].
   - Probes that map to the sex chromosomes (X and Y).

   It is important to mention that other authors have also filtered out probes with high detection p-value or low bead counts across samples [9, 10]. However, I did not include these filters since it was not pointed out in the *minfi* guidelines [12, 15] and it could complicate further downstream analyses (e.g. different sets of probes missing across different batches, discarding probes that were needed for cell composition estimation, ...).

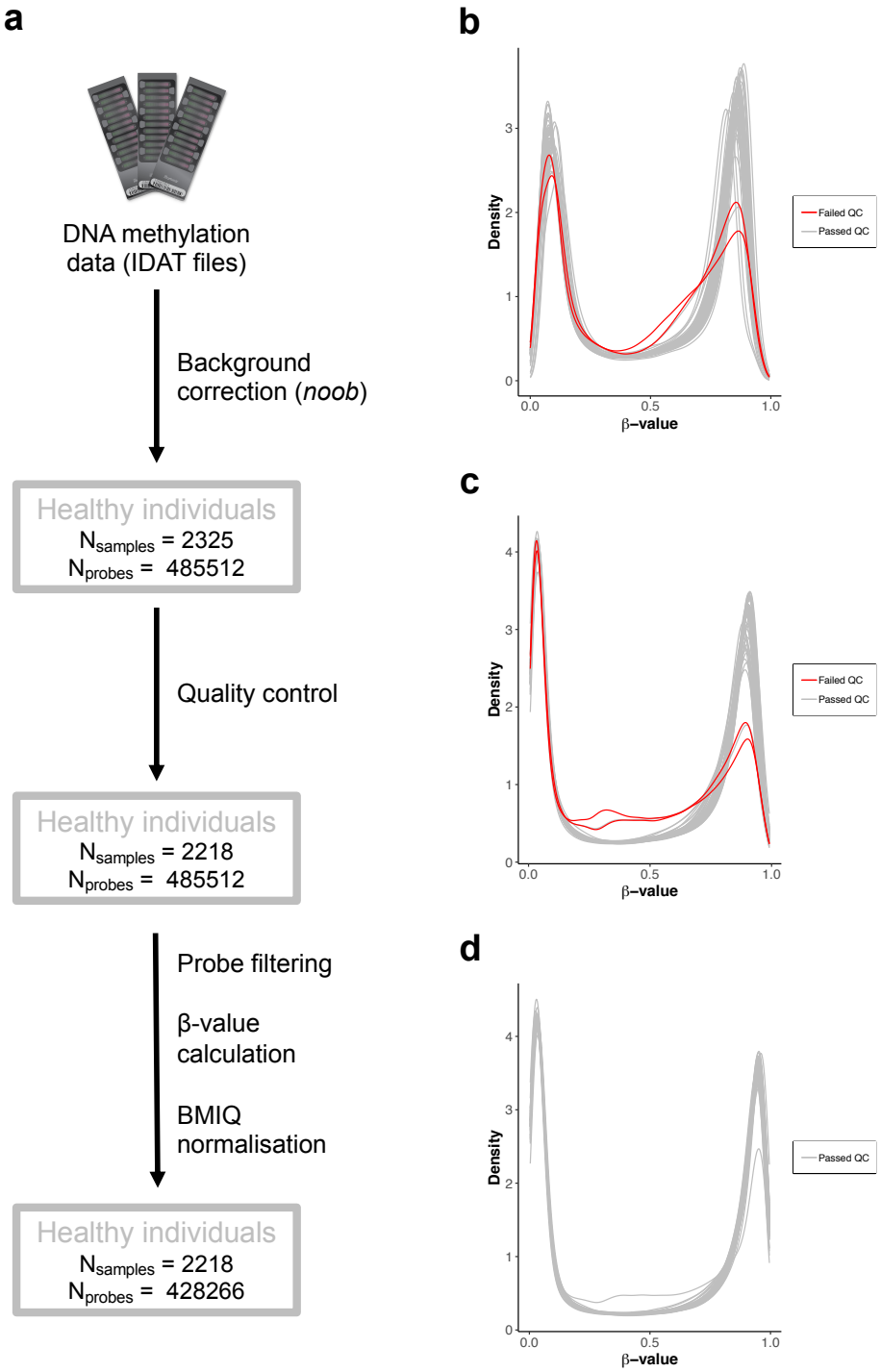


Fig. 2.2 Main DNA methylation data pre-processing pipeline. **a.** Flowchart showing the main steps that I implemented to pre-process the DNA methylation data for the healthy individuals. The number of samples ($N_{samples}$) and the number of array probes (N_{probes}) left after each step are also specified. **b.** β-value distributions, calculated using the raw fluorescence intensities (i.e. before any pre-processing), for the samples in the GSE41273 batch. Each curve represents a different sample. In grey: 51 samples that passed quality control (QC). In red: 2 samples that failed QC. **c.** As in b., but calculating the β-values after background correction. **d.** As in b., but calculating the β-values after background correction, QC, probe filtering and BMIQ normalisation (i.e. the final β-values that I used for downstream analyses). Note that the samples that failed QC have been removed.

4. **β-value calculation**. The methylation status of a given CpG site in one of the array probes is normally quantified using the β-value statistic (β), which can be calculated as [9, 16]:

$$\beta_i = \frac{\max(M_i, 0)}{\max(M_i, 0) + \max(U_i, 0) + \alpha} \tag{2.4}$$

where M_i and U_i represent the methylated and unmethylated intensity measurements for the ith-probe and α is a constant offset (in this work $\alpha = 100$, as recommended by Illumina) [16].

In a DNA copy (allele) of a single cell, a specific CpG site is either unmethylated or methylated (categorical / binary variable). However, given that a bulk DNA sample from a tissue is composed of thousands of cells (which can include different cell types with different methylation patterns), β-values result in a continuous variable between 0 and 1. A value of 0 means that all the measured DNA molecules are unmethylated (0%) and a value of 1 means that all the measured DNA molecules are methylated (100%), which is roughly equivalent to say that 100% of the cells are either unmethylated or methylated respectively in that CpG site for the sampled tissue. The β-values for a given sample normally follow a bimodal distribution, where the two peaks are centred around 0 and 1 (Fig. 2.2d).

Other authors have used M-values to quantify methylation levels in arrays (Fig. S1.3), which can be calculated as:

$$\text{M-value}_i = \log_2 \left(\frac{\max(M_i, 0) + \alpha}{\max(U_i, 0) + \alpha} \right) \tag{2.5}$$

with a default offset value of $\alpha = 1$. Du *et al.* reported that β-values suffer from severe heteroscedasticity for highly methylated or unmethylated CpG sites and therefore the M-values have more desirable statistical properties [16]. However, Zhuang *et al.* later showed that this only becomes a problem in studies with small sample sizes [17] (which is not the case for my analyses). Furthermore, β-values are easier to interpret biologically and can be readily used in the context of BMIQ normalisation (see below). For these reasons, I choose β-values as the main methylation variable for this work.

5. **Beta-mixture quantile normalisation** (BMIQ). As mentioned in Chapter 1, in the case of the 450K arrays two types of probes / chemistry coexist in the same platform.

Infinium I probes and Infinium II probes have different β-values distributions (a.k.a. Infinium II probe bias). BMIQ is an intra-array normalisation strategy that allows to correct for this bias and has been shown to outperform other methods used in this context [18–21]. BMIQ fits a three-state beta-mixture model to Infinium I and Infinium II probes separately and then maps the Infinium II probes distribution into the Infinium I probe distribution (Fig. 2.3). In the case of unmethylated (β-values close to 0) and methylated (β-values close to 1) probes, this is done by transforming probabilities into quantiles. In the case of 'hemimethylated' probes (intermediate β-values), a dilation transformation is applied to preserve the monotonicity and continuity of the data [18]. I applied BMIQ to my samples and discarded those that failed the normalisation step.

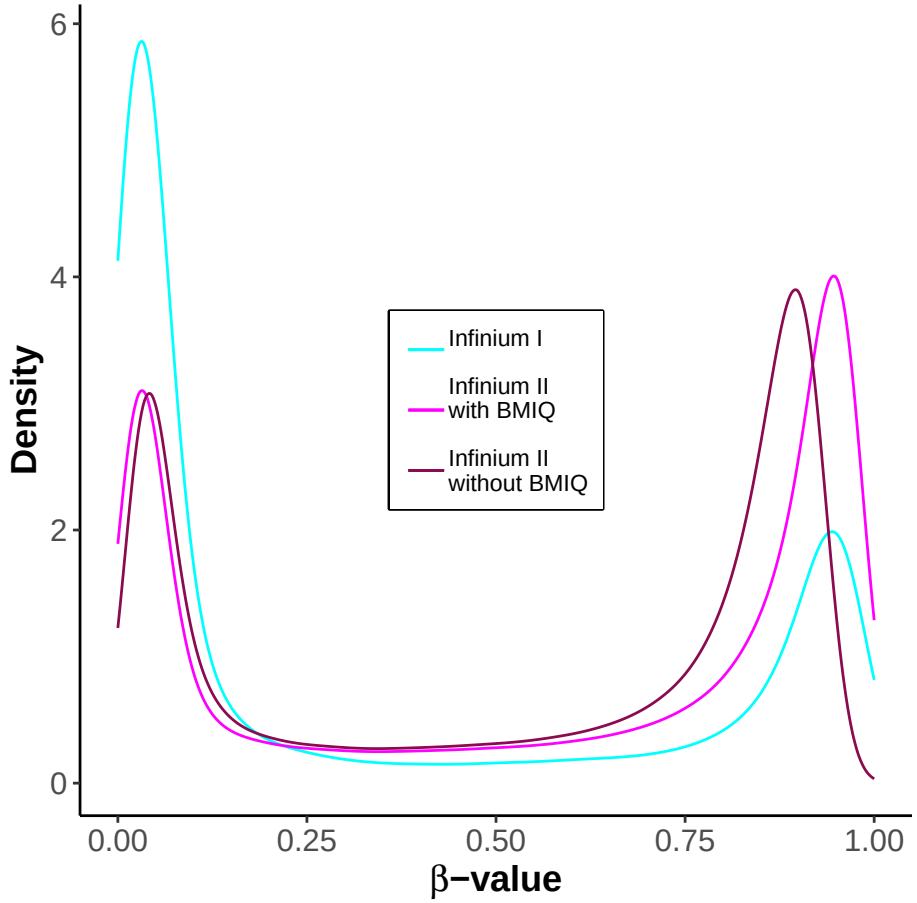


Fig. 2.3 Effect of BMIQ normalisation on the β-value distribution. The β-value distributions for different subsets of array probes in a DNA methylation sample from the GSE41273 batch. It can be appreciated how BMIQ transforms the distribution of the Infinium II probes into a distribution more similar to the Infinium I probes.

2.2 Behaviour of Horvath's epigenetic clock during ageing

2.3 Behaviour of other epigenetic clocks during ageing

2.4 Additional methods

Experimental procedures for DNA methylation data generation

Chapter 3

Biological aspects of the epigenetic clock

3.1 What is loren ipsum?

Chapter 4

Technological aspects of epigenetic clocks

4.1 Background

With the advent of next-generation sequencing, scientists are studying the biology of life at unprecedented resolution [22]. Unfortunately, owing to the large size of many commonly studied genomes (human, mouse and tobacco plant for example are all > 2.5 Gbp in size) [23–25], it is often still prohibitively expensive to conduct whole genome sequencing at high coverage. This creates a trade-off that negatively impacts the number of replicates that can be included and, therefore, it challenges the statistical power and the reproducibility of the studies [26, 27]. This is true in particular for DNA methylation, where differentially methylated regions (DMRs) are typically called by identifying changes as small as 10% and where $70-80\%$ of the reads of Whole Genome Bisulfite Sequencing (WGBS) methods contain little to no relevant information on the DNA methylation status [28].

To address these cost inefficiencies, many methods have been developed to reduce the number of genomic fragments that need to be sequenced for a given biological system [29–33]. These methods can be broadly split into those that positively select for genomic fragments of interest and those that deplete for fragments that are not of interest. Positive selection-based methods involve the sites of interest being enriched from the background. This usually occurs through pull-down of these sites via an antibody (e.g. anti-5mC antibody) [34], a recombinant binding protein (e.g. methyl-CpG-binding domains or MBD) [35], covalent biotin tagging [36], capture probes/baits for the sites of interest [37–39], array-based approaches (e.g. 27K, 450K and EPIC arrays in human) [5–7, 40] or PCR-based approaches [41–46]. These methods have many limitations, including enrichment biases, complex protocols and difficulties in quantification [29, 31].

Current evidence shows that depletion-based methods do not have enrichment biases, tend to be simpler and are more readily quantifiable [29, 32]. The most common depletion-based approaches use restriction enzymes to exploit the fact that the nucleotide composition in a given genome is non-random and that the fragment lengths produced from a given digestion will thus reflect this [47–51]. In the case of 5-methylcytosine (5mC), the most common depletion-based method is Reduced Representation Bisulfite Sequencing (RRBS) using the methylation-insensitive restriction enzyme MspI (with the recognition sequence C|CGG) [52, 53], although enzymes such as BglII [54], XmaI [55], Taq$^{\alpha}$I [56, 57], MspJI [58] , ApeKI [59], HpyCH4IV or HpaII [60] have also been used. RRBS has proven extremely useful for cost-effective, global studies of DNA methylation [52, 56, 61, 62], capturing around 10% of CpG sites within mammalian genomes but with up to a 30-fold reduction in the number of fragments sequenced in comparison to WGBS [63].

In the context of epigenetic clocks, most studies have used methylation arrays in humans [64–66] and MspI-based RRBS in mice, dogs and wolves [62, 67–70]. The utility of the MspI-based RRBS approach is limited to a specific subset of CpG sites in the genome, mainly found within CpG islands and promoters [52]. Nevertheless, it is known that many age-related changes in the methylome occur in other genomic regions (such as enhancers) [71–73], and current technologies could be biasing our discoveries. Furthermore, epigenetic clocks could be used in the near future to perform high-throughput screening of anti-ageing drugs or employed as ageing biomarkers in clinical trials [74]. However, the current assay costs could preclude the use of epigenetic clocks in this context.

Given that restriction enzyme-based approaches are versatile and simple, we developed a new computational method called customised Reduced Representation Bisulfite Sequencing (cuRRBS), which allows researchers to optimise the RRBS protocol for a specific experiment. cuRRBS generalises the problem of genomic enrichment with restriction enzymes by allowing the user to define both the genome and the particular sites of interest, before outputting the optimal enzyme combinations and size ranges to target these sites. In addition, cuRRBS provides the user with a variety of metrics to compare the various suggested protocols, including an estimate of the fold-reduction in sequencing costs compared to WGBS and a robustness value to assess the impact of experimental error in the size selection step.

Here, we have tested the enrichment ability of cuRRBS in several biological systems (including the Horvath epigenetic clock), with sites in both CpG and CHG contexts and multiple species, to showcase the generalisability and utility of the software [65, 75–80]. In addition, we take advantage of two recently published independent RRBS datasets to demonstrate the accuracy of the software predictions in both single and double enzyme

experimental settings [55, 57]. We hope that cuRRBS will be useful as a tool for designing cost-effective, genome-wide studies in the future, to help in the development of new epigenetic-based predictors and to validate previous results from whole genome approaches in a simple, cheap and timely fashion.

4.2 Restriction enzyme digestion as a tool for genomic enrichment

Restriction enzymes represent an incredibly effective tool for the enrichment of certain sites of interest in a genome. This is possible due to the wide variety of motifs that commercially-available restriction enzymes can recognise (Fig. 4.1) combined with the non-random nature of the genome composition itself. Fig. 4.1 highlights that this motif diversity is driven both by the sequence composition (GC content) and the length of the recognition sequence. Thus, different restriction enzymes will generate different fragment length distributions, dependent upon how frequently their recognition site is present in a given genome (Fig. 4.2a, Fig. S3.1).

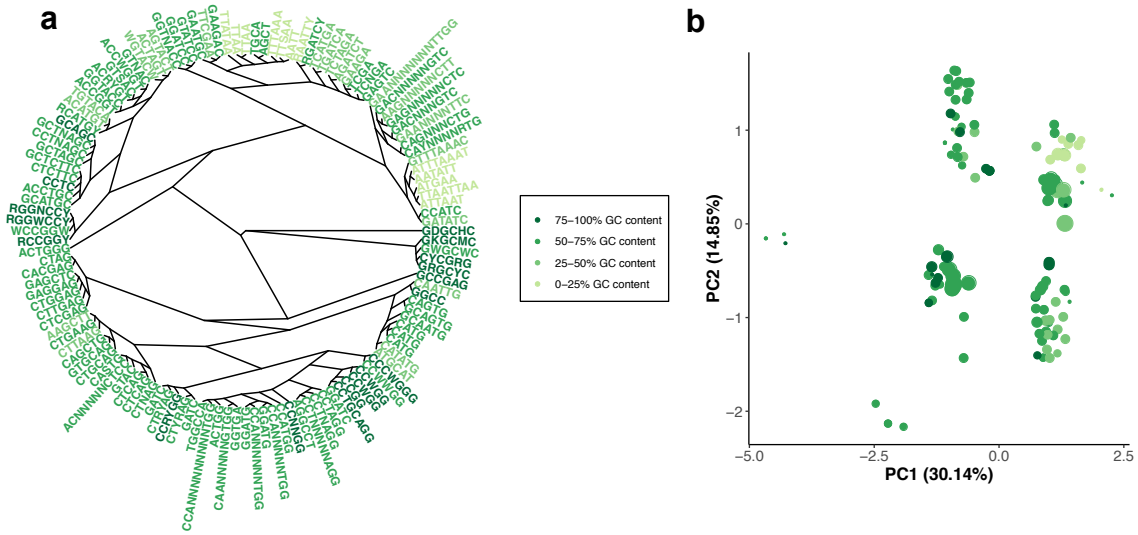


Fig. 4.1 The landscape of restriction enzyme motifs. **a.** Phylogenetic analysis of the motifs that are recognised by the different commercially-available restriction enzymes which are insensitive to CpG methylation. Each sequence represents a different isoschizomer family considered in this study. A neighbour-joining method was used to construct the tree. Motifs with different GC content are shown with different colours. **b.** Principal component analysis (PCA) performed on the matrix of pairwise distances from the aligned motifs. Each circle represents a different motif. The coordinates of the different motifs on the first two principal components are plotted on the x- and y-axes. Motifs with different GC content are shown with different colours (same as in a.) and the motif length is represented by the diameter of the circle.

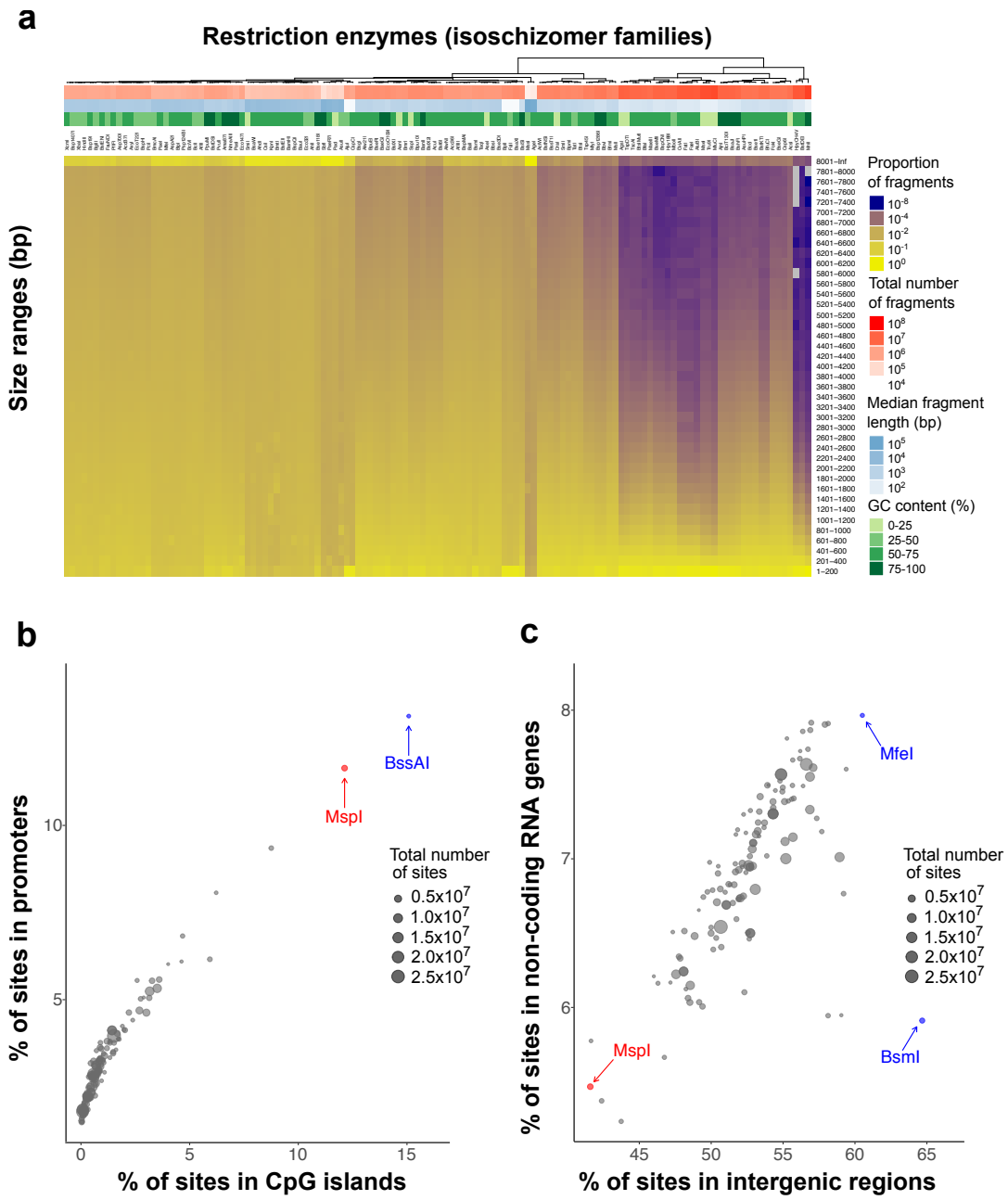


Fig. 4.2 Restriction enzyme digestion as a tool for genomic enrichment. **a.** Heatmap showing the fragment length distributions generated by different restriction enzymes in the human genome (hg38). Each column represents the distribution for an isoschizomer family of restriction enzymes that contains at least one member which is methylation-insensitive in a CpG context. The distributions are binned in size ranges of 200 bp, ordered as they would appear in an electrophoretic gel. Additional row annotations on top of the heatmap contain information regarding the total number of fragments (in red) and the median fragment length (in blue) produced by each in silico digestion, together with the GC content of the recognition motif in the isoschizomer family (in green). Legend is displayed on the right hand side. **b.** Scatterplot showing the percentage of cleavage sites from different restriction enzymes that overlaps with CpG islands (x-axis) and promoters (y-axis) in the human genome (hg38). The size of the circles represents the total number of cleavage sites generated by each enzyme. The enzymes MspI and BssAI are highlighted in red and blue respectively. Legend is displayed on the right hand side. **c.** Scatterplot showing the percentage of cleavage sites from different restriction enzymes that overlaps with intergenic regions (x-axis) and non-coding RNA genes (y-axis) in the human genome (hg38). The size of the circles represents the total number of cleavage sites generated by each enzyme. The enzyme MspI is highlighted in red. The enzymes BsmI and MfeI are both highlighted in blue. Legend is displayed on the right hand side.

In DNA methylation studies the most common application is the use of MspI (cutting at C|CGG) in RRBS (Reduced Representation Bisulfite Sequencing), which is used to enrich for CG dinucleotides (CpGs) contained in promoters and CpG islands [52] (Fig. 4.2b). However, in many cases, MspI is by no means the most effective restriction enzyme that could be used. For instance, MspI would be a poor restriction enzyme to choose for the enrichment of CpGs found in intergenic regions or non-coding RNA genes, which would be far better enriched for using BsmI or MfeI respectively (Fig. 4.2c). In fact, it turns out that across many genomic features MspI is rarely the most optimal methylation-insensitive restriction enzyme (Fig. S3.2).

Previous studies have tested the potential of other restriction enzymes and enzyme combinations to expand the range of CpG sites that can be targeted in a genome [47, 49–51, 55, 56, 59, 60]. However, to our knowledge, there is currently no computational method that systematically explores the capacity of all commercially-available restriction enzymes to generate 'personalised' reduced-representations of the genome whilst minimising the experimental cost (Fig. S3.3).

4.3 cuRRBS: customised Reduced Representation Bisulfite Sequencing

We have developed a novel computational method (cuRRBS) that determines the optimal combination of restriction enzymes and size range to enrich for any given set of sites of interest in any genome. In other words, by modifying two of the steps in the original RRBS protocol (Fig. 4.3a), cuRRBS generalises RRBS.

The software takes as input the genomic coordinates that the user wants to target (Fig. 4.3b, Fig. S3.4a). Afterwards, cuRRBS assesses *in silico* the potential of all single enzymes and double-enzyme combinations to enrich for the sites of interest using the following variables:

- *NF*, which reflects the theoretical number of genomic fragments that will be sequenced after the size selection step (i.e. those whose lengths after the *in silico* digestion are within the size range). Assuming that the sequencing cost is proportional to *NF*, cuRRBS attempts to minimise this value.
- *Score*, which reflects the theoretical number of sites of interest that will be sequenced after the size selection step. cuRRBS attempts to maximise this value, which can be calculated as:

$$Score = \sum_{i=1}^{n} w_i \cdot \gamma_i \tag{4.1}$$

where n is the total number of sites of interest, w_i is the weight of the ith site of interest and γ_i is 1 if the ith site would be theoretically sequenced (i.e. present in a size selected fragment and $\leq$ *read length* base pairs away from one of the ends of the fragment) and 0 otherwise.

- *Enrichment Value* (*EV*), which combines both *NF* and *Score* into a single number. The objective of cuRRBS is to minimise *EV*, which can be calculated as:

$$EV = -log_{10}\left(\frac{Score}{NF} \cdot \frac{n}{max\_Score}\right) \tag{4.2}$$

where *max_Score* is the *Score* obtained if all the sites of interest were sequenced.

The *NF* and *Score* variables are positively correlated with one another, such that the more genomic fragments sequenced, the more sites of interest are likely to be contained within the reduced representation (Fig. 4.3c, Fig. S3.4b). However, this relationship disappears at higher *NF* values, where the *Score* variable becomes saturated such that any additional fragments sequenced will result in a reduction in the overall enrichment of the sites of interest. This *Score* saturation at high *NF* is mainly due to additional sites of interest being buried within long fragments that will not be sequenced due to limitations in the read length (cuRRBS parameter –r, see Table 4.1). For a given enzyme or enzyme combination, the *NF* and the *Score* variables depend on the *size range* chosen, since only the genomic fragments within the size range will be present in the reduced representation of the genome.

cuRRBS requires that the user sets *thresholds* for the maximum *NF* (i.e. minimum *CRF*, see below) and minimum *Score* that would be acceptable for a given application (Fig. 4.3b, Fig. S3.4a). These *thresholds* allow cuRRBS to search through all possible *size ranges* for a given enzyme or enzyme combination and to find the one that minimises the *Enrichment Value* (*EV*). cuRRBS repeats this procedure for every single enzyme and enzyme combination and reports those with the best hits (i.e. those with the lowest *EV*s) (Fig. S3.4a).

The output file contains the best scoring enzymes with their correspondent size ranges and some other useful variables for each one of the hits, such as:

- *Cost Reduction Factor* (*CRF*), which estimates the theoretical fold-reduction in sequencing costs for the cuRRBS protocol when compared to Whole Genome Bisulfite Sequencing (WGBS). The *CRF* for a given cuRRBS protocol can be calculated as:

$$CRF = \frac{NF_{ref}}{NF} = \frac{g/r}{NF} \tag{4.3}$$

where NF_{ref} is the estimated number of fragments that would be sequenced in a WGBS experiment, that can be roughly calculated as the genome size (g) divided by the read length (r).

- *Robustness* (*R*). This assesses how much the cuRRBS prediction varies if a slightly different size range is used (Fig. 4.3d). The results for robust enzymes will not be greatly affected as a consequence of experimental error during the size selection step. This will help the user to make an informed decision on which enzyme combination to choose for the system of interest (Fig. S3.4c). The *robustness* of a given enzyme (combination) is calculated as:

$$R = e^{-\theta} \tag{4.4}$$

with

$$\theta = \frac{\sum_{x\in\{a-\delta,a,a+\delta\}} \sum_{y\in\{b-\delta,b,b+\delta\}} |EV_{x,y} - EV_{a,b}|}{EV_{a,b}} \tag{4.5}$$

where $EV_{a,b}$ is the EV for the optimal size range (a: lower limit in size range, b: breadth) and δ is the experimental error (in bp) that is assumed during the size selection step. The *robustness* will take values in the interval $(0,1]$, with higher values identifying robust cuRRBS protocols.

4.4 Running cuRRBS in different biological systems

cuRRBS provides a way to effectively interrogate DNA methylation in any biological system (including the CpG sites that constitute different epigenetic clocks) for which the reference

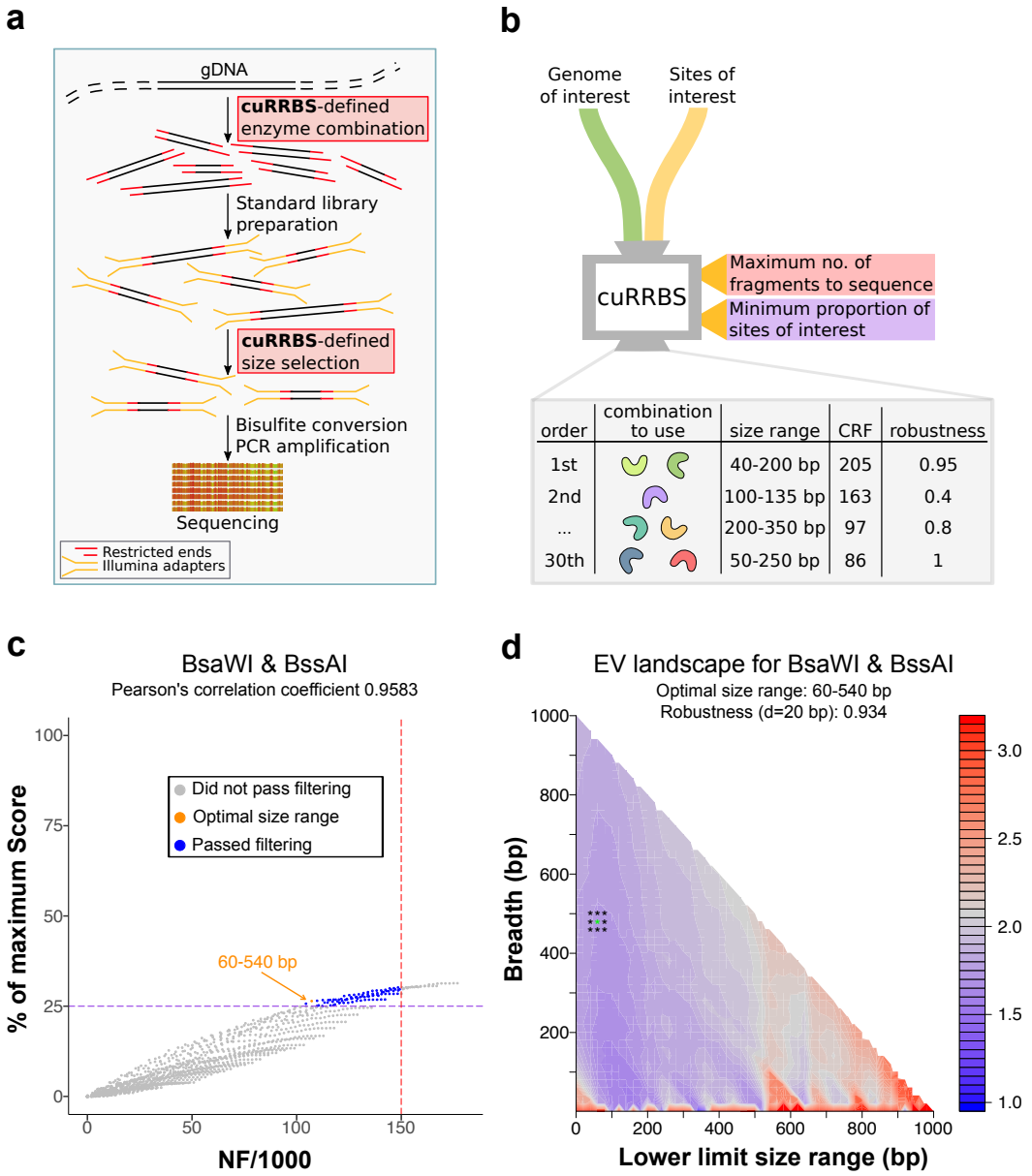


Fig. 4.3 cuRRBS overview. **a.** Outline of an RRBS protocol. Highlighted are the two steps that would be modified according to the output produced by cuRRBS (i.e. the restriction enzymes used for the genomic digestion and the size selection). Legend is displayed on the bottom left. **b.** Schematic of cuRRBS. Highlighted are the two main inputs required for the software and the two *thresholds* that the user has to define (red and purple tags). The default output for cuRRBS is a table containing the top hits (restriction enzyme combination and size range) along with additional information that might be useful to the user (such as *Cost Reduction Factor* and *robustness*). **c.** Scatterplot showing the trade-off between the number of fragments (*NF*) and the *Score* for the best enzyme combination (BsaWI & BssAI) that targets the CpGs present in the human placental-specific imprinted regions [75]. *NF* is divided by 1000 for visualization purposes. Each point represents a different *size range*. Shown in dark blue and grey are the size ranges that would and would not pass filtering respectively. Shown in orange is the optimal size range in the filtered search space. The dotted lines depict the *thresholds* that need to be specified by the user (red: maximum *NF*; purple: minimum percentage of the maximum *Score*). In this mock example we specified an *NF threshold* of 150000 fragments and a *Score threshold* of 25% of the maximum *Score*. Legend is displayed below the plot title. **d.** Contour plot that depicts how the *robustness* (*R*) variable is calculated for the optimal enzyme combination (BsaWI & BssAI; size range: 60-540 bp) that targets the CpGs present in the human placental-specific imprinted regions [75]. *Enrichment values* (*EVs*) are calculated for all possible size ranges in order to create an *EV* 'landscape'. In this landscape, cuRRBS finds the size range with the lowest *EV* that still satisfies the *thresholds* (asterisk in green). Afterwards, cuRRBS samples *EV*s around the optimum (asterisks in black). The points that are sampled depend on the experimental error (in this case, $\delta = 20$ bp). A high *robustness* value means that the sampled *EV*s do not change a lot when compared to the optimum, which implies that cuRRBS prediction will not be greatly affected by experimental errors during the size selection step.

genome is available. Besides reducing the cost for organisms currently under intensive study (e.g. human, mouse), cuRRBS opens the door to the cost-effective study of DNA methylation in species with large genomes or where DNA methylation in non-CpG contexts is common, such as plants [81], which currently lack an MspI-based RRBS protocol, owing to the enzyme's CHG methylation sensitivity [82].

We decided to test the ability of cuRRBS to enrich for genomic sites that have important functional roles in different systems. Some of the systems that we tested *in silico* include genomic regions whose methylation status is important during cellular reprogramming [76], the Horvath human epigenetic clock [65], transcription factor binding sites that are affected by DNA methylation [78, 80], imprinted loci [75], CpGs found in the exon-intron boundaries [79] and CHG sites that are differentially methylated between different arabidopsis accessions [77] (Fig. S3.5). For these in silico systems we chose to run the software with the threshold set to 25% of the maximum *Score*.

In all cases, cuRRBS is able to dramatically reduce the cost associated with the sequencing by several orders of magnitude compared to WGBS, which is assessed using the *Cost Reduction Factor* (*CRF*) (Fig. 4.4). In addition, for cases where a comparison to MspI-based RRBS could be made, cuRRBS is able to improve the *CRF*, again, by orders of magnitude. As an example, for the placental-specific imprints, the sequencing costs are reduced by approximately 400-fold when compared to WGBS and by 12.5-fold when compared to the traditional MspI-based RRBS.

Furthermore, we have also observed that many of the top hits reported by cuRRBS are digestions of two restriction enzymes (Fig. S3.5), highlighting the combinatorial power of restriction enzymes to produce optimal reduced representations of the genome [49]. Excitingly, we are able to show that using cuRRBS it is possible to assay a far larger number of target sites, in a far simpler experimental design than would normally be achieved using amplicon-based bisulfite sequencing.

4.5 Experimental validation of cuRRBS

To assess in an unbiased manner how well predictions from cuRRBS perform in an experimental setting, we employed two independent non-canonical RRBS datasets: one generated from a single enzyme (XmaI) and the other from a combination of two restriction enzymes (MspI and Taq$^{\alpha}$I) [55, 57]. By evaluating the predictive power of cuRRBS in these two

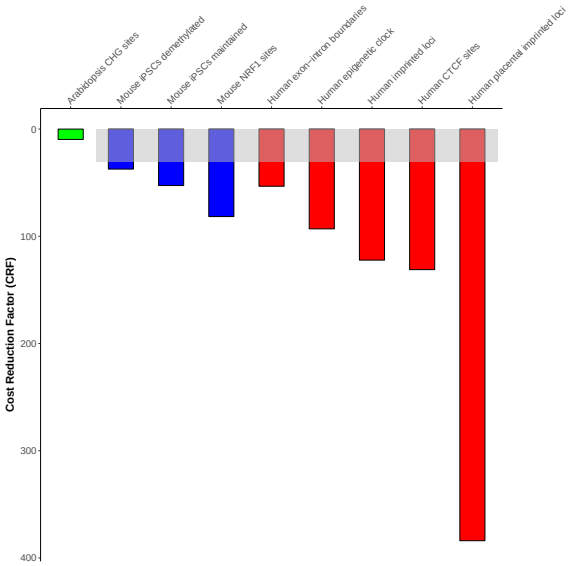


Fig. 4.4 Running cuRRBS in different biological systems. Barplot showing the values for the *Cost Reduction Factor* (*CRF*) in the different biological systems that were tested (see Fig. S3.5) [65, 75–80]. The colours in the bars represent the different species interrogated (green: *Arabidopsis thaliana*, blue: *Mus musculus*, red: *Homo sapiens*). The *CRF* for the traditional RRBS protocol (MspI in the human genome, using a bead size selection step of 20-800 bp, $CRF = 30.65$) is displayed as a grey area, which is not compared with the *A. thaliana* system (since MspI is sensitive to CHG methylation).

datasets, we were able to observe cuRRBS' performance in both single and double enzyme contexts and across different genomes.

To test the accuracy of cuRRBS predictions in the context of a single enzyme digestion, we utilised the non-canonical RRBS dataset generated from human DNA using the restriction enzyme XmaI [55]. This dataset was previously used to show that XmaI could enrich for CpG islands (CGIs), while reducing the overall sequencing cost relative to MspI, making the protocol more cost-effective. To validate cuRRBS using this system, we therefore chose to enrich for all CpG sites that overlapped with a CGI (CGI-CpGs) in the human genome using a predetermined theoretical size range equivalent to the 'reproducible library fragment lengths' reported in [55] (i.e. 90-185 bp). cuRRBS predicted with high accuracy the CpG sites that were observed in the experimental XmaI-RRBS dataset (Fig. 4.5a). In particular, only a small proportion of the total number of CGI-CpGs should be theoretically sequenced (102253 out of 2164614), and this was indeed the case (Fig. 4.5a). Furthermore, upon filtering out sites with low depth of coverage, which commonly represent noise in RRBS datasets, the sensitivity increased up to approximately 80%. Importantly, the specificity remained constant at almost 100% independent of the threshold set for depth of coverage (Fig. 4.5b). Thus,

cuRRBS produces a prediction that is relatively conservative, as highlighted by the low numbers of false positives (Fig. 4.5a), at the expense of a small decrease in sensitivity.

Interestingly, the original theoretical size range that the study was aiming for (110-200 bp) was slightly different to the one achieved in the actual experiments (90-185 bp) [55]. We ran cuRRBS using the original size range target and obtained slightly worse results for the sensitivity but not the specificity of the prediction (Fig. S3.6). This demonstrates that the correct execution of the size selection step during the experimental protocol is key for obtaining the sites predicted by cuRRBS and highlights the importance of the *robustness* variable as part of the cuRRBS output in order to judge the consequences of these experimental errors.

To test the accuracy of cuRRBS predictions in the context of a double enzyme digestion, we utilised the non-canonical RRBS dataset generated from mouse DNA using the restriction enzymes MspI and Taq$^{\alpha}$I [57]. To compare the accuracy of cuRRBS prediction in this double enzyme system to that of the XmaI-RRBS system, we again ran cuRRBS for CGI-CpGs, this time in the mouse genome with a theoretical size range of 80-160 bp [57]. cuRRBS predicted with high accuracy the CpG sites that were observed in this double enzyme experiment (Fig. 4.5c). In addition, the results for sensitivity and specificity were very similar to the ones reported for the XmaI-RRBS dataset (Fig. 4.5d). Therefore, we conclude that cuRRBS produces robust predictions for the sites of interest that will be sequenced in RRBS protocols both for single and double enzyme combinations independent of the genome under study.

Lastly, the number of fragments that were theoretically recoverable in each of our experimental systems ranged from $NF = 12780$ (for XmaI) to $NF = 331058$ (for MspI and Taq$^{\alpha}$I). This represents approximately a 30-fold difference in the number of recoverable fragments and demonstrates that cuRRBS predictions, even for low NF values, are experimentally feasible. Importantly, in the nine theoretical examples that we report (Fig. S3.5), the number of fragments required by each cuRRBS protocol ranges from 107248 to 974050. Thus, the number of fragments required to achieve the stated CRF comfortably exceeds the minimum experimentally validated NF value (>8-fold).

4.6 Conclusions and future directions

cuRRBS provides a new framework that allows the user to optimise RRBS for the biological system of interest by using novel combinations of restriction enzymes. Therefore, cuRRBS makes the study of DNA methylation more affordable across all species for which genomic

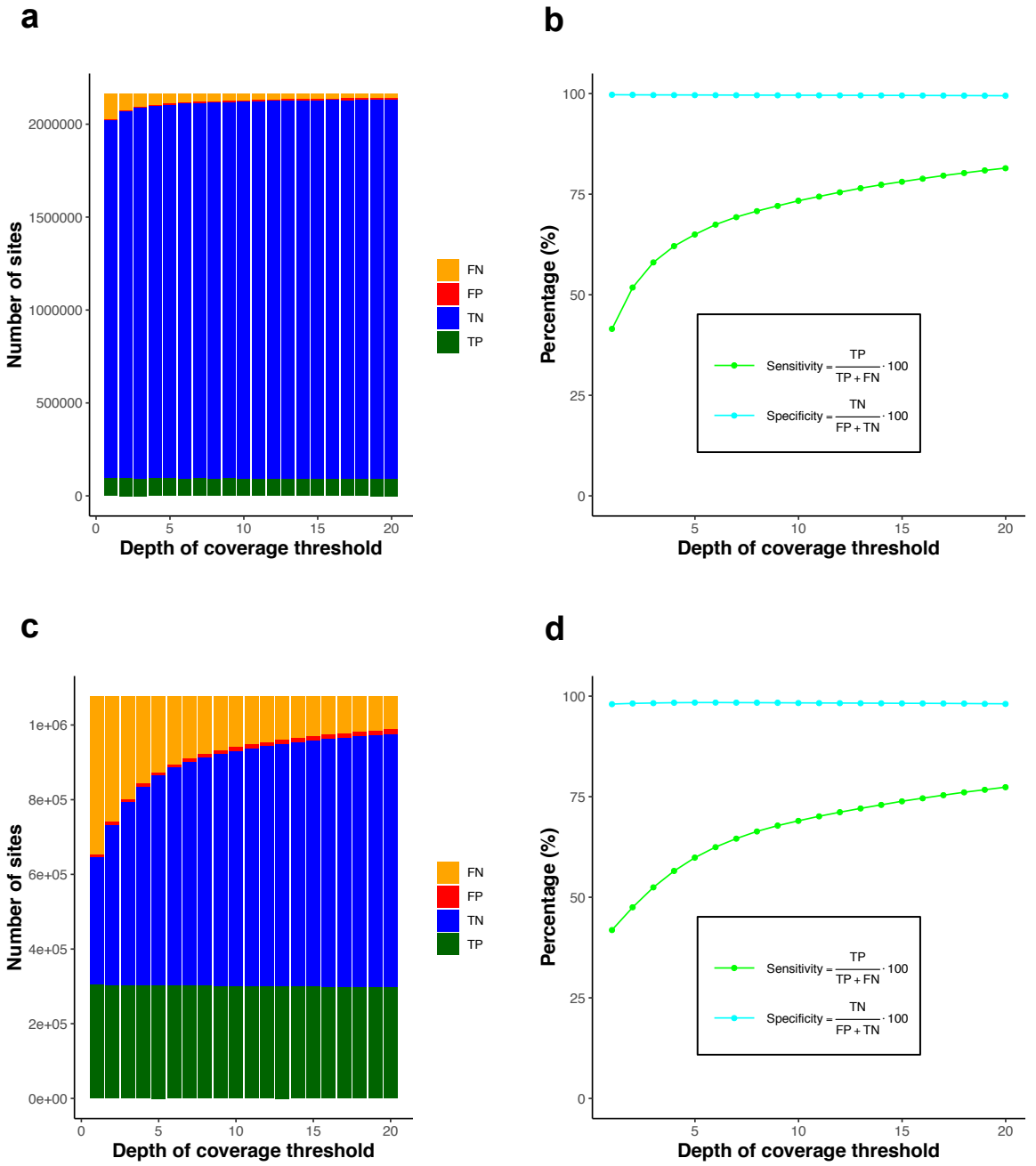


Fig. 4.5 Experimental validation of cuRRBS. **a.** Barplots showing the number of true positives (TP, in green), true negatives (TN, in blue), false positives (FP, in red) and false negatives (FN, in orange) when comparing cuRRBS theoretical prediction with the actual XmaI-RRBS experimental data [55]. The number of sites in each category is calculated for different thresholds in the depth of coverage (number of reads covering a CpG site as reported by Bismark). cuRRBS prediction for the CpG sites in human CpG islands was obtained enforcing a theoretical size range of 90-185 bp and running the software for XmaI with all the default parameters (with a *read length* of 200 bp). Legend is displayed on the right hand side. **b.** Plot showing values of cuRRBS sensitivity (in light green) and specificity (in cyan) as a function of the depth of coverage threshold employed to filter the experimental data [55]. The number of true positives (TP), true negatives (TN), false positives (FP) and false negatives (FN) are the same as in a. Legend is displayed below the plot curves. **c.** Same as in a. but for the MspI&Taq$^{\alpha}$I-RRBS experimental data [57]. cuRRBS prediction for the CpG sites in mouse CpG islands was obtained enforcing a theoretical size range of 80-160 bp and running the software for MspI&Taq$^{\alpha}$I with all the default parameters (with a *read length* of 75 bp). **d.** Same as in b. but for the MspI&Taq$^{\alpha}$I-RRBS experimental data [57].

sequences are available. Furthermore, it can open the door to the design of future studies in a clinical context [56], which require cost-effective and robust protocols.

Currently, cuRRBS only considers combinations of up to two restriction enzymes. However, in the future, it would be possible to adapt the software to explore combinations that contain higher numbers of enzymes, which could theoretically allow targeting the sites of interest even more efficiently [49]. Moreover, there are several methods that are able to impute DNA methylation levels in sites that are not covered experimentally [83, 84]. These methods could expand the set of sites of interest that are finally measured by making use of the additional DNA methylation information that is retrieved in a cuRRBS experiment.

Finally, the potential of restriction enzymes to target different genomic coordinates is not limited to DNA methylation. As such, it would be conceivable for cuRRBS to be adapted to enrich for SNPs of interest [85, 86] or to optimise chromosome conformation capture techniques [87, 88]. By reducing the cost associated with sequencing, we believe that cuRRBS will help to democratise high-throughput genomic studies.

4.7 Additional methods

Restriction enzymes annotation

All the information regarding the commercially-available restriction enzymes that are used by cuRRBS was extracted from REBASE [89, 90]. Restriction enzymes were grouped in isoschizomer families (i.e. enzymes that recognise the same sequence and generate identical fragment length distributions) and each enzyme was manually annotated for different types of methylation-sensitivity (CpG, CHG, CHH). Only isoschizomer families that contained at least one methylation-insensitive enzyme were considered for the examples described here.

Genome assemblies and genomic annotation

All the analyses presented here were performed in the following genome assemblies: *Homo sapiens* (hg38), *Mus musculus* (mm10) and *Arabidopsis thaliana* (TAIR10). Scaffolds not assembled into the main chromosomes were discarded. Genomic annotation for the human genome (hg38) was obtained from GENCODE (v25, basic gene annotation) [91], with the exception of CpG islands (CGIs), which were extracted from the UCSC Genome Browser [92]. GC content and CpG content were calculated, around each restriction enzyme cleavage site, taking windows of ± 25 bp and ± 500 bp respectively. For each enzyme, the mean of

all cleavage sites was calculated to obtain the mean GC content and the mean CpG content. Intragenic regions were defined as those regions within ± 2.5 kb of a protein-coding gene, whilst the rest of the genome was considered to be intergenic. CpG shores were defined as regions 0 to 2 kb away from CGIs in both directions and CpG shelves as regions 2 to 4 kb away from CGIs in both directions [83]. Promoters were defined as encompassing a 3 kb region (2.5 kb upstream and 0.5 kb downstream of the TSS) relative to the TSS of all protein-coding transcripts in GENCODE, similar to the strategy used in Taher *et al.* [93]. Genomic annotation for the CGIs in the mouse genome (mm10) was also obtained from the UCSC Genome Browser [92]. All annotations were handled using the *pybedtools* library [94, 95].

Performing *in silico* digestions of a given genome

We used the *Restriction* package from Biopython v1.68 to digest the different genomes with the appropriate restriction enzymes *in silico* [96]. Only the first member of a given isoschizomer family (which contained at least one methylation-insensitive enzyme) was processed to avoid redundant computations. The output of the *in silico* digestions was stored (pre-computed files) and subsequently read by cuRRBS when needed to reduce the computational time (see 'cuRRBS heuristics and computational efficiency'). When assessing enzyme combinations, the information from the appropriate individual pre-computed files (i.e. the genomic coordinates where the enzyme theoretically cuts) were combined by the software to compute all the necessary variables.

cuRRBS' enzyme flexibility

To ensure the user has full control over the enzymes that cuRRBS will use to derive the desired enrichments, one of the inputs given to cuRRBS is an enzyme annotation file. This file contains the desired isoschizomer families that the user wishes to be tested by cuRRBS. In my GitHub repository we have already defined enzyme annotation files for enzymes that are methylation-insensitive in a CG context and in CG, CHG and CHH contexts [97]. However, it is also possible for the user to define a personalised set of enzymes by providing a self-generated annotation file. This can be useful, for instance, to reduce the chance of any star activity in the reported cuRRBS protocols.

In addition, the output file from cuRRBS contains, by default, 30 cuRRBS protocols that would enrich for the user's sites of interest. Therefore, the user can determine which enzyme combination and size range would be the simplest and most appropriate for the given

cuRRBS parameter (abbrev.)	Significance	Default	Range
Enzymes to check (-e)	Defines the enzymes (isoschizomer families) that cuRRBS will look at	-	-
Annotation for the sites of interest (-a)	Allows identification and weighting of the sites of interest	-	-
Read length (-r)	Defines the positions in the theoretical fragments that can be 'seen' after sequencing	-	30-300
Adapters size (-s)	Ensures correct experimental size selection	-	-
C_Score constant (-c)	Sets the minimum acceptable *Score*	-	0-1
Genome size (-g)	Needed to calculate the *CRF*	-	-
C_NF/1000 constant (-k)	Sets the minimum acceptable *CRF*	0.2	0-1
Experimental error (-d)	Sets the assumed experimental error (δ)	20	5-500
Size range breadth (-b)	Constrains the breadth of the size range	980	-
Output size (-t)	Defines the number of cuRRBS protocols the user can compare	30	-
Site IDs (-i)	Enables the identification of the recovered sites of interest	No	-

Table 4.1 Flexible user-defined cuRRBS parameters. This table details the flexible user-defined parameters that cuRRBS will accept as arguments. The cuRRBS parameter full name and command line abbreviation (in brackets) are provided alongside a simplified description of the significance of these arguments to the user. Where applicable, the defaults and ranges of these arguments are also detailed.

application. This provides the user with the opportunity to consider experimental factors that may complicate the protocol, such as buffer compatibility and whether consecutive digestions would be required.

Flexible user-defined cuRRBS parameters

cuRRBS contains a number of user-defined parameters to ensure the greatest possible flexibility and ease of use. A table of these parameters is provided to highlight the versatility that the user has and why such versatility is useful (Table 4.1).

cuRRBS heuristics and computational efficiency

cuRRBS employs several strategies to reduce the computational time needed in each run:

- Restriction enzymes are grouped in isoschizomer families. Since isoschizomers generate the same genomic digestions, only one member of each family needs to be processed.

- *In silico* digestions are read from pre-computed files. Digesting the genomes would be a limiting factor in the cuRRBS pipeline. The user can download the pre-computed files [97] and the information that they contain is read every time that an enzyme needs to be assessed.
- The number of size ranges that are sampled is minimised. Since the experimental size selection step is generally imperfect, size ranges are sampled with a sliding window whose 'resolution' is equivalent to the experimental error specified by the user.
- Parallelization. cuRRBS can use several cores to decrease the CPU time.

Moreover, we have observed that, in many enzyme combinations, one of the enzymes is providing most of the enrichment for the sites of interest, while the second one complements the targeting. Therefore, it would be possible to implement a 'heuristic' mode, where only those enzymes that perform well individually are used as 'seeds' to construct combinations (as opposed to the current implementation, where all the enzyme combinations are checked exhaustively). This could further reduce the computational time, especially if combinations of more than two enzymes were being evaluated.

The CPU time required by cuRRBS depends on several parameters, including the number of enzymes checked, the experimental error, the number of sites of interest or the genome size (Fig. S3.7). The RAM used will be approximately equal to the size of the pre-computed files that are read by the software. A standard cuRRBS run (e.g. for a few thousand sites of interest in the human genome, checking 128 CpG methylation-insensitive isoschizomer families) takes around 0.5-1 hours and uses around 4 GB RAM, which allows the user to easily run it on a dual-core laptop or desktop computer.

Obtaining the sites of interest for different biological systems

We have tested *in silico* the ability of cuRRBS to enrich for the sites of interest in a selection of different biological systems where DNA methylation has an important functional role. In some of these systems, described below, previous analysis was performed in order to obtain the genomic coordinates for the sites:

- Exon-intron boundaries in human. Exons and introns were obtained from protein-coding genes using GENCODE annotation data. Those CpG sites that were found within ± 5 bp of a canonical splice site (5'-GT, 3'-AG) were selected.
- Epigenetic clock in human. These sites were obtained from the Horvath epigenetic clock [65] and were lifted over to hg38 [98] before running cuRRBS.

- Canonical and placental imprints in human. These loci were obtained from Hanna *et al.* [75]. The sites were lifted over to hg38 [98] and the CpG sites were then extracted for the analysis.

- CTCF binding sites in human. We obtained the CpG sites that overlap with *in vivo* CTCF binding sites. Peaks from sites that seem to be affected by methylation (upregulated, reactivated) were kindly provided by Dr. M. T. Maurano [78]. We scanned the peaks for high-scoring motifs according to the CTCF JASPAR model [99]. Finally, we extracted those CpGs that were found in positions 5 and 15 of the motif, whose methylation status is supposed to influence the binding of the transcription factor [78].

- Induced pluripotent stem cells (iPSCs) demethylated and maintained sites in mouse. These were obtained by comparing mouse embryonic fibroblasts (MEFs) to iPSCs as described previously [76], with an additional filter for magnitude of methylation change (>50% methylation change).

- NRF1 binding sites in mouse. We obtained the CpG sites that overlap with *in vivo* NRF1 binding sites in mouse. ChIP-seq data was processed as described in the original publication [80], where peaks were called using Peakzilla [100]. We took as our final set of peaks the overlap between the two TKO replicates. Next, we scanned the peaks for high-scoring motifs according to the NRF1 JASPAR model [99]. Finally, we extracted those CpGs that were found in positions 2 and 8 of the motif, whose methylation status is supposed to influence the binding of the transcription factor [99].

- CHG sites in *Arabidopsis thaliana*. Non-CpG DMRs arising from the epigenomic diversity between *Arabidopsis thaliana* accessions were obtained from Kawakatsu *et al.* [77]. The coordinates for C sites in non-CpG context were extracted.

In all the cases the sites were equally weighted ($w_i = 1$), with the exception of the human epigenetic clock system, where the sites were assigned the absolute value of the weights in the linear model [65]. All the site annotation files can be found in my GitHub repository [97]

Running cuRRBS for the different biological systems

cuRRBS was run in the different systems described above using the default parameters ($k = 0.2$, $d = 20$, $b = 980$, $t = 30$), for a *read length* (r) of 75 bp and a *Score threshold* (c) of 0.25. In the mouse and human examples we considered 128 isoschizomer families that contained enzymes that were not sensitive to CpG methylation. In the case of *Arabidopsis*

thaliana we used 28 isoschizomer families that contained enzymes that were not sensitive to 5mC in any context (CG, CHG, CHH).

Mapping of RRBS samples

XmaI-RRBS data generated on the Ion Torrent platform [55] and MspI&Taq$^{\alpha}$I -RRBS data generated on the Illumina HiSeq platform [57] were quality trimmed using Trim Galore (www.bioinformatics.babraham.ac.uk/projects/trim_galore/) and had base pairs removed from the 3' end to avoid including filled-in nucleotides with artificial methylation states (the filled-in XmaI, MspI and Taq$^{\alpha}$I cut sites include the nucleotide sequence CCGG, CG and CG respectively). The data was then mapped to the human genome (for XmaI data, parameters: –non_directional) or the mouse genome (for MspI&Taq$^{\alpha}$I data, parameters: –directional) using Bismark (0.18.0) [101]. In each of the two cases data from different experiments or replicates was merged into the same FASTQ file prior to quality trimming.

Estimating cuRRBS' sensitivity and specificity

We assessed the performance of cuRRBS predictions in two independent experimental datasets [55, 57] (see 'Experimental validation of cuRRBS'). We ran cuRRBS fixing the theoretical size ranges tested to the ones reported in the publications [55, 57] and we used as our sites of interest the CpGs that overlapped with CpG islands (CGI-CpGs) in the human [55] and the mouse genomes [57] respectively. From the cuRRBS output files we recovered the IDs of the sites that should be theoretically sequenced. Moreover, using the experimental RRBS data [55, 57], we could obtain the IDs of the sites that were actually sequenced (filtered by a given depth of coverage threshold). Afterwards, we calculated the following variables for each one of the datasets:

- True positives (TP): number of CGI-CpGs that cuRRBS predicted to be sequenced and were indeed found in the RRBS data.
- True negatives (TN): number of CGI-CpGs that cuRRBS predicted to be absent and were not found in the RRBS data.
- False positives (FP): number of CGI-CpGs that cuRRBS predicted to be sequenced but were not found in the RRBS data.
- False negatives (FN): number of CGI-CpGs that cuRRBS predicted to be absent but were found in the RRBS data.

Finally, we estimated the sensitivity and specificity, for a given dataset, as follows:

$$Sensitivity = \frac{TP}{TP+FN} \cdot 100 \tag{4.6}$$

$$Specificity = \frac{TN}{FP+TN} \cdot 100 \tag{4.7}$$

Software availability

cuRRBS and its documentation are freely distributed under GNU General Public License v3.0 and can be accessed in my GitHub repository [97].

Appendix

Supplementary figures and tables

S.1 Statistical aspects of the epigenetic clock

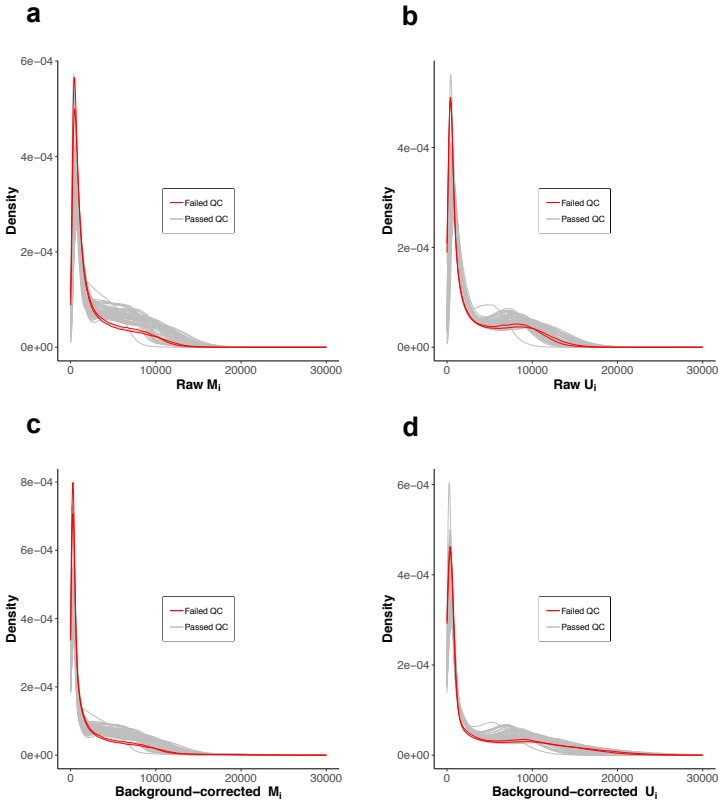


Fig. S1.1 Effects of *noob* background correction on the array flurescence intensities. Distributions of the array fluorescence intensities for the **a.** methylated signals (M_i) before background correction; **b.** unmethylated signals (U_i) before background correction; **c.** methylated signals (M_i) after background correction and **d.** unmethylated signals (U_i) after background correction. Each curve represents a DNA methylation sample from the GSE41273 batch. In grey: 51 samples that passed quality control (QC). In red: 2 samples that failed QC.

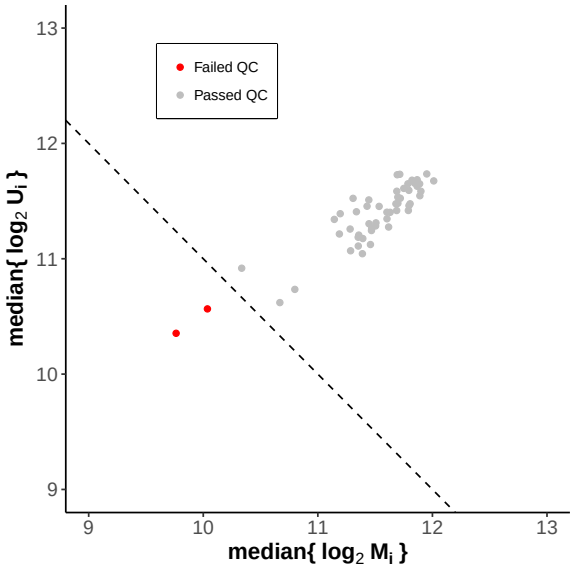


Fig. S1.2 Quality control (QC) strategy to identify outlier samples, according to their global intensity values, in the GSE41273 batch. Those samples with low median intensity values (see criteria in the main text) were discarded from downstream analyses (2/53, in red). Each sample is represented by one point. The dashed line represents the intensity threshold. M_i and U_i represent the background-corrected methylated and unmethylated intensity measurements for all the 450K array probes in a given sample.

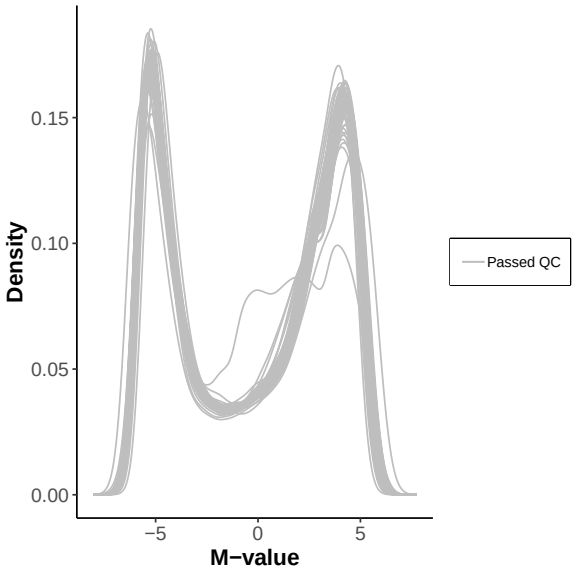


Fig. S1.3 M-value distributions in the samples of the GSE41273 batch, after all the pre-processing steps have been carried out (background correction, quality control, probe filtering and BMIQ normalisation). M-values were calculated applying the logistic transformation to the β-values, as described in Du *et al.* [16]. Each curve represents a different sample.

S.2 Biological aspects of the epigenetic clock

S.3 Technological aspects of epigenetic clocks

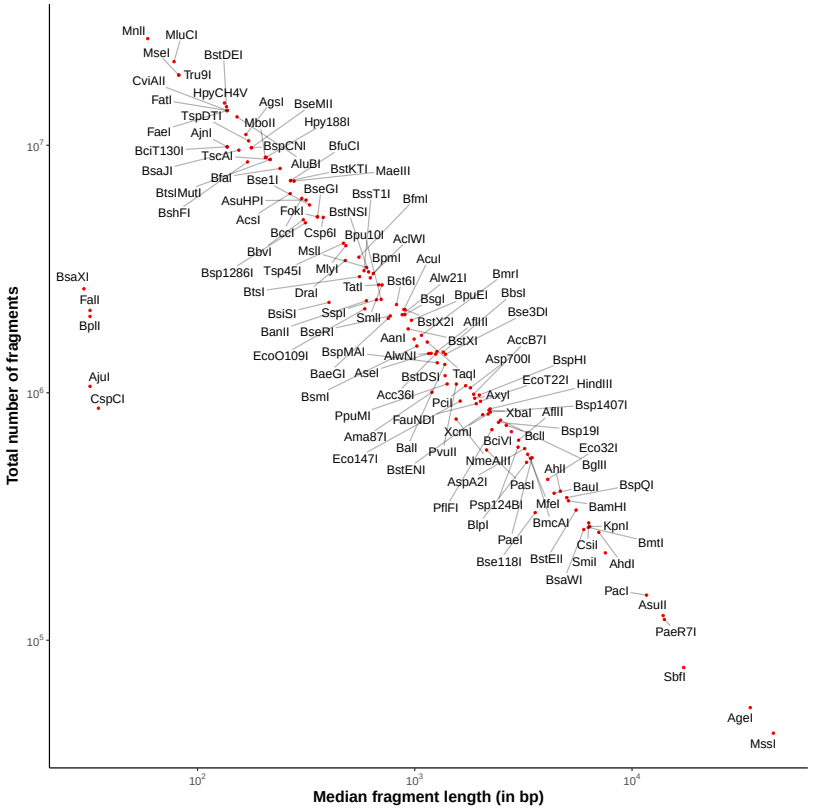


Fig. S3.1 Scatterplot which summarises the fragment length distributions for the same isoschizomer families portrayed in Fig 4.2a. The red dots represent the actual values of median fragment length and total number of fragments for each family. The black lines assign each name label to the correspondent red point for visualization purposes.

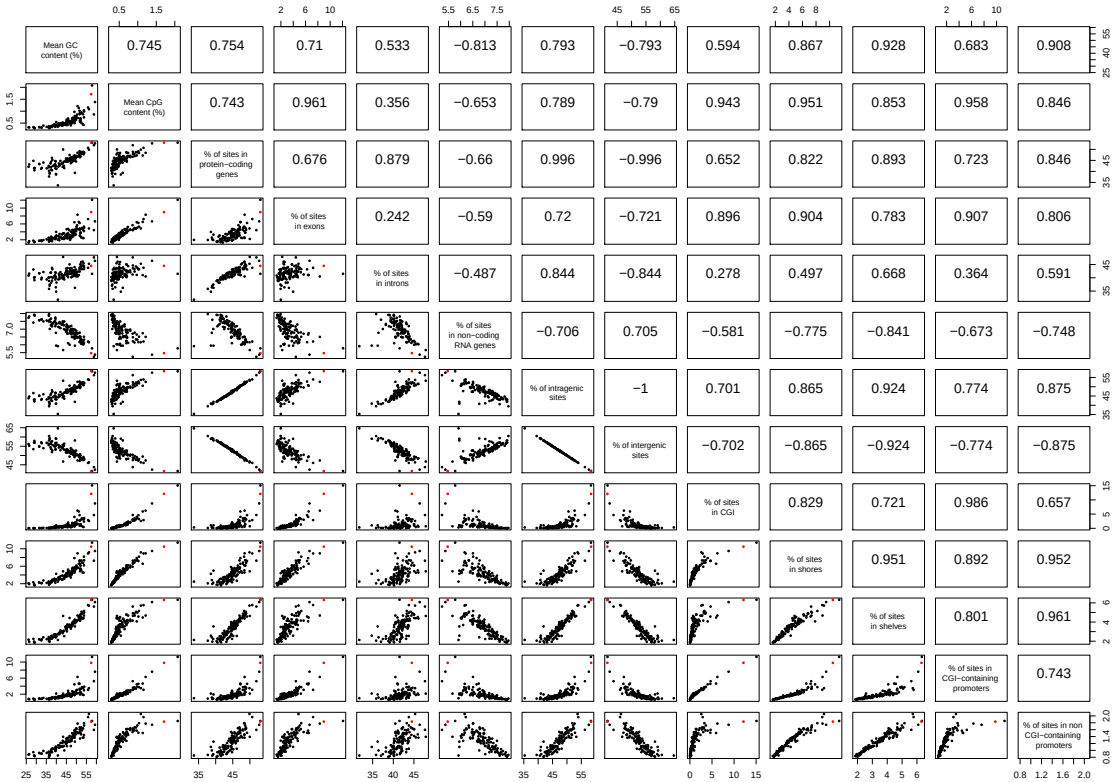


Fig. S3.2 Matrix of scatterplots showing the percentages of cleavage sites from different restriction enzymes that overlap with several genomic features (listed on the diagonal) in the human genome (hg38). The red dot in each scatterplot represents the values for MspI. The numbers above the diagonal are the Pearson correlation coefficients between all the possible pairs of genomic features.

First author(s)	Title	Date	Single enzymes checked	Double enzymes checked	Size ranges interrogated	Genomic regions targeted	Organism(s)	Read lengths tested	For sequencing	Code available
Cedar H	Direct detection of methylated cytosine in DNA by use of the restriction enzyme MspI	1979	YES	NO	NA	NA	*Neurospora crassa*, herpes virus, fly, bovine	NA	N	N
Yu L	A NotI–EcoRV promoter library for studies of genetic and epigenetic alterations in mouse models of human malignancies	2004	YES	YES	NA	CpG islands, protein-coding genes	Human (hg16), mouse (mm4)	NA	Y	N
Wang J and Xia Y	Double restriction-enzyme digestion improves the coverage and accuracy of genome-wide CpG methylation profiling by reduced representation bisulfite sequencing	2013	YES	YES	2	Increase CpG coverage genome-wide	Human (hg18), mouse(mm9)	50 bp PE, 90 bp PE	Y	N
Bystrykh L	A combinatorial approach to the restriction of a mouse genome	2013	YES	YES	NA	NA	Mouse (mm10)	NA	N	N
Martinez-Arguelles DB	In silico analysis identifies novel restriction enzyme combinations that expand reduced representation bisulfite sequencing CpG coverage	2014	YES	YES	1	Increase CpG coverage genome-wide	Human (hg38), mouse (mm10), rat (NCBI build 4.2)	50 bp PE	Y	N
Lee YK and Jin S	Improved reduced representation bisulfite sequencing for epigenomic profiling of clinical samples	2014	YES	YES	1	Increase CpG coverage genome-wide	Human (hg19)	36 bp PE	Y	N
Kirschner SA	Focussing reduced representation CpG sequencing through judicious restriction enzyme choice	2016	YES	YES	2	Increase CpG coverage genome-wide	Mouse (mm10)	NA	Y	N
Tanas AS	Rapid and affordable genome-wide bisulfite DNA sequencing by XmaI-reduced representation bisulfite sequencing	2017	YES	NO	1	CpG islands	Human (hg19)	NA	Y	N
Martin-Herranz DE and Stubbs TM	cuRRBS	2017	YES	YES	Defined by the user	Defined by the user	Defined by the user	Defined by the user	Y	Y

Fig. S3.3 Table showing the comparison of different studies that have attempted to use restriction enzymes to target different regions in the genome.

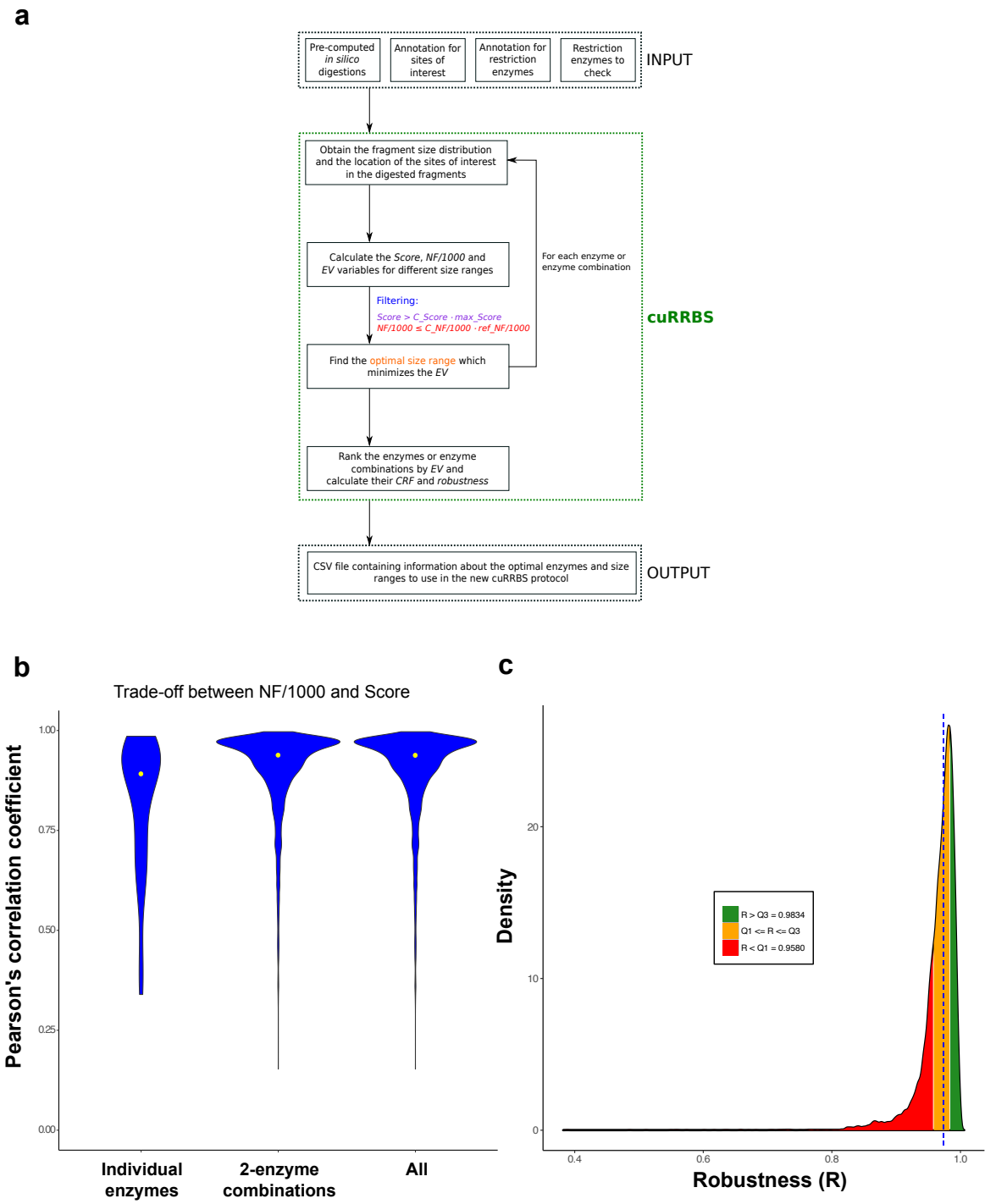


Fig. S3.4 Additional insights into cuRRBS. **a.** Detailed flowchart showing the input, main steps in cuRRBS and the output of the software. **b.** Violin plots showing the distribution of Pearson's correlation coefficients between the number of fragments (*NF*) and the *Score* for all the different enzymes tested with cuRRBS (single-enzyme, double-enzyme, all). In this example we used the Horvath epigenetic clock system [65], checking all the *size ranges* between 20 and 1000 bp, with an *experimental error* of 10 bp and a *read length* of 75 bp. Each yellow point represents the median for the Pearson's correlation coefficients under consideration. **c.** Density plot showing the distribution of the *robustness* (R) values when assuming an *experimental error* (δ) of 20 bp. cuRRBS was run for all the biological systems under study (Fig. S3.5) [65, 75–80] with the same parameters as described in 'Running cuRRBS for different *in silico* systems' (all the hits that satisfied the *thresholds* were reported in this case). The dashed blue line represents the median (0.9734). The different colours provide a way to judge the *robustness* values: bad (in red, $R < Q_1 = 0.9580$), medium (in orange, $Q_1 \leq R \leq Q_3 = 0.9834$) and good (in green, $R > Q_3$); where Q_1 and Q_3 represent the first and the third quartiles respectively.

Species	System	PMID where applicable	Additional information about the system	Total number of sites targeted	Optimal restriction enzyme combination	Optimal theoretical size range (in bp)	% max *Score*	*NF* /1000	*Enrichment Value* (*EV*)	*Cost Reduction Factor* (*CRF*)	*Robustness* (*R*)
Homo sapiens	Exon-intron boundaries		DNA methylation has been shown to affect alternative splicing. Therefore, we focused on targeting CpGs close to canonical splicing sites.	26211	(BsiSI OR MspI) AND (SbfI OR SdaI OR Sse8387I)	80_500	25.4	772.23	2.06446811	53.32	0.94704403
Homo sapiens	Horvath epigenetic clock	24138928	The Horvath epigenetic clock is the best predictor of biological age available in humans. We have attempted to target the 353 CpG sites that are used in the model in order to reduce the cost associated with the assay.	353	(BsiSI OR MspI) AND (BspQI OR LguI OR SapI)	60_160	27.57	442.456	3.65771916	93.06	0.91305072
Homo sapiens	Imprinted loci	26769960	Genomic imprinting is an epigenetic phenomenon that results in gene expression occuring in a parent-of-origin fashion. We have attempted to target Cs in CpG context that are found within the canonical human imprints.	2810	(BmeT110I OR BsoBI) AND (BsaWI)	60_540	25.12	336.88	2.67867053	122.23	0.98085689
Homo sapiens	Placental imprinted loci	26769960	Genomic imprinting is an epigenetic phenomenon that results in gene expression occuring in a parent-of-origin fashion. However, until recently many extraembryonic imprints were still unknown. We have targetted Cs in CpG context that are found within these novel human placental imprints.	7591	(BsaWI) AND (BssAI)	60_540	26.41	107.248	1.72827483	383.94	0.93382453
Homo sapiens	CTCF sites	26257180	CTCF is an important architectural protein that helps to organise chromatin domains. Since its binding has been shown to be dependent on DNA methylation in some of its recognition sequences, we have targeted the CpG sites within these regions of the genome.	2000	(BmeT110I OR BsoBI) AND (BssAI)	40_360	25.5	314.079	2.78946872	131.1	0.88798165
Mus musculus	iPSCs demethylated	28147265	iPSC reprogramming in mouse is characterised by global changes in DNA methylation. Sites that tend to undergo demethylation faster than the genome average tend to be within ESC-Super Enhancers. We targetted the Cs in CpG context in these regions, as they are interesting for the reprogramming field.	1449	(BmeT110I OR BsoBI) AND (BsiSI OR MspI)	80_980	25.19	974.05	3.42628839	37.31	0.96792238
Mus musculus	iPSCs maintained	28147265	iPSC reprogramming in mouse is characterised by global changes in DNA methylation. Sites that tend to be resistant to the genome-wide demethylation tend to be within Intercisernal A-particle containing regions. We targetted the Cs in CpG context in these regions, as they are interesting for the reprogramming field.	3896	(BmeT110I OR BsoBI) AND (BsiSI OR MspI)	80_560	25.85	690.088	2.835875	52.66	0.94227711
Mus musculus	NRF1 sites	26675734	NRF1 is a transcription factor whose binding to the DNA is dependent on the methylation status of its recognition sequences. We have tried to enrich for those CpG sites that overlap with *in vivo* NRF1 binding sites.	17018	(BmeT110I OR BsoBI) AND (PaeI OR SphI)	20_760	25.04	445.36	2.01909776	81.6	0.99634045
Arabidopsis thaliana	CHG sites	27419873	Non-CpG methylation is an important epigenetic modification in plants. In this study a huge number of regions containing non-CpG methylation were found to vary between different Arabidopsis accessions in the 1001 Epigenomes Project. We targetted Cs in non-CpG context within these non-CpG DMRs.	21801	(AanI OR PsiI) AND (Csp6I OR CviQI)	100_520	25.05	165.313	1.48095531	9.65	0.94999336

Fig. S3.5 Table showing the information regarding the different biological systems [65, 75–80] for which cuRRBS was run *in silico*. Some variables from the top hits in cuRRBS output are also reported.

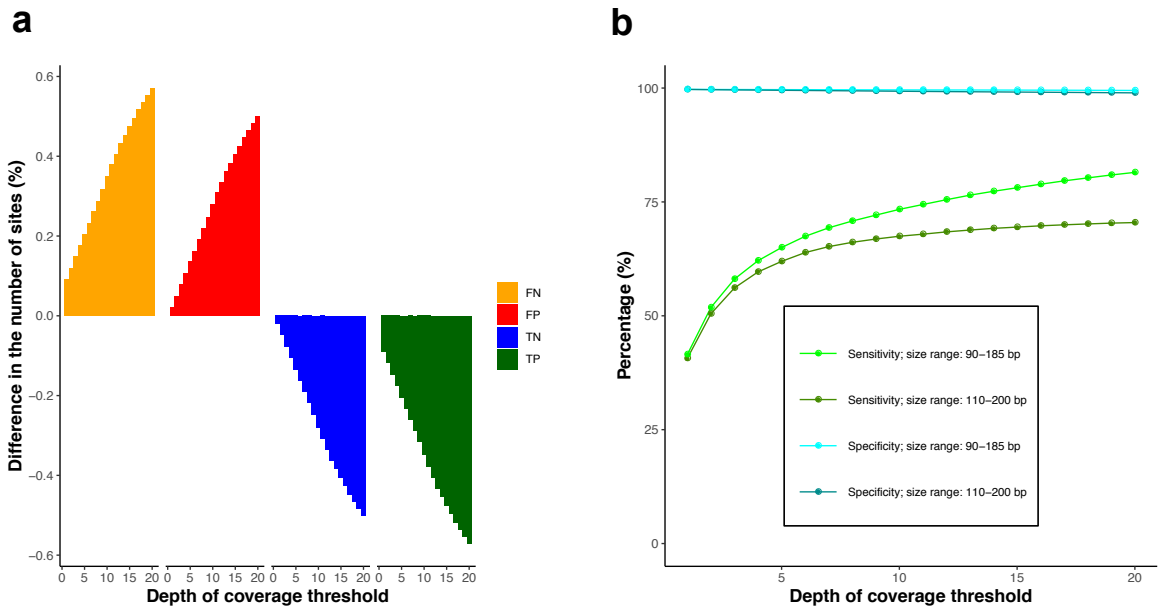


Fig. S3.6 Effect of experimental errors during size selection in cuRRBS predictions. **a.** Barplots showing the difference in the number of true positives (TP, in green), true negatives (TN, in blue), false positives (FP, in red) and false negatives (FN, in yellow) derived from cuRRBS theoretical predictions for the XmaI-RRBS data [55] using two different size ranges: 110-200 bp (aimed size range) and 90-185 bp (real size range). The difference observed between the two size ranges (aimed - real) is expressed as the percentage of the total number of sites considered (i.e. all CGI- CpGs). The number of sites in each category is calculated for different thresholds in the depth of coverage (number of reads covering a CpG site as reported by Bismark). cuRRBS was run for XmaI with all the default parameters (with a *read length* of 200 bp). Legend is displayed on the right hand side. **b.** Plot showing values of cuRRBS sensitivity and specificity as a function of the depth of coverage threshold employed to filter the experimental data [55]. The two size ranges considered in a. (aimed: 110-200 bp; real: 90-185 bp) are used for the calculations. Legend is displayed below the plot curves.

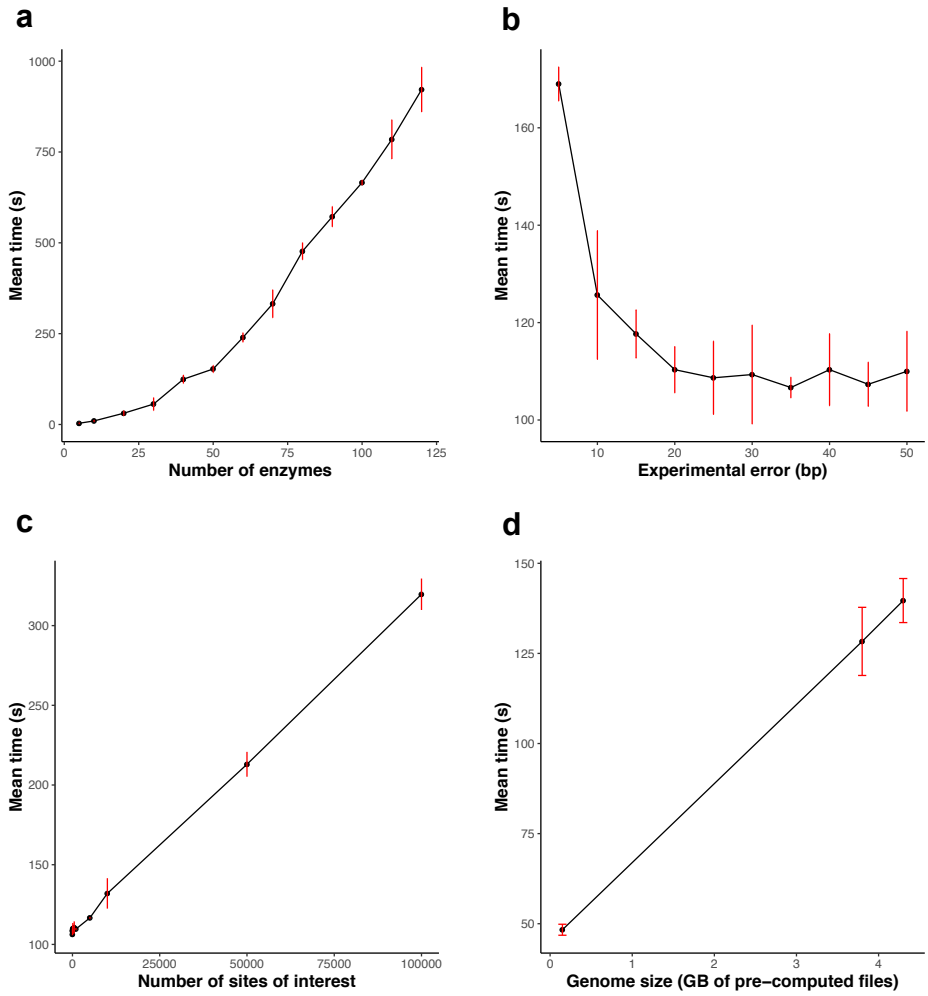


Fig. S3.7 cuRRBS computational efficiency. **a.** Plot showing the dependency between the number of enzymes checked and the computational (real) time required by the software (mean between 3 independent runs). cuRRBS was run for the Horvath epigenetic clock system [65] with a *read length* of 75 bp, a *Score threshold* of 25% and an *experimental error* of 10 bp. A laptop with an Intel® Core^{TM} i7-6600U CPU was used, which allowed cuRRBS to employ 4 parallel threads. The red error bars display the mean ± SD for the 3 independent runs. **b.** Plot showing the dependency between the *experimental error* (which determines how many size ranges are sampled) and the computational (real) time required by the software (mean between 3 independent runs). cuRRBS was run for the Horvath epigenetic clock system [65] with a *read length* of 75 bp, a *Score threshold* of 25% and a list with 40 enzymes. A laptop with an Intel® Core^{TM} i7-6600U CPU was used, which allowed cuRRBS to employ 4 parallel threads. The red error bars display the mean ± SD for the 3 independent runs. **c.** Plot showing the dependency between the number of sites of interest and the computational (real) time required by the software (mean between 3 independent runs). cuRRBS was run with a *read length* of 75 bp, a *Score threshold* of 25%, an *experimental error* of 10 bp and a list with 40 enzymes. A laptop with an Intel® Core^{TM} i7-6600U CPU was used, which allowed cuRRBS to employ 4 parallel threads. The red error bars display the mean ± SD for the 3 independent runs. **d.** Plot showing the dependency between genome size (measured as the size in GB of all the pre-computed files) and the computational (real) time required by the software (mean between 3 independent runs). cuRRBS was run with a *read length* of 75 bp, a *Score threshold* of 25%, an experimental error of 10 bp and a list with 40 enzymes. A laptop with an Intel® Core^{TM} i7-6600U CPU was used, which allowed cuRRBS to employ 4 parallel threads. The red error bars display the mean ± SD for the 3 independent runs.

References

[1] V K Rakyan, T A Down, D J Balding, and S Beck. Epigenome-wide association studies for common human diseases. *Nat Rev Genet*, 12:529–541, 2011.

[2] James M Flanagan. Epigenome-Wide Association Studies (EWAS): Past, Present, and Future. In Mukesh Verma, editor, *Cancer Epigenetics: Risk Assessment, Diagnosis, Treatment and Prognosis*, pages 51–63. Springer New York, New York, NY, 2015.

[3] R Edgar, M Domrachev, and AE Lash. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Research*, 30(1):207–210, 2002.

[4] Erfan Aref-Eshghi, David I. Rodenhiser, Laila C. Schenkel, Hanxin Lin, Cindy Skinner, Peter Ainsworth, Guillaume Paré, Rebecca L. Hood, Dennis E. Bulman, Kristin D. Kernohan, Kym M. Boycott, Philippe M. Campeau, Charles Schwartz, and Bekim Sadikovic. Genomic DNA Methylation Signatures Enable Concurrent Diagnosis and Clinical Genetic Variant Classification in Neurodevelopmental Syndromes. *American Journal of Human Genetics*, 102(1):156–174, 2018.

[5] M Bibikova, J Le, B Barnes, S Saedinia-Melnyk, L Zhou, R Shen, and K L Gunderson. Genome-wide DNA methylation profiling using Infinium® assay. *Epigenomics*, 1(1):177–200, 2009.

[6] M Bibikova, B Barnes, C Tsan, V Ho, B Klotzle, J M Le, D Delano, L Zhang, G P Schroth, K L Gunderson, J B Fan, and R Shen. High density DNA methylation array with single CpG site resolution. *Genomics*, 98(4):288–295, 2011.

[7] Ruth Pidsley, Elena Zotenko, Timothy J Peters, Mitchell G Lawrence, Gail P Risbridger, Peter Molloy, Susan Van Djik, Beverly Muhlhausler, Clare Stirzaker, and Susan J Clark. Critical evaluation of the Illumina MethylationEPIC BeadChip microarray for whole-genome DNA methylation profiling. *Genome Biology*, 17(1):208, 2016.

[8] Sean Davis and Paul S Meltzer. GEOquery: a bridge between the Gene Expression Omnibus (GEO) and BioConductor. *Bioinformatics*, 23(14):1846–1847, 2007.

[9] C S Wilhelm-Benartzi, D C Koestler, M R Karagas, J M Flanagan, B C Christensen, K T Kelsey, C J Marsit, E A Houseman, and R Brown. Review of processing and analysis methods for DNA methylation array data. *Br J Cancer*, 109(6):1394–1402, 2013.

[10] Tiffany J Morris and Stephan Beck. Analysis pipelines and packages for Infinium HumanMethylation450 BeadChip (450k) data. *Methods*, 72:3–8, 2015.

[11] Jie Liu and Kimberly D Siegmund. An evaluation of processing methods for HumanMethylation450 BeadChip data. *BMC Genomics*, 17(1):469, 2016.

[12] Martin J. Aryee, Andrew E. Jaffe, Hector Corrada-Bravo, Christine Ladd-Acosta, Andrew P. Feinberg, Kasper D. Hansen, and Rafael A. Irizarry. Minfi: A flexible and comprehensive Bioconductor package for the analysis of Infinium DNA methylation microarrays. *Bioinformatics*, 30(10):1363–1369, 2014.

[13] Timothy J Triche Jr, Daniel J Weisenberger, David Van Den Berg, Peter W Laird, and Kimberly D Siegmund. Low-level processing of Illumina Infinium DNA Methylation BeadArrays. *Nucleic Acids Research*, 41(7):e90, 2013.

[14] Yi-an Chen, Mathieu Lemire, Sanaa Choufani, Darci T Butcher, Daria Grafodatskaya, Brent W Zanke, Steven Gallinger, Thomas J Hudson, and Rosanna Weksberg. Discovery of cross-reactive probes and polymorphic CpGs in the Illumina Infinium HumanMethylation450 microarray. *Epigenetics*, 8(2):203–209, 2013.

[15] Jean-Philippe Fortin and Kasper D. Hansen. minfi guidelines: analysis of 450K data using minfi, 2015.

[16] P Du, X Zhang, C . C Huang, N Jafari, W A Kibbe, L Hou, and S M Lin. Comparison of Beta-value and M-value methods for quantifying methylation levels by microarray analysis. *BMC Bioinformatics*, 11:587, 2010.

[17] Joanna Zhuang, Martin Widschwendter, and Andrew E Teschendorff. A comparison of feature selection and classification methods in DNA methylation studies using the Illumina Infinium platform. *BMC Bioinformatics*, 13(1):59, 2012.

[18] Andrew E Teschendorff, Francesco Marabita, Matthias Lechner, Thomas Bartlett, Jesper Tegner, David Gomez-Cabrero, and Stephan Beck. A Beta-Mixture Quantile Normalisation method for correcting probe design bias in Illumina Infinium 450k DNA methylation data. *Bioinformatics (Oxford, England)*, 29(2):189–196, 2012.

[19] Sarah Dedeurwaerder, Matthieu Defrance, Emilie Calonne, Hélène Denis, Christos Sotiriou, and François Fuks. Evaluation of the Infinium Methylation 450K technology. *Epigenomics*, 3(6):771–784, 2011.

[20] Nizar Touleimat and Jörg Tost. Complete pipeline for Infinium® Human Methylation 450K BeadChip data processing using subset quantile normalization for accurate DNA methylation estimation. *Epigenomics*, 4(3):325–341, 2012.

[21] Jovana Maksimovic, Lavinia Gordon, and Alicia Oshlack. SWAN: Subset-quantile Within Array Normalization for Illumina Infinium HumanMethylation450 BeadChips. *Genome Biology*, 13(6):1–12, 2012.

[22] Jay Shendure and Hanlee Ji. Next-generation DNA sequencing. *Nature Biotechnology*, 26:1135, 2008.

[23] International Human Genome Sequencing Consortium, Eric S Lander, Lauren M Linton, Bruce Birren, Chad Nusbaum, Michael C Zody, Jennifer Baldwin, Keri Devon, Ken Dewar, Michael Doyle, William FitzHugh, Roel Funke, Diane Gage, Katrina Harris, Andrew Heaford, John Howland, Lisa Kann, Jessica Lehoczky, Rosie LeVine, Paul McEwan, Kevin McKernan, James Meldrim, Jill P Mesirov, Cher Miranda, William Morris, Jerome Naylor, Christina Raymond, Mark Rosetti, Ralph Santos, Andrew Sheridan, Carrie Sougnez, Nicole Stange-Thomann, Nikola Stojanovic, Aravind Subramanian, Dudley Wyman, Jane Rogers, John Sulston, Rachael Ainscough, Stephan Beck, David Bentley, John Burton, Christopher Clee, Nigel Carter, Alan Coulson, Rebecca Deadman, Panos Deloukas, Andrew Dunham, Ian Dunham, Richard Durbin, Lisa French, Darren Grafham, Simon Gregory, Tim Hubbard, Sean Humphray, Adrienne Hunt, Matthew Jones, Christine Lloyd, Amanda McMurray, Lucy Matthews, Simon Mercer, Sarah Milne, James C Mullikin, Andrew Mungall, Robert Plumb, Mark Ross, Ratna Shownkeen, Sarah Sims, Robert H Waterston, Richard K Wilson, LaDeana W Hillier, John D McPherson, Marco A Marra, Elaine R Mardis, Lucinda A Fulton, Asif T Chinwalla, Kymberlie H Pepin, Warren R Gish, Stephanie L Chissoe, Michael C Wendl, Kim D Delehaunty, Tracie L Miner, Andrew Delehaunty, Jason B Kramer, Lisa L Cook, Robert S Fulton, Douglas L Johnson, Patrick J Minx, Sandra W Clifton, Trevor Hawkins, Elbert Branscomb, Paul Predki, Paul Richardson, Sarah Wenning, Tom Slezak, Norman Doggett, Jan-Fang Cheng, Anne Olsen, Susan Lucas, Christopher Elkin, Edward Uberbacher, Marvin Frazier, Richard A Gibbs, Donna M Muzny, Steven E Scherer, John B Bouck, Erica J Sodergren, Kim C Worley, Catherine M Rives, James H Gorrell, Michael L Metzker, Susan L Naylor, Raju S Kucherlapati, David L Nelson, George M Weinstock, Yoshiyuki Sakaki, Asao Fujiyama, Masahira Hattori, Tetsushi Yada, Atsushi Toyoda, Takehiko Itoh, Chiharu Kawagoe, Hidemi Watanabe, Yasushi Totoki, Todd Taylor, Jean Weissenbach, Roland Heilig, William Saurin, Francois Artiguenave, Philippe Brottier, Thomas Bruls, Eric Pelletier, Catherine Robert, Patrick Wincker, André Rosenthal, Matthias Platzer, Gerald Nyakatura, Stefan Taudien, Andreas Rump, Douglas R Smith, Lynn Doucette-Stamm, Marc Rubenfield, Keith Weinstock, Hong Mei Lee, JoAnn Dubois, Huanming Yang, Jun Yu, Jian Wang, Guyang Huang, Jun Gu, Leroy Hood, Lee Rowen, Anup Madan, Shizen Qin, Ronald W Davis, Nancy A Federspiel, A Pia Abola, Michael J Proctor, Bruce A Roe, Feng Chen, Huaqin Pan, Juliane Ramser, Hans Lehrach, Richard Reinhardt, W Richard McCombie, Melissa de la Bastide, Neilay Dedhia, Helmut Blöcker, Klaus Hornischer, Gabriele Nordsiek, Richa Agarwala, L Aravind, Jeffrey A Bailey, Alex Bateman, Serafim Batzoglou, Ewan Birney, Peer Bork, Daniel G Brown, Christopher B Burge, Lorenzo Cerutti, Hsiu-Chuan Chen, Deanna Church, Michele Clamp, Richard R Copley, Tobias Doerks, Sean R Eddy, Evan E Eichler, Terrence S Furey, James Galagan, James G R Gilbert, Cyrus Harmon, Yoshihide Hayashizaki, David Haussler, Henning Hermjakob, Karsten Hokamp, Wonhee Jang, L Steven Johnson, Thomas A Jones, Simon Kasif, Arek Kaspryzk, Scot Kennedy, W James Kent, Paul Kitts, Eugene V Koonin, Ian Korf, David Kulp, Doron Lancet, Todd M Lowe, Aoife McLysaght, Tarjei Mikkelsen, John V Moran, Nicola Mulder, Victor J Pollara, Chris P Ponting, Greg Schuler, Jörg Schultz, Guy Slater, Arian F A Smit, Elia Stupka, Joseph Szustakowki, Danielle Thierry-Mieg, Jean Thierry-Mieg, Lukas Wagner, John Wallis, Raymond Wheeler, Alan Williams, Yuri I Wolf, Kenneth H Wolfe, Shiaw-Pyng Yang, Ru-Fang Yeh, Francis Collins, Mark S Guyer, Jane Peterson, Adam Felsenfeld, Kris A Wetterstrand, Richard M Myers, Jeremy Schmutz, Mark Dickson, Jane Grimwood,

David R Cox, Maynard V Olson, Rajinder Kaul, Christopher Raymond, Nobuyoshi Shimizu, Kazuhiko Kawasaki, Shinsei Minoshima, Glen A Evans, Maria Athanasiou, Roger Schultz, Aristides Patrinos, and Michael J Morgan. Initial sequencing and analysis of the human genome. *Nature*, 409:860–921, 2001.

[24] Mouse Genome Sequencing Consortium, Asif T Chinwalla, Lisa L Cook, Kimberly D Delehaunty, Ginger A Fewell, Lucinda A Fulton, Robert S Fulton, Tina A Graves, LaDeana W Hillier, Elaine R Mardis, John D McPherson, Tracie L Miner, William E Nash, Joanne O Nelson, Michael N Nhan, Kymberlie H Pepin, Craig S Pohl, Tracy C Ponce, Brian Schultz, Johanna Thompson, Evanne Trevaskis, Robert H Waterston, Michael C Wendl, Richard K Wilson, Shiaw-Pyng Yang, Peter An, Eric Berry, Bruce Birren, Toby Bloom, Daniel G Brown, Jonathan Butler, Mark Daly, Robert David, Justin Deri, Sheila Dodge, Karen Foley, Diane Gage, Sante Gnerre, Timothy Holzer, David B Jaffe, Michael Kamal, Elinor K Karlsson, Cristyn Kells, Andrew Kirby, Edward J Kulbokas III, Eric S Lander, Tom Landers, J P Leger, Rosie Levine, Kerstin Lindblad-Toh, Evan Mauceli, John H Mayer, Megan McCarthy, Jim Meldrim, Jim Meldrim, Jill P Mesirov, Robert Nicol, Chad Nusbaum, Steven Seaman, Ted Sharpe, Andrew Sheridan, Jonathan B Singer, Ralph Santos, Brian Spencer, Nicole Stange-Thomann, Jade P Vinson, Claire M Wade, Jamey Wierzbowski, Dudley Wyman, Michael C Zody, Ewan Birney, Nick Goldman, Arkadiusz Kasprzyk, Emmanuel Mongin, Alistair G Rust, Guy Slater, Arne Stabenau, Abel Ureta-Vidal, Simon Whelan, Rachel Ainscough, John Attwood, Jonathon Bailey, Karen Barlow, Stephan Beck, John Burton, Michele Clamp, Christopher Clee, Alan Coulson, James Cuff, Val Curwen, Tim Cutts, Joy Davies, Eduardo Eyras, Darren Grafham, Simon Gregory, Tim Hubbard, Adrienne Hunt, Matthew Jones, Ann Joy, Steven Leonard, Christine Lloyd, Lucy Matthews, Stuart McLaren, Kirsten McLay, Beverley Meredith, James C Mullikin, Zemin Ning, Karen Oliver, Emma Overton-Larty, Robert Plumb, Simon Potter, Michael Quail, Jane Rogers, Carol Scott, Steve Searle, Ratna Shownkeen, Sarah Sims, Melanie Wall, Anthony P West, David Willey, Sophie Williams, Josep F Abril, Roderic Guigó, Genís Parra, Pankaj Agarwal, Richa Agarwala, Deanna M Church, Wratko Hlavina, Donna R Maglott, Victor Sapojnikov, Marina Alexandersson, Lior Pachter, Stylianos E Antonarakis, Emmanouil T Dermitzakis, Alexandre Reymond, Catherine Ucla, Robert Baertsch, Mark Diekhans, Terrence S Furey, Angela Hinrichs, Fan Hsu, Donna Karolchik, W James Kent, Krishna M Roskin, Matthias S Schwartz, Charles Sugnet, Ryan J Weber, Peer Bork, Ivica Letunic, Mikita Suyama, David Torrents, Evgeny M Zdobnov, Marc Botcherby, Stephen D Brown, Robert D Campbell, Ian Jackson, Nicolas Bray, Olivier Couronne, Inna Dubchak, Alex Poliakov, Edward M Rubin, Michael R Brent, Paul Flicek, Evan Keibler, Ian Korf, S Batalov, Carol Bult, Wayne N Frankel, Piero Carninci, Yoshihide Hayashizaki, Jun Kawai, Yasushi Okazaki, Simon Cawley, David Kulp, Raymond Wheeler, Francesca Chiaromonte, Francis S Collins, Adam Felsenfeld, Mark Guyer, Jane Peterson, Kris Wetterstrand, Richard R Copley, Richard Mott, Colin Dewey, Nicholas J Dickens, Richard D Emes, Leo Goodstadt, Chris P Ponting, Eitan Winter, Diane M Dunn, Andrew C von Niederhausern, Robert B Weiss, Sean R Eddy, L Steven Johnson, Thomas A Jones, Laura Elnitski, Diana L Kolbe, Pallavi Eswara, Webb Miller, Michael J O'Connor, Scott Schwartz, Richard A Gibbs, Donna M Muzny, Gustavo Glusman, Arian Smit, Eric D Green, Ross C Hardison, Shan Yang, David Haussler, Axin Hua, Bruce A Roe, Raju S Kucherlapati, Kate T Montgomery, Jia Li, Ming Li, Susan Lucas, Bin Ma, W Richard McCombie, Michael Morgan, Pavel Pevzner, Glenn Tesler, Jörg Schultz,

Douglas R Smith, John Tromp, Kim C Worley, Eric S Lander, Josep F Abril, Pankaj Agarwal, Marina Alexandersson, Stylianos E Antonarakis, Robert Baertsch, Eric Berry, Ewan Birney, Peer Bork, Nicolas Bray, Michael R Brent, Daniel G Brown, Jonathan Butler, Carol Bult, Francesca Chiaromonte, Asif T Chinwalla, Deanna M Church, Michele Clamp, Francis S Collins, Richard R Copley, Olivier Couronne, Simon Cawley, James Cuff, Val Curwen, Tim Cutts, Mark Daly, Emmanouil T Dermitzakis, Colin Dewey, Nicholas J Dickens, Mark Diekhans, Inna Dubchak, Sean R Eddy, Laura Elnitski, Richard D Emes, Pallavi Eswara, Eduardo Eyras, Adam Felsenfeld, Paul Flicek, Wayne N Frankel, Lucinda A Fulton, Terrence S Furey, Sante Gnerre, Gustavo Glusman, Nick Goldman, Leo Goodstadt, Eric D Green, Simon Gregory, Roderic Guigó, Ross C Hardison, David Haussler, LaDeana W Hillier, Angela Hinrichs, Wratko Hlavina, Fan Hsu, Tim Hubbard, David B Jaffe, Michael Kamal, Donna Karolchik, Elinor K Karlsson, Arkadiusz Kasprzyk, Evan Keibler, W James Kent, Andrew Kirby, Diana L Kolbe, Ian Korf, Edward J Kulbokas III, David Kulp, Eric S Lander, Ivica Letunic, Ming Li, Kerstin Lindblad-Toh, Bin Ma, Donna R Maglott, Evan Mauceli, Jill P Mesirov, Webb Miller, Richard Mott, James C Mullikin, Zemin Ning, Lior Pachter, Genís Parra, Pavel Pevzner, Alex Poliakov, Chris P Ponting, Simon Potter, Alexandre Reymond, Krishna M Roskin, Victor Sapojnikov, Jörg Schultz, Matthias S Schwartz, Scott Schwartz, Steve Searle, Jonathan B Singer, Guy Slater, Arian Smit, Arne Stabenau, Charles Sugnet, Mikita Suyama, Glenn Tesler, David Torrents, John Tromp, Catherine Ucla, Jade P Vinson, Claire M Wade, Ryan J Weber, Raymond Wheeler, Eitan Winter, Shiaw-Pyng Yang, Evgeny M Zdobnov, Robert H Waterston, Simon Whelan, Kim C Worley, and Michael C Zody. Initial sequencing and comparative analysis of the mouse genome. *Nature*, 420:520–562, 2002.

[25] Nicolas Sierro, James N D Battey, Sonia Ouadi, Nicolas Bakaher, Lucien Bovet, Adrian Willig, Simon Goepfert, Manuel C Peitsch, and Nikolai V Ivanov. The tobacco genome sequence and its comparison with those of tomato and potato. *Nature Communications*, 5:3833, 2014.

[26] Matteo Fumagalli. Assessing the Effect of Sequencing Depth and Sample Size in Population Genetics Inferences. *PLOS ONE*, 8(11):e79667, 2013.

[27] Hao Wu, Tianlei Xu, Hao Feng, Li Chen, Ben Li, Bing Yao, Zhaohui Qin, Peng Jin, and Karen N Conneely. Detection of differentially methylated regions from whole-genome bisulfite sequencing data without replicates. *Nucleic Acids Research*, 43(21):e141–e141, 2015.

[28] M J Ziller, H Gu, F Mueller, J Donaghey, L T Tsai, and O Kohlbacher. Charting a dynamic DNA methylation landscape of the human genome. *Nature*, 500:477–481, 2013.

[29] Masako Suzuki and John M Greally. Genome-wide DNA Methylation Analysis Using Massively Parallel Sequencing Technologies. *Seminars in Hematology*, 50(1):70–77, 2013.

[30] Nongluk Plongthongkum, Dinh H Diep, and Kun Zhang. Advances in the profiling of DNA modifications: cytosine methylation and beyond. *Nat Rev Genet*, 15(10):647–661, 2014.

[31] Wai-Shin Yong, Fei-Man Hsu, and Pao-Yang Chen. Profiling genome-wide DNA methylation. *Epigenetics & Chromatin*, 9(1):26, 2016.

[32] S. Kurdyukov and M. Bullock. DNA Methylation Analysis: Choosing the Right Method. *Biology*, 5(1):3, 2016.

[33] Thadeous J Kacmarczyk, Mame P Fall, Xihui Zhang, Yuan Xin, Yushan Li, Alicia Alonso, and Doron Betel. "Same difference": comprehensive evaluation of four DNA methylation measurement platforms. *Epigenetics & Chromatin*, 11(1):21, 2018.

[34] Oluwatosin Taiwo, Gareth A Wilson, Tiffany Morris, Stefanie Seisenberger, Wolf Reik, Daniel Pearce, Stephan Beck, and Lee M Butcher. Methylome analysis using MeDIP-seq with low DNA concentrations. *Nature Protocols*, 7:617–636, 2012.

[35] Arie B Brinkman, Femke Simmer, Kelong Ma, Anita Kaan, Jingde Zhu, and Hendrik G Stunnenberg. Whole-genome DNA methylation profiling using MethylCap-seq. *Methods*, 52(3):232–236, 2010.

[36] Edita Kriukienė, Viviane Labrie, Tarang Khare, Giedrė Urbanavičiūtė, Audronė Lapinaitė, Karolis Koncevičius, Daofeng Li, Ting Wang, Shraddha Pai, Carolyn Ptak, Juozas Gordevičius, Sun-Chong Wang, Artūras Petronis, and Saulius Klimašauskas. DNA unmethylome profiling by covalent capture of CpG sites. *Nature Communications*, 4:2190, 2013.

[37] Maxim Ivanov, Mart Kals, Marina Kacevska, Andres Metspalu, Magnus Ingelman-Sundberg, and Lili Milani. In-solution hybrid capture of bisulfite-converted DNA for targeted bisulfite sequencing of 174 ADME genes. *Nucleic Acids Research*, 41(6):e72, 2013.

[38] Fiona Allum, Xiaojian Shao, Frédéric Guénard, Marie-Michelle Simon, Stephan Busche, Maxime Caron, John Lambourne, Julie Lessard, Karolina Tandre, Åsa K Hedman, Tony Kwan, Bing Ge, The Multiple Tissue Human Expression Resource Consortium, Kourosh R Ahmadi, Chrysanthi Ainali, Amy Barrett, Veronique Bataille, Jordana T Bell, Alfonso Buil, Emmanouil T Dermitzakis, Antigone S Dimas, Richard Durbin, Daniel Glass, Neelam Hassanali, Catherine Ingle, David Knowles, Maria Krestyaninova, Cecilia M Lindgren, Christopher E Lowe, Eshwar Meduri, Paola di Meglio, Josine L Min, Stephen B Montgomery, Frank O Nestle, Alexandra C Nica, James Nisbet, Stephen O'Rahilly, Leopold Parts, Simon Potter, Johanna Sandling, Magdalena Sekowska, So-Youn Shin, Kerrin S Small, Nicole Soranzo, Gabriela Surdulescu, Mary E Travers, Loukia Tsaprouni, Sophia Tsoka, Alicja Wilk, Tsun-Po Yang, Krina T Zondervan, Lars Rönnblom, Mark I McCarthy, Panos Deloukas, Todd Richmond, Daniel Burgess, Timothy D Spector, André Tchernof, Simon Marceau, Mark Lathrop, Marie-Claude Vohl, Tomi Pastinen, and Elin Grundberg. Characterization of functional methylomes by next-generation capture sequencing identifies novel disease-associated variants. *Nature Communications*, 6:7211, 2015.

[39] Warren A Cheung, Xiaojian Shao, Andréanne Morin, Valérie Siroux, Tony Kwan, Bing Ge, Dylan Aïssi, Lu Chen, Louella Vasquez, Fiona Allum, Frédéric Guénard, Emmanuelle Bouzigon, Marie-Michelle Simon, Elodie Boulier, Adriana Redensek, Stephen Watt, Avik Datta, Laura Clarke, Paul Flicek, Daniel Mead, Dirk S Paul,

Stephan Beck, Guillaume Bourque, Mark Lathrop, André Tchernof, Marie-Claude Vohl, Florence Demenais, Isabelle Pin, Kate Downes, Hendrick G Stunnenberg, Nicole Soranzo, Tomi Pastinen, and Elin Grundberg. Functional variation in allelic methylomes underscores a strong genetic contribution and reveals novel epigenetic alterations in the human epigenome. *Genome Biology*, 18(1):50, 2017.

[40] Emily Hodges, Andrew D. Smith, Jude Kendall, Zhenyu Xuan, Kandasamy Ravi, Michelle Rooks, Michael Q. Zhang, Kenny Ye, Arindam Bhattacharjee, Leonardo Brizuela, W. Richard McCombie, Michael Wigler, Gregory J. Hannon, and James B. Hicks. High definition profiling of mammalian DNA methylation by array capture and single molecule bisulfite sequencing. *Genome Research*, 19:1593–1605, 2009.

[41] Jie Deng, Robert Shoemaker, Bin Xie, Athurva Gore, Emily M LeProust, Jessica Antosiewicz-Bourget, Dieter Egli, Nimet Maherali, In-Hyun Park, Junying Yu, George Q Daley, Kevin Eggan, Konrad Hochedlinger, James Thomson, Wei Wang, Yuan Gao, and Kun Zhang. Targeted bisulfite sequencing reveals changes in DNA methylation associated with nuclear reprogramming. *Nature Biotechnology*, 27:353–360, 2009.

[42] Dinh Diep, Nongluk Plongthongkum, Athurva Gore, Ho-Lim Fung, Robert Shoemaker, and Kun Zhang. Library-free methylation sequencing with bisulfite padlock probes. *Nature Methods*, 9:270–272, 2012.

[43] H. Kiyomi Komori, Sarah A. LaMere, Ali Torkamani, G. Traver Hart, Steve Kotsopoulos, Jason Warner, Michael L. Samuels, Jeff Olson, Steven R. Head, Phillip Ordoukhanian, Pauline L. Lee, Darren R. Link, and Daniel R. Salomon. Application of microdroplet PCR for large-scale targeted bisulfite sequencing. *Genome Research*, 21(10):1738–1745, 2011.

[44] Dirk S. Paul, Paul Guilhamon, Anna Karpathakis, Lee M. Butcher, Christina Thirlwell, Andrew Feber, and Stephan Beck. Assessment of raindrop BS-seq as a method for large-scale, targeted bisulfite sequencing. *Epigenetics*, 9(5):678–684, 2014.

[45] Diana L Bernstein, Vasumathi Kameswaran, John E Le Lay, Karyn L Sheaffer, and Klaus H Kaestner. The BisPCR2 method for targeted bisulfite sequencing. *Epigenetics {&} Chromatin*, 8:27, 2015.

[46] Yao Yang, Robert Sebra, Benjamin S Pullman, Wanqiong Qiao, Inga Peter, Robert J Desnick, C Ronald Geyer, John F DeCoteau, and Stuart A Scott. Quantitative and multiplexed DNA methylation analysis using long-read single-molecule real-time bisulfite sequencing (SMRT-BS). *BMC Genomics*, 16(1):350, 2015.

[47] Howard Cedar, Adina Solage, Gad Glaser, and Aharon Razin. Direct detection of methylated cytosine in DNA by use of the restriction enzyme MspI. *Nucleic Acids Research*, 6(6):2125–2132, 1979.

[48] Devora Cohen-Karni, Derrick Xu, Lynne Apone, Alexey Fomenkov, Zhiyi Sun, Paul J Davis, Shannon R Morey Kinney, Megumu Yamada-Mabuchi, Shuang-yong Xu, Theodore Davis, Sriharsa Pradhan, Richard J Roberts, and Yu Zheng. The MspJI family of modification-dependent restriction endonucleases for epigenetic studies. *Proceedings of the National Academy of Sciences*, 108(27):11040–11045, 2011.

[49] Leonid V Bystrykh. A combinatorial approach to the restriction of a mouse genome. *BMC Research Notes*, 6(1):284, 2013.

[50] Daniel B Martinez-Arguelles, Sunghoon Lee, and Vassilios Papadopoulos. In silico analysis identifies novel restriction enzyme combinations that expand reduced representation bisulfite sequencing CpG coverage. *BMC research notes*, 7(1):534, 2014.

[51] Li Yu, Chunhui Liu, Kristi Bennett, Yue-Zhong Wu, Zunyan Dai, Jeff Vandeusen, Rene Opavsky, Aparna Raval, Prashant Trikha, Ben Rodriguez, Brian Becknell, Charlene Mao, Stephen Lee, Ramana V Davuluri, Gustavo Leone, Ignatia B Van den Veyver, Michael A Caligiuri, and Christoph Plass. A NotI–EcoRV promoter library for studies of genetic and epigenetic alterations in mouse models of human malignancies. *Genomics*, 84(4):647–660, 2004.

[52] Alexander Meissner, Tarjei S Mikkelsen, Hongcang Gu, Marius Wernig, Jacob Hanna, Andrey Sivachenko, Xiaolan Zhang, Bradley E Bernstein, Chad Nusbaum, David B Jaffe, Andreas Gnirke, Rudolf Jaenisch, and Eric S Lander. Genome-scale DNA methylation maps of pluripotent and differentiated cells. *Nature*, 454(7205):766–70, 2008.

[53] Patrick Boyle, Kendell Clement, Hongcang Gu, Zachary D Smith, Michael Ziller, Jennifer L Fostel, Laurie Holmes, Jim Meldrim, Fontina Kelley, Andreas Gnirke, and Alexander Meissner. Gel-free multiplexed reduced representation bisulfite sequencing for large-scale DNA methylation profiling. *Genome Biology*, 13(10):R92, 2012.

[54] Alexander Meissner, Andreas Gnirke, George W. Bell, Bernard Ramsahoye, Eric S. Lander, and Rudolf Jaenisch. Reduced representation bisulfite sequencing for comparative high-resolution DNA methylation analysis. *Nucleic Acids Research*, 33(18):5868–5877, 2005.

[55] Alexander S Tanas, Marina E Borisova, Ekaterina B Kuznetsova, Viktoria V Rudenko, Kristina O Karandasheva, Marina V Nemtsova, Vera L Izhevskaya, Olga A Simonova, Sergey S Larin, Dmitry V Zaletaev, and Vladimir V Strelnikov. Rapid and affordable genome-wide bisulfite DNA sequencing by XmaI-reduced representation bisulfite sequencing. *Epigenomics*, 9(6):833–847, 2017.

[56] Yew Kok Lee, Shengnan Jin, Shiwei Duan, Yen Ching Lim, Desmond P Y Ng, Xueqin Michelle Lin, George S H Yeo, and Chunming Ding. Improved reduced representation bisulfite sequencing for epigenomic profiling of clinical samples. *Biological Procedures Online*, 16(1):1, 2014.

[57] Yen Ching Lim, Sook Yoong Chia, Shengnan Jin, Weiping Han, Chunming Ding, and Lei Sun. Dynamic DNA methylation landscape defines brown and white cell specificity during adipogenesis. *Molecular Metabolism*, 5(10):1033–1041, 2016.

[58] Xiaojun Huang, Hanlin Lu, Jun-Wen Wang, Liqin Xu, Siyang Liu, Jihua Sun, and Fei Gao. High-throughput sequencing of methylated cytosine enriched by modification-dependent restriction endonuclease MspJI. *BMC Genetics*, 14(1):56, 2013.

[59] Junwen Wang, Yudong Xia, Lili Li, Desheng Gong, Yu Yao, Huijuan Luo, Hanlin Lu, Na Yi, Honglong Wu, Xiuqing Zhang, Qian Tao, and Fei Gao. Double restriction-enzyme digestion improves the coverage and accuracy of genome-wide CpG methylation profiling by reduced representation bisulfite sequencing. *BMC genomics*, 14:11, 2013.

[60] Sophie A Kirschner, Oliver Hunewald, Sophie B Mériaux, Regina Brunnhoefer, Claude P Muller, and Jonathan D Turner. Focussing reduced representation CpG sequencing through judicious restriction enzyme choice. *Genomics*, 107(4):109–119, 2016.

[61] Hongcang Gu, Christoph Bock, Tarjei S Mikkelsen, Natalie Jäger, Zachary D Smith, Eleni Tomazou, Andreas Gnirke, Eric S Lander, and Alexander Meissner. Genome-scale DNA methylation mapping of clinical samples at single-nucleotide resolution. *Nature methods*, 7(2):133–136, 2010.

[62] Thomas M. Stubbs, Marc Jan Bonder, Anne-Katrien Stark, Felix Krueger, Ferdinand von Meyenn, Oliver Stegle, and Wolf Reik. Multi-tissue DNA methylation age predictor in mouse. *Genome Biology*, 18(1):68, 2017.

[63] Zachary D. Smith, Hongcang Gu, Christoph Bock, Andreas Gnirke, and Alexander Meissner. High-throughput bisulfite sequencing in mammalian genomes. *Methods*, 48(3):226–232, 2009.

[64] Carmen M Koch and Wolfgang Wagner. Epigenetic-aging-signature to determine age in different tissues. *Aging*, 3(10):1018–1027, 2011.

[65] Steve Horvath. DNA methylation age of human tissues and cell types. *Genome Biology*, 14(10):3156, 2013.

[66] G Hannum, J Guinney, L Zhao, L Zhang, G Hughes, and S Sadda. Genome-wide methylation profiles reveal quantitative views of human aging rates. *Mol Cell*, 49(2):359–367, 2013.

[67] Daniel A Petkovich, Dmitriy I Podolskiy, Alexei V Lobanov, Sang-Goo Lee, Richard A Miller, and Vadim N Gladyshev. Using DNA Methylation Profiling to Evaluate Biological Age and Longevity Interventions. *Cell Metabolism*, 25(4):954–960.e6, 2017.

[68] Michael J. Thompson, Karolina Chwiałkowska, Liudmilla Rubbi, Aldons J. Lusis, Richard C. Davis, Anuj Srivastava, Ron Korstanje, Gary A. Churchill, Steve Horvath, and Matteo Pellegrini. A multi-tissue full lifespan epigenetic clock for mice. *Aging*, 10(10):2832–2854, 2018.

[69] Margarita V Meer, Dmitriy I Podolskiy, Alexander Tyshkovskiy, and Vadim N Gladyshev. A whole lifespan mouse multi-tissue DNA methylation clock. *eLife*, 7:e40675, 2018.

[70] Michael J. Thompson, Bridgett von Holdt, Steve Horvath, and Matteo Pellegrini. An epigenetic aging clock for dogs and wolves. *Aging*, 9(3):1055–1068, 2017.

[71] Roderick C Slieker, Maarten van Iterson, René Luijk, Marian Beekman, Daria V Zhernakova, Matthijs H Moed, Hailiang Mei, Michiel van Galen, Patrick Deelen, Marc Jan Bonder, Alexandra Zhernakova, André G Uitterlinden, Ettje F Tigchelaar, Coen D A Stehouwer, Casper G Schalkwijk, Carla J H van der Kallen, Albert Hofman, Diana van Heemst, Eco J de Geus, Jenny van Dongen, Joris Deelen, Leonard H van den Berg, Joyce van Meurs, Rick Jansen, Peter A C 't Hoen, Lude Franke, Cisca Wijmenga, Jan H Veldink, Morris A Swertz, Marleen M J van Greevenbroek, Cornelia M van Duijn, Dorret I Boomsma, P Eline Slagboom, Bastiaan T Heijmans, and BIOS Consortium. Age-related accrual of methylomic variability is linked to fundamental ageing mechanisms. *Genome Biology*, 17(1):191, 2016.

[72] Roderick C Slieker, Caroline L Relton, Tom R Gaunt, P Eline Slagboom, and Bastiaan T Heijmans. Age-related DNA methylation changes are tissue-specific with ELOVL2 promoter methylation as exception. *Epigenetics & Chromatin*, 11(1):25, 2018.

[73] John J Cole, Neil A Robertson, Mohammed Iqbal Rather, John P Thomson, Tony McBryan, Duncan Sproul, Tina Wang, Claire Brock, William Clark, Trey Ideker, Richard R Meehan, Richard A Miller, Holly M Brown-Borg, and Peter D Adams. Diverse interventions that extend mouse lifespan suppress shared age-associated epigenetic changes at critical gene regulatory regions. *Genome Biology*, 18(1):58, 2017.

[74] Steve Horvath, Junko Oshima, George M. Martin, Ake T. Lu, Austin Quach, Howard Cohen, Sarah Felton, Mieko Matsuyama, Donna Lowe, Sylwia Kabacik, James G. Wilson, Alex P. Reiner, Anna Maierhofer, Julia Flunkert, Abraham Aviv, Lifang Hou, Andrea A. Baccarelli, Yun Li, James D. Stewart, Eric A. Whitsel, Luigi Ferrucci, Shigemi Matsuyama, and Kenneth Raj. Epigenetic clock for skin and blood cells applied to Hutchinson Gilford Progeria Syndrome and ex vivo studies. *Aging*, 10(7):1758–1775, 2018.

[75] Courtney W. Hanna, Maria S. Peï¿½aherrera, Heba Saadeh, Simon Andrews, Deborah E. McFadden, Gavin Kelsey, and Wendy P. Robinson. Pervasive polymorphic imprinted methylation in the human placenta. *Genome Research*, 26:756–767, 2016.

[76] Inês Milagre, Thomas M Stubbs, Michelle R King, Julia Spindel, Fátima Santos, Felix Krueger, Martin Bachman, Anne Segonds-Pichon, Shankar Balasubramanian, Simon R Andrews, Wendy Dean, and Wolf Reik. Gender Differences in Global but Not Targeted Demethylation in iPSC Reprogramming. *Cell Reports*, 18(5):1079–1089, 2017.

[77] Taiji Kawakatsu, Shao-shan Carol Huang, Florian Jupe, Eriko Sasaki, Robert J Schmitz, Mark A Urich, Rosa Castanon, Joseph R Nery, Cesar Barragan, Yupeng He, Huaming Chen, Manu Dubin, Cheng-Ruei Lee, Congmao Wang, Felix Bemm, Claude Becker, Ryan O'Neil, Ronan C O'Malley, Danjuma X Quarless, Carlos Alonso-Blanco, Jorge Andrade, Claude Becker, Felix Bemm, Joy Bergelson, Karsten Borgwardt, Eunyoung Chae, Todd Dezwaan, Wei Ding, Joseph R Ecker, Moisés Expósito-Alonso, Ashley Farlow, Joffrey Fitz, Xiangchao Gan, Dominik G Grimm, Angela Hancock, Stefan R Henz, Svante Holm, Matthew Horton, Mike Jarsulic, Randall A Kerstetter, Arthur Korte, Pamela Korte, Christa Lanz, Chen-Ruei Lee, Dazhe Meng, Todd P

Michael, Richard Mott, Ni Wayan Muliyati, Thomas Nägele, Matthias Nagler, Viktoria Nizhynska, Magnus Nordborg, Polina Novikova, F Xavier Picó, Alexander Platzer, Fernando A Rabanal, Alex Rodriguez, Beth A Rowan, Patrice A Salomé, Karl Schmid, Robert J Schmitz, Ümit Seren, Felice Gianluca Sperone, Mitchell Sudkamp, Hannes Svardal, Matt M Tanzer, Donald Todd, Samuel L Volchenboum, Congmao Wang, George Wang, Xi Wang, Wolfram Weckwerth, Detlef Weigel, Xuefeng Zhou, Nicholas J Schork, Detlef Weigel, Magnus Nordborg, and Joseph R Ecker. Epigenomic Diversity in a Global Collection of Arabidopsis thaliana Accessions. *Cell*, 166(2):492–505, 2016.

[78] Matthew T. Maurano, Hao Wang, Sam John, Anthony Shafer, Theresa Canfield, Kristen Lee, and John A. Stamatoyannopoulos. Role of DNA Methylation in Modulating Transcription Factor Occupancy. *Cell Reports*, 12(7):1184–1195, 2015.

[79] Galit Lev Maor, Ahuvi Yearim, and Gil Ast. The alternative role of DNA methylation in splicing regulation. *Trends in Genetics*, 31(5):274–280, 2015.

[80] Silvia Domcke, Anaïs Flore Bardet, Paul Adrian Ginno, Dominik Hartl, Lukas Burger, and Dirk Schübeler. Competition between DNA methylation and transcription factors determines binding of NRF1. *Nature*, 528(7583):575–579, 2015.

[81] Hume Stroud, Truman Do, Jiamu Du, Xuehua Zhong, Suhua Feng, Lianna Johnson, Dinshaw J Patel, and Steven E Jacobsen. Non-CG methylation patterns shape the epigenetic landscape in Arabidopsis. *Nature Structural &Amp; Molecular Biology*, 21:64–72, 2013.

[82] Yan Sun, Rui Hou, Xiaoteng Fu, Changsen Sun, Shi Wang, Chen Wang, Ning Li, Lingling Zhang, and Zhenmin Bao. Genome-Wide Analysis of DNA Methylation in Five Tissues of Zhikong Scallop, Chlamys farreri. *PLOS ONE*, 9(1):e86232, 2014.

[83] Weiwei Zhang, Tim D Spector, Panos Deloukas, Jordana T Bell, and Barbara E Engelhardt. Predicting genome-wide DNA methylation using methylation marks, genomic position, and DNA regulatory elements. *Genome biology*, 16(1):14, 2015.

[84] Christof Angermueller, Heather J Lee, Wolf Reik, and Oliver Stegle. DeepCpG: accurate prediction of single-cell DNA methylation states using deep learning. *Genome Biology*, 18(1):67, 2017.

[85] John W Davey and Mark L Blaxter. RADSeq: next-generation population genetics. *Briefings in Functional Genomics*, 9(5-6):416–423, 2011.

[86] John W Davey, Paul A Hohenlohe, Paul D Etter, Jason Q Boone, Julian M Catchen, and Mark L Blaxter. Genome-wide genetic marker discovery and genotyping using next-generation sequencing. *Nature Reviews Genetics*, 12:499–510, 2011.

[87] Natalia Naumova, Emily M Smith, Ye Zhan, and Job Dekker. Analysis of long-range chromatin interactions using Chromosome Conformation Capture. *Methods*, 58(3):192–203, 2012.

[88] Job Dekker, Marc A Marti-Renom, and Leonid A Mirny. Exploring the three-dimensional organization of genomes: interpreting chromatin interaction data. *Nature Reviews Genetics*, 14:390–403, 2013.

[89] Richard J Roberts, Tamas Vincze, Janos Posfai, and Dana Macelis. REBASE—restriction enzymes and DNA methyltransferases. *Nucleic Acids Research*, 33(suppl_1):D230–D232, 2005.

[90] Richard J Roberts, Tamas Vincze, Janos Posfai, and Dana Macelis. REBASE—a database for DNA restriction and modification: enzymes, genes and genomes. *Nucleic Acids Research*, 43(D1):D298–D299, 2015.

[91] Jennifer Harrow, Adam Frankish, Jose M. Gonzalez, Electra Tapanari, Mark Diekhans, Felix Kokocinski, Bronwen L. Aken, Daniel Barrell, Amonida Zadissa, Stephen Searle, If Barnes, Alexandra Bignell, Veronika Boychenko, Toby Hunt, Mike Kay, Gaurab Mukherjee, Jeena Rajan, Gloria Despacio-Reyes, Gary Saunders, Charles Steward, Rachel Harte, Michael Lin, Cédric Howald, Andrea Tanzer, Thomas Derrien, Jacqueline Chrast, Nathalie Walters, Suganthi Balasubramanian, Baikang Pei, Michael Tress, Jose Manuel Rodriguez, Iakes Ezkurdia, Jeltje Van Baren, Michael Brent, David Haussler, Manolis Kellis, Alfonso Valencia, Alexandre Reymond, Mark Gerstein, Roderic Guigó, and Tim J. Hubbard. GENCODE: The reference human genome annotation for the ENCODE project. *Genome Research*, 22:1760–1774, 2012.

[92] C Bock, J Walter, M Paulsen, and T Lengauer. CpG island mapping by epigenome prediction. *PLoS Comput Biol*, 3(6):e110, 2007.

[93] Leila Taher, Robin P Smith, Mee J Kim, Nadav Ahituv, and Ivan Ovcharenko. Sequence signatures extracted from proximal promoters can be used to predict distal enhancers. *Genome Biology*, 14(10):R117, 2013.

[94] Ryan K Dale, Brent S Pedersen, and Aaron R Quinlan. Pybedtools: a flexible Python library for manipulating genomic datasets and annotations. *Bioinformatics*, 27(24):3423–3424, 2011.

[95] Aaron R Quinlan and Ira M Hall. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, 26(6):841–842, 2010.

[96] Peter J A Cock, Tiago Antao, Jeffrey T. Chang, Brad A. Chapman, Cymon J. Cox, Andrew Dalke, Iddo Friedberg, Thomas Hamelryck, Frank Kauff, Bartek Wilczynski, and Michiel J L De Hoon. Biopython: Freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics*, 25(11):1422–1423, 2009.

[97] Daniel E Martin-Herranz, Antonio JM Ribeiro, and Thomas M Stubbs. demh/cuRRBS: cuRRBS V1.0.4, aug 2017.

[98] Robert M Kuhn, David Haussler, and W James Kent. The UCSC genome browser and associated tools. *Briefings in Bioinformatics*, 14(2):144–161, 2012.

[99] Anthony Mathelier, Oriol Fornes, David J Arenillas, Chih-yu Chen, Grégoire Denay, Jessica Lee, Wenqiang Shi, Casper Shyr, Ge Tan, Rebecca Worsley-Hunt, Allen W Zhang, François Parcy, Boris Lenhard, Albin Sandelin, and Wyeth W Wasserman. JASPAR 2016: a major expansion and update of the open-access database of transcription factor binding profiles. *Nucleic Acids Research*, 44(D1):D110–D115, 2015.

[100] Anaïs F. Bardet, Jonas Steinmann, Sangeeta Bafna, Juergen A. Knoblich, Julia Zeitlinger, and Alexander Stark. Identification of transcription factor binding sites from ChIP-seq data at high resolution. *Bioinformatics*, 29(21):2705–2713, 2013.

[101] Felix Krueger and Simon R Andrews. Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications. *Bioinformatics*, 27(11):1571–1572, 2011.