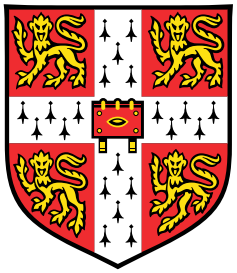
On the epigenetic ageing clock in humans



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
aDMPs	Differentially methylated positions during ageing
aVMPs	Variably methylated positions during ageing
B	$CD19^+$ B cells
BMIQ	Beta-mixture quantile normalisation
bp	Base pairs
CCC	Cell composition correction
CD4T	$CD4^+$ T cells
CD8T	$CD8^+$ T cells
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
CP/QP	Constrained projection/quadratic programming
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
DHS	DNase Hypersensitive Sites
DHS-DMCs	In cell-type deconvolution strategies, reference probes identified using information from differential methylation and chromatin accessibility
DMCs	Differentially methylated cytosines
DMCTs	Differentially methylated cytosines in individual cell types
DMPs	Differentially methylated positions
DMRs	Differentially methylated regions
DNA	Deoxyribonucleic acid
DNAmAge	DNA methylation age i.e. epigenetic age calculated with Horvath's epigenetic clock
EAA	Epigenetic age acceleration
EPIC	Illumina Infinium MethylationEPIC array
epiTOC	epigenetic Timer of Cancer (i.e. the epigenetic mitotic clock)
EV	Enrichment Value in cuRRBS
EWAS	Epigenome-wide association studies
FN	False negatives
FP	False positives
GB	Gigabytes
Gbp	Giga base pairs
GC content	Guanine + cytosine content
GEO	Gene Expression Omnibus repository
Gran	Granulocytes
hg38	Reference human genome assembly 38
IDOL	IDentifying Optimal DNA methylation Libraries, a strategy to build cell-type deconvolution references

IEAA	Intrinsic epigenetic age acceleration
iPSCs	Induced pluripotent stem cells
kb	Kilo base pairs
KNN	*k*-nearest neighbours
MAE	Mean absolute error (in the context of cell-type deconvolution benchmarking) or median absolute error (in the context of Horvath's epigenetic clock)
MBD	Methyl-CpG-binding domain
MEFs	Mouse embryonic fibroblasts
Mono	$CD14^+$ monocytes
NF	Theoretical number of fragments sequenced in cuRRBS
NK	$CD56^+$ natural killer cells
NRF1	Nuclear respiratory factor 1
OOB	Out-of-band fluorescence intensities in the Infinium I probes of Illumina arrays
PBMC	Peripheral blood mononuclear cells
PC	Principal component
PCA	Principal component analysis
PCC	Pearson's correlation coefficient
pcgtAge	Mitotic age according to the epigenetic mitotic clock (epiTOC)
PCR	Polymerase chain reaction
PRC2	Polycomb repressing complex 2
QC	Quality control
R	It can have two meanings: robustness variable in cuRRBS or the R programming language
R^2	Coefficient of determination
RAM	Random-access memory

RMSE	Root mean squared error
RNA	Ribonucleic acid
RPC	Robust partial correlations
RRBS	Reduced Representation Bisulfite Sequencing
SCC	Spearman's correlation coefficient
SD	Standard deviation
Sex_p	Sex predicted for a sample using DNA methylation data
SNP	Single-nucleotide polymorphism
SQN	Stratified quantile normalisation
TKO	Triple knockout
TN	True negatives
TP	True positives
TSS	Transcription start site
WGBS	Whole Genome Bisulfite Sequencing

Chapter 1

Introduction

1.1 Ageing stuff

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, biases and number of probes [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 from human blood that I could find, including samples from both whole blood and peripheral blood mononuclear cells (PBMC). Furthermore, the data also had to satisfy 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 section 2.2.3) 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 assemble 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.

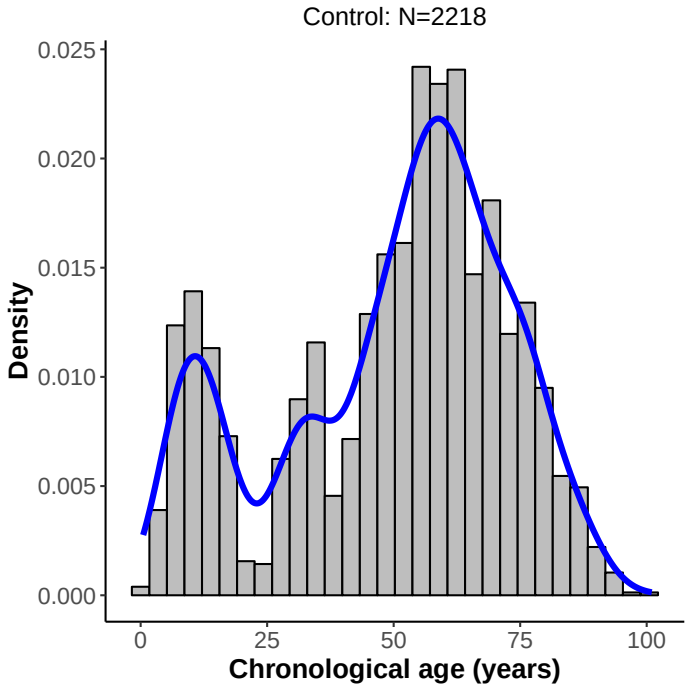


Fig. 2.1 Chronological age distribution in the healthy individuals. Histogram showing the chronological age distribution for all the healthy individuals included in the DNA methylation dataset. The blue line represents the 1D kernel density estimate, as calculate by the *stat_density* function in R with 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 section 2.1.2).

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 methylation values (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 $X_s \sim Exp(\gamma)$ and $X_b \sim N(\mu, \sigma^2)$) 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).

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, which can be calculated as [9, 16]:

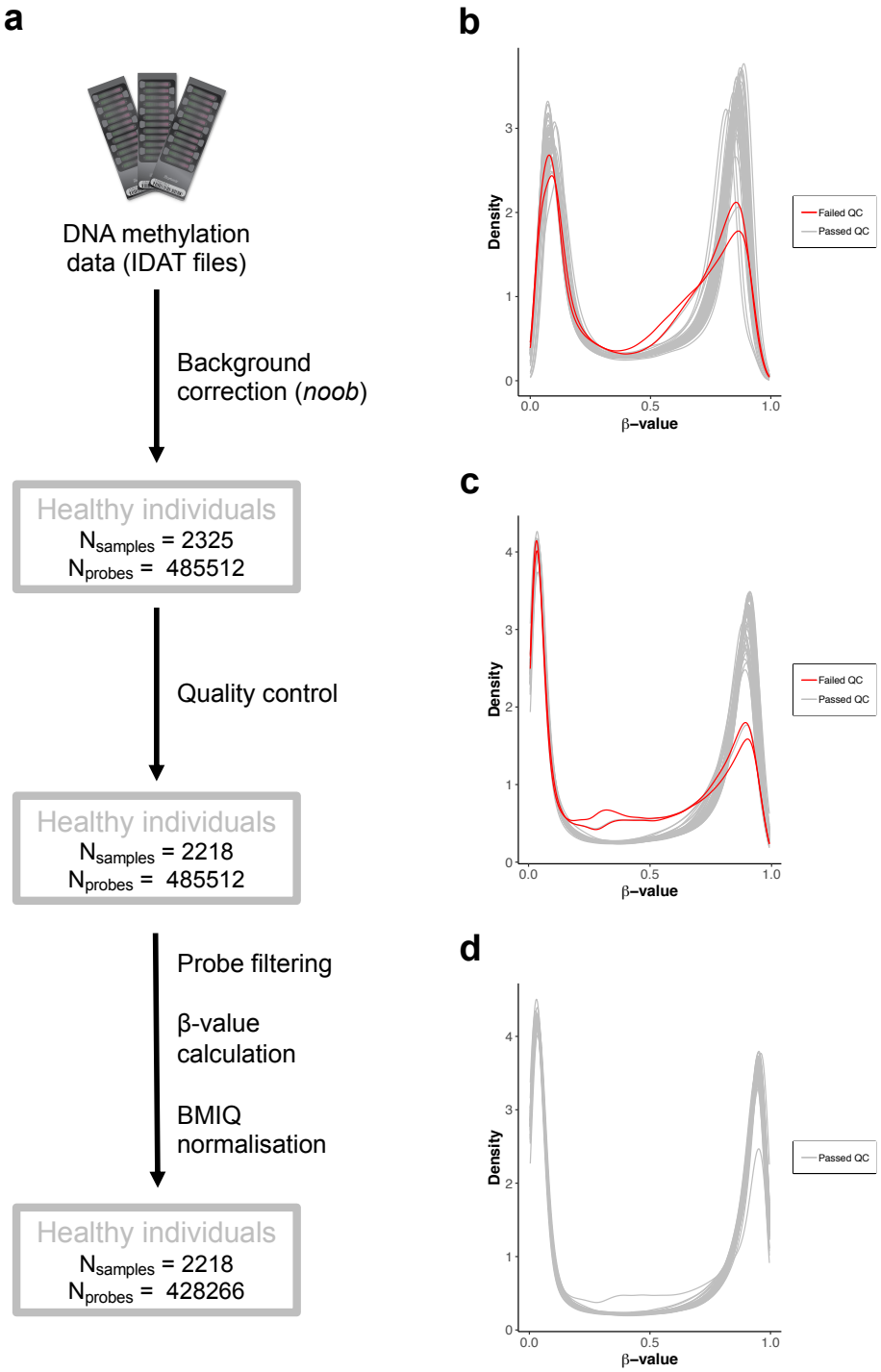


Fig. 2.2 Main DNA methylation data pre-processing pipeline. **a.** Flowchart showing the main steps implemented to pre-process the DNA methylation data from the 450K methylation arrays. The number of samples ($N_{samples}$) and the number of array probes (N_{probes}) left after each step are also specified for the samples from the healthy individuals. **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.

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

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 molecule of a single cell, a specific cytosine 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%) in that cytosine, which is roughly equivalent to say that 100% of the cells are either unmethylated or methylated respectively in that cytosine for the sampled tissue. The β-values for a given sample (i.e. considering all the cytosines measured, normally in a CpG context) usually 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.

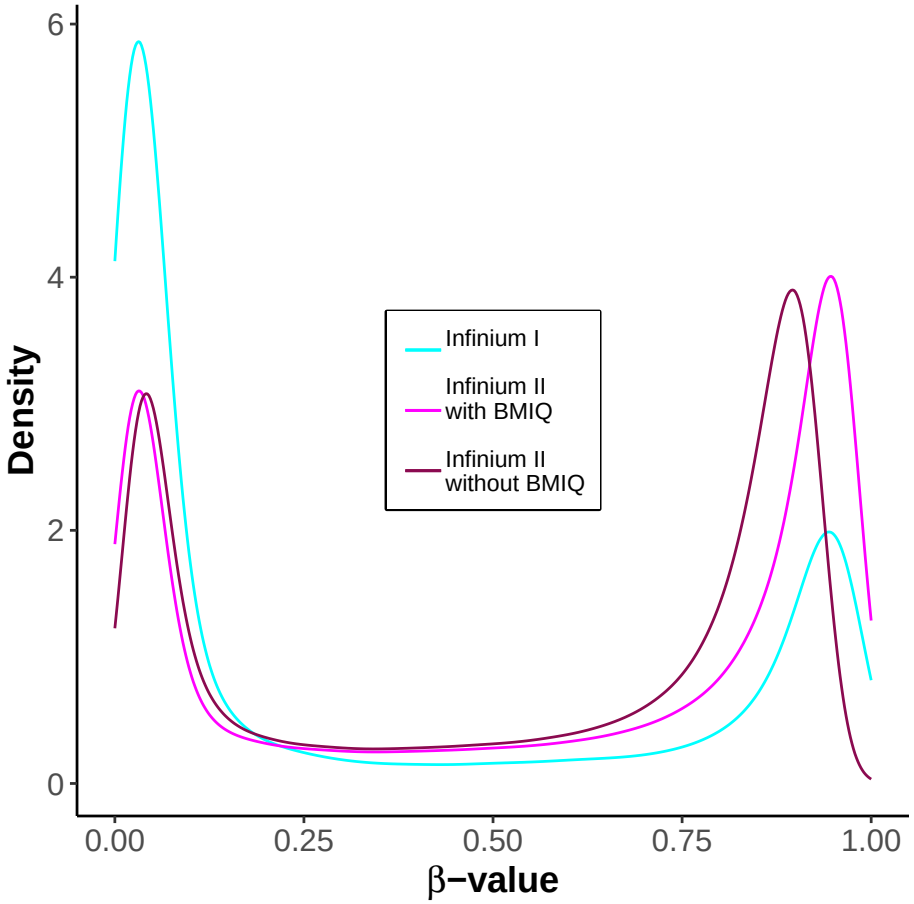


Fig. 2.3 Effect of BMIQ normalisation on the β-value distribution of different subsets of array probes with different chemistries (Infinium I, Infinium II). These results correspond to a DNA methylation sample from the GSE41273 batch. It can be appreciated how BMIQ transforms the distribution of the Infinium II probes into a distribution more similar to the Infinium I probes.

2.1.3 Accounting for blood cell composition changes during ageing

Whole blood is composed of several cell types that contain a nucleus, including neutrophils, eosinophils, basophils, $CD14^+$ monocytes, $CD4^+$ T cells, $CD8^+$ T cells, $CD19^+$ B cells and $CD56^+$ natural killer (NK) cells [22]. These cell types have different epigenetic profiles and, as a consequence, changes in their proportions (i.e. changes in blood cell composition) can affect bulk DNA methylation measurements [23].

Accounting for this cellular heterogeneity is really important in epigenome-wide association studies (EWAS) [24–26]. Furthermore, previous research has highlighted changes in blood cell composition with age, which could be one of the causes behind immunosenescence [27–31]. Therefore, considering blood cell composition in the context of ageing-related studies and the epigenetic clock is fundamental in order to make sure that the observed age-related changes in the methylome are not a direct consequence of the changes in blood cell composition during ageing [25, 32, 33].

Several methods have been developed to estimate the cell composition of a blood sample given a bulk DNA methylation measurement (a.k.a. cell-type deconvolution) [22, 34–36]. These methods can be broadly split in two categories:

- **Reference-based approaches**. They use a pre-defined set of DNA methylation reference profiles for the cell types that are supposed to be present in the tissue. In the case of methylation arrays, these reference profiles can be constituted by the β-values for a subset of array probes that are highly discriminative of the underlying cell types. Assuming that the blood sample is a weighted linear sum of the reference profiles, the objective of the method is to find these weights (w_c), which should be equivalent to the actual cell type proportions (given the assumption $\sum_{c=1}^{C} w_c \leq 1$) [22]. In mathematical terms:

$$\mathbf{y} = \sum_{c=1}^{C} w_c \mathbf{b_c} + \varepsilon \tag{2.6}$$

  where $\mathbf{y}$ is the DNA methylation profile of the sample being considered, C is the number of underlying cell types, $\mathbf{b_c}$ is the DNA methylation profile for the cth cell type and ε is the error [35]. Different algorithms have been applied to estimate the values of w_c, with the approach by Houseman *et al.* (which uses a linear constrained projection) [37] being the most widely used.

- **Reference-free approaches**. Instead of making use of reference profiles for the cell types of interest, these methods generally calculate latent variables that capture variation driven by cell type composition, although the strategy and assumptions to derive these latent variables from the DNA methylation data is highly method-specific [22]. These methods become particularly useful when no references are available for the cell types that constitute the tissue [22].

However, reference-free approaches rarely provide with estimates for the specific cell types in a given sample [22] (which are needed in the current modelling framework of the epigenetic clock) and they often rely on the assumption that the top components of variation correlate with cell composition [35], something that is not always true (especially in the case of developmental disorders, see Chapter 3), Thus, I decided to benchmark different reference-based cell-type deconvolution strategies in blood. In this context I tested (Fig. S1.4):

- **Different blood references**. As pointed out before, the quality of the reference, containing the DNA methylation profiles of the cell types to be inferred, is crucial [35, 38]. The reference must be composed of those CpG sites (in this case, array probes) that are able to better discriminate between the different cell types. In my case I considered 6 major blood 'cell types' for the inference: granulocytes ('Gran'), $CD4^+$ T cells ('CD4T'), $CD8^+$ T cells ('CD8T'), $CD19^+$ B cells ('B'), $CD14^+$ monocytes ('Mono') and $CD56^+$ natural killer cells ('NK'). It is important to point out that granulocytes are not themselves a 'biological cell type' (since they are composed of neutrophils, eosinophils and basophils), but will be considered as a 'computational cell type' as previously done [32, 33]. I tested three different blood references whose constitutive probes were selected using different strategies:

    1. The reference implemented in the *estimateCellCounts* function from the *minfi* R package [12], which is widely used in the epigenetic literature. The reference probes were selected using t-statistics, by finding those probes that were differentially methylated in each cell type when compared with the rest of the cell types. Among those probes that showed differences at p-value $< 10^{-8}$, the 100 most differentially methylated probes by effect size (50 hypermethylated and 50 hypomethylated) were chosen for each cell type (making a total of 600 probes for the reference) [25].

    2. The reference implemented in the *EpiDISH* R package (*centDHSbloodDMC.m*) [39]. The reference probes (DHS-DMCs, 333 in total) were selected by leveraging information of both differentially methylated cytosines (DMCs, using moderated

t-statistics) and chromatin accessibility (DNase Hypersensitive Sites or DHS) for each cell type [35].

3. The reference implemented as part of the IDOL strategy (IDentifying Optimal DNA methylation Libraries) [38]. In this case, the reference probes (300 in total) were originally selected based on differential methylation criteria and are updated in an iterative manner, with the probability of being selected based on their contribution to prediction accuracy [38].

The three references were built using the dataset from Reinius *et al.* (GSE35069) [23], which I obtained directly from the *FlowSorted.Blood.450k* R package [40]. This dataset contains DNA methylation data generated in the 450K array for the 6 cell types considered, that were isolated using flow cytometry [23]. The β-values for the selected probes were averaged across the biological replicates for each cell type.

- **Different DNA methylation pre-processing pipelines**. I tested different configurations for the pre-processing of both the gold-standard (see below) and the reference data. For example, I tested whether probe filtering according to the criteria outlined in the previous section (section 2.1.2) is desirable, since this leads to the removal of some of the probes originally selected for the reference in the original publications [35, 38] (Fig. S1.4). Furthermore, I also tested whether the prediction benefits from a similar pre-processing of both the gold-standard (or the dataset where the prediction will be made) and the reference.

- **Different deconvolution algorithms**. I tested the performance of the following algorithms: CP/QP (constrained projection/quadratic programming, originally implemented by Houseman *et al.* [37]), RPC (robust partial correlations) [35] and CIBERSORT (which was originally developed for cell-type deconvolution using RNA expression data) [35, 41]. One of the key differences between the algorithms is how the normalisation constrain ($\sum_{c=1}^{C} w_c \leq 1$) is implemented [35]. All the algorithms were run using the implementations in the *epidish* function from the *EpiDISH* R package [39], with the exception of the run in the *minfi* reference, for which I used the *estimateCellCounts* function with default parameters for the 450K array [12].

In order to compare the results from the predictions against real cell composition values, I used a **gold-standard** dataset (GSE77797) containing 12 samples where known proportions of DNA isolated from the different blood cell types were mixed [38]. I assessed the accuracy of the predictions using 3 different metrics:

- Root mean squared error (*RMSE*), which can be calculated as (for a given cell type *c*):

$$RMSE_c = \sqrt{\frac{\sum_{n=1}^{N}(\hat{y}_{cn} - y_{cn})^2}{N}} \tag{2.7}$$

  where $\hat{y}_{cn}$ is the predicted proportion of the *c*th cell type in the *n*th sample, y_{cn} is the real proportion of the *c*th cell type in the *n*th sample and N is the total number of samples in the gold-standard dataset ($N = 12$). A perfect prediction for a cell type would minimise the value of $RMSE_c$ (i.e. $RMSE_c = 0$).

- Mean absolute error (*MAE*), which can be calculated as (for a given cell type *c*):

$$MAE_c = \frac{\sum_{n=1}^{N}|\hat{y}_{cn} - y_{cn}|}{N} \tag{2.8}$$

  A perfect prediction for a cell type would minimise the value of MAE_c (i.e. $MAE_c = 0$).

- Coefficient of determination (R^2), which can be calculated as (for a given cell type *c*):

$$R_c^2 = \frac{\sum_{n=1}^{N}(\hat{y}_{cn} - \bar{y}_c)^2}{\sum_{i=1}^{N}(y_{cn} - \bar{y}_c)^2} \tag{2.9}$$

  where $\bar{y}_c = \frac{\sum_{n=1}^{N} y_{cn}}{N}$. A perfect prediction would maximise the value of R_c^2 (i.e. $R_c^2 = 1$).

The most accurate strategy, according to the *RMSE* (mean across cell types: 1.9270) and *MAE* (mean across cell types: 1.5498), is 'idol_NFB_houseman' (Fig. 2.4, Fig. S1.5) i.e. the strategy that uses the IDOL reference, with all the pre-processing steps from my main pipeline for both reference and gold-standard (*noob* background correction, probe filtering and BMIQ normalisation) and employs Houseman's CP/CQ algorithm (Fig. S1.4). This strategy performed well in all the cell types (Fig. 2.5) and I selected it for my cell-type deconvolution analyses.

It is important to mention that the gold-standard dataset was generated as part of the same study where the IDOL reference was also derived [38]. However, the gold-standard samples were used as an independent validation of the IDOL reference and should not have an influence on the conclusions of the benchmarking that I performed. In the future, it will

be interesting to validate these conclusions using new gold-standard datasets generated from whole blood.

Next, I ran the optimal blood cell-type deconvolution strategy in the DNA methylation dataset that I built from healthy individuals (Table 2.1). The main goal of this analysis was to provide blood cell type proportions that can be used as covariates as part of the epigenetic clock modelling (see section 2.2.2). However, this also allowed me to broadly quantify the **changes in blood composition that occur during human ageing** (Fig. 2.6). The mammalian immune system undergoes dramatic changes during ageing. These changes are normally referred as *immunosenescence* and can be broadly defined as a decline in immune system functionality and its ability to fight infections, which results in an increase in morbidity and mortality with age [42]. Furthermore, human ageing is also characterised by an increase in chronic, low-grade inflammation referred as *inflammageing*, which is thought to contribute to the development of age-related diseases (such as atherosclerosis, type 2 diabetes, Alzheimer's disease and osteoporosis) [43].

In my dataset, I observe the following (Fig. 2.6):

- A relative decrease in cell types from the adaptive immune system ($CD4^+$ T cells, $CD8^+$ T cells and $CD19^+$ B cells). Interestingly, the decline in $CD8^+$ T cells was more pronounced (i.e. higher absolute value of the slope) than in the case of $CD4^+$ T cells, which has been previously pointed out [27].

- A relative increase in cell types from the innate immune system (granulocytes, $CD14^+$ monocytes and $CD56^+$ natural killer cells).

These results are highly consistent with the literature [25, 27–31], which validates the methodology for cell-type deconvolution that I have used. These variations in blood cell composition may be caused by the age-related changes that happen in the two primary lymphoid organs: the bone marrow (whose hematopoietic stem cells exhibit reduced self-renewal potential and increased skewing towards myelopoiesis) and the thymus (which undergoes tissue involution) [44].

This analysis provides a preliminary overview of the blood composition landscape during human ageing. However, only relative changes in blood composition were quantified and the analysis is limited by the 'cell types' that I have deconvoluted (e.g. granulocytes include different cell types, different subsets of monocytes exist, ...), which means that these conclusions must be taken with care [42]. Furthermore, the sex of the individual can influence the proportions of blood leukocytes [29] and it should be taken into account in future analyses.

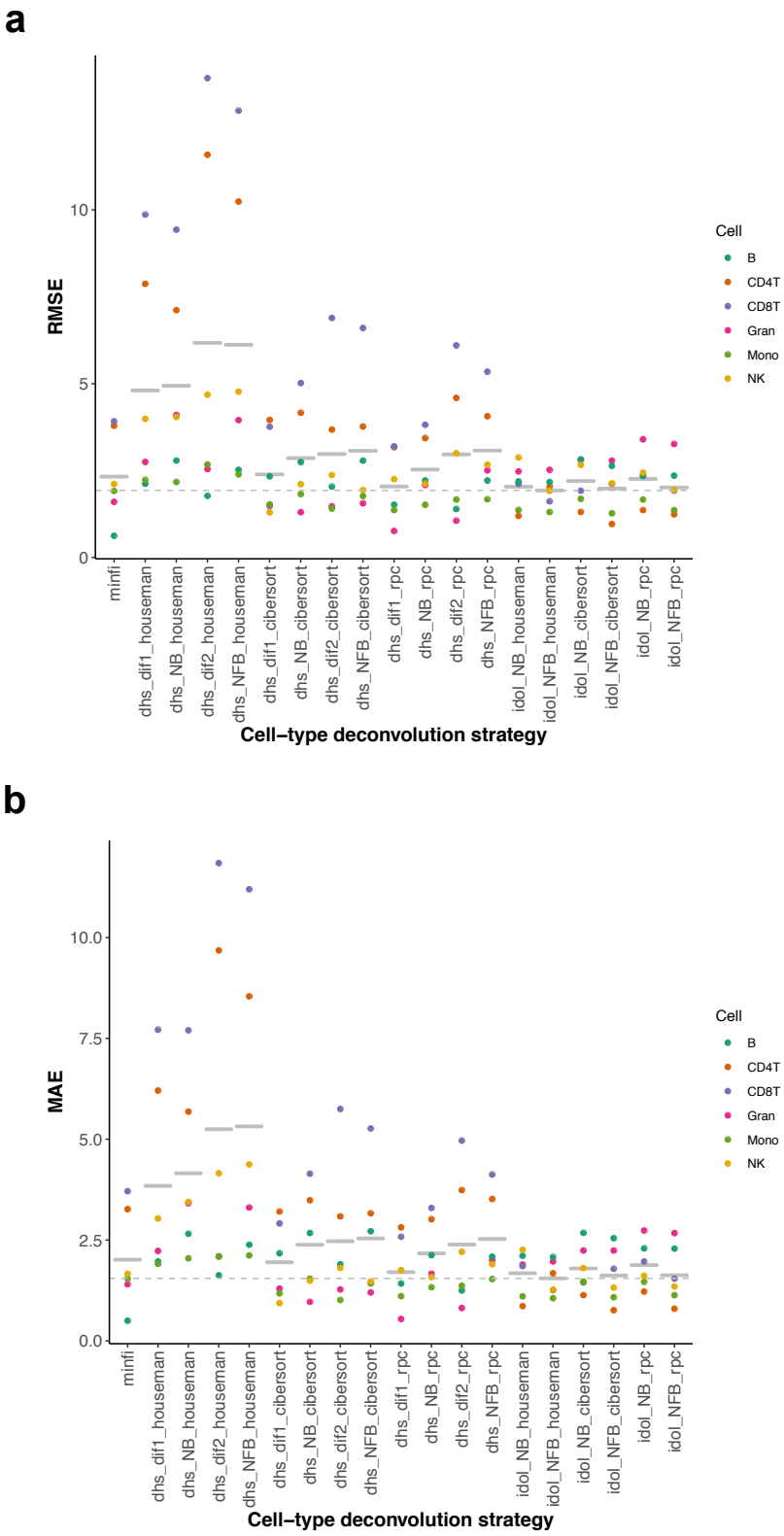


Fig. 2.4 Benchmarking of the cell-type deconvolution strategies in blood. The x-axis shows the different strategies that were tested (for a detailed description see Fig. S1.4). The y-axis shows the results for **a.** the root mean squared error (*RMSE*) and **b.** the mean absolute error (*MAE*) when comparing the predictions with the real proportions of cells in a gold-standard dataset (GSE77797) [38]. The grey horizontal solid lines represent the mean for the *RMSE* or the *MAE* across cell types and the grey dashed line the minimum of these values.

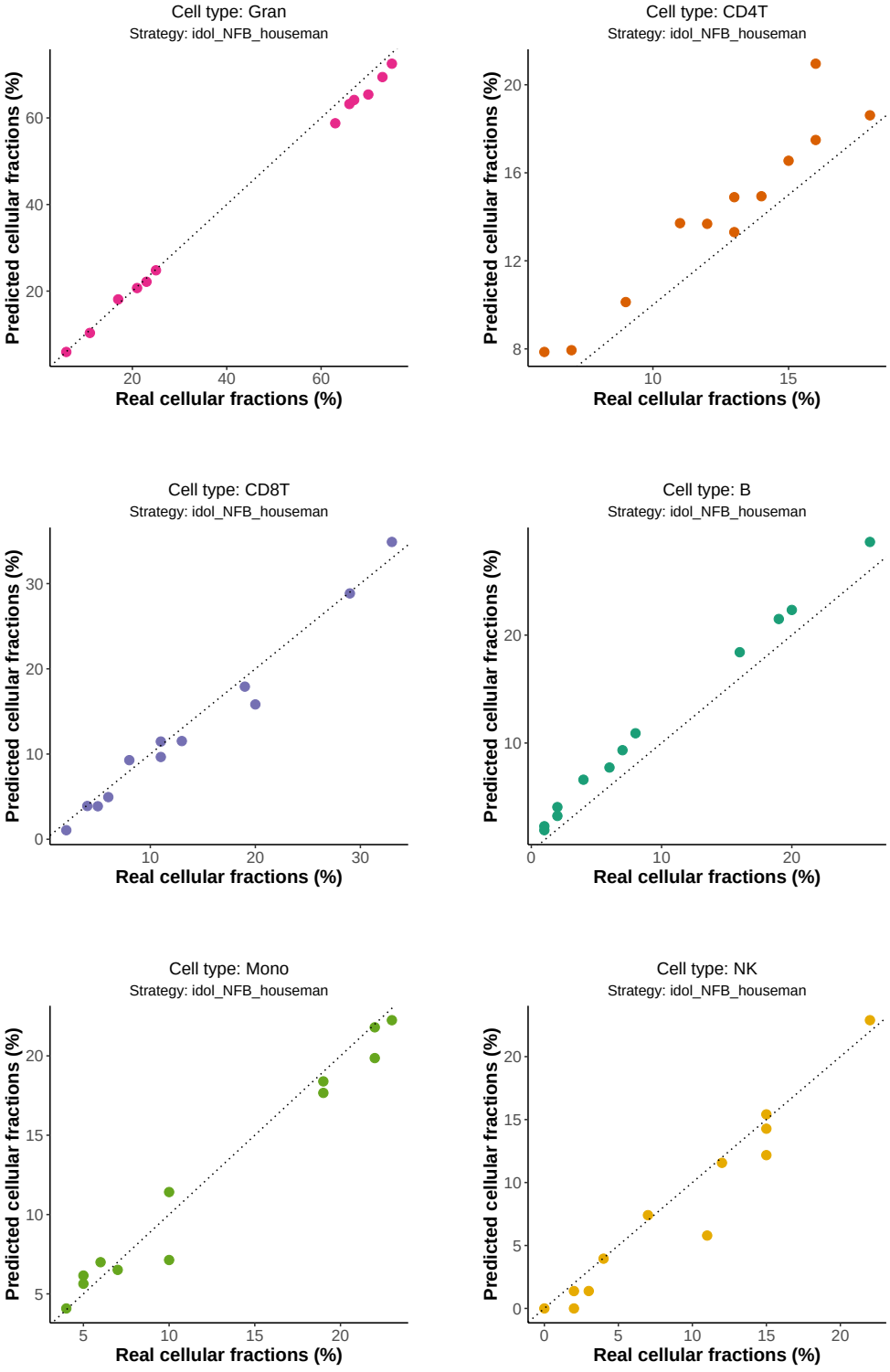


Fig. 2.5 Comparison of the predictions for the different cell types using the optimal deconvolution strategy ('idol_NFB_houseman') with the real cell type fractions in the gold-standard dataset (GSE77797) [38]. Each point corresponds to a different sample in the gold-standard. The black dashed line represents the diagonal to aid visual interpretation.

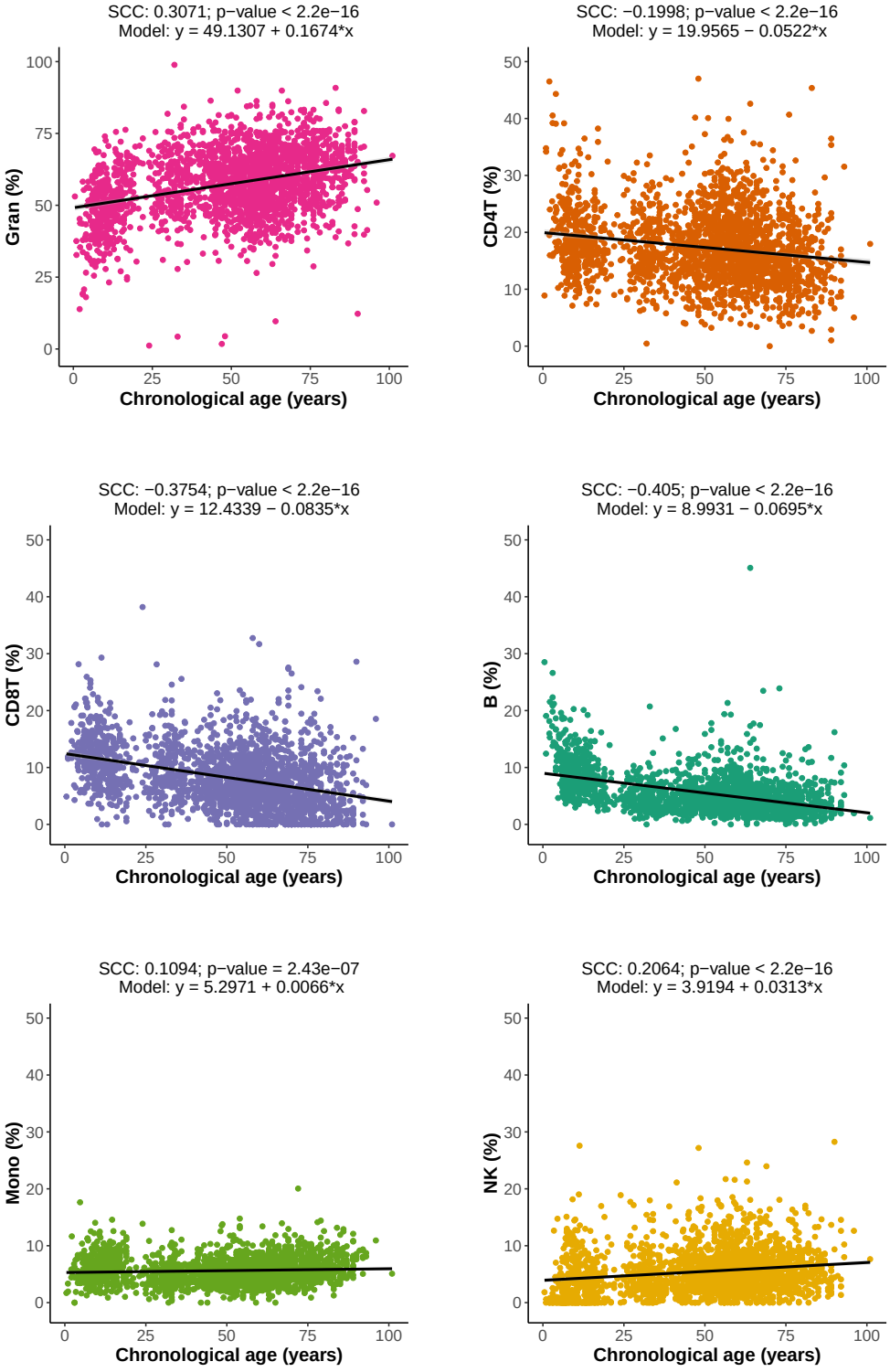


Fig. 2.6 Changes in blood cell composition during human ageing. Scatterplots showing the changes in the proportions of the 6 cell types considered (inferred using the cell-type deconvolution strategy) as a function of chronological age. Each point represents a different DNA methylation human sample from Table 2.1. The black line displays the linear model %cell_type $\sim$ Age (see section 2.4 for more details on linear modelling), with the slope and intercept shown in the titles. The Spearman's correlation coefficient (SCC) and the p-value associated with it are also displayed.

2.1.4 Identifying differentially methylated positions during ageing

Differential methylation analysis is one of the most common types of downstream analyses in the context of DNA methylation data [9, 10, 36]. It involves finding associations between the DNA methylation levels at specific CpG sites in the genome (a.k.a. differentially methylation positions or DMPs) and a given phenotypic variable of interest (e.g. a specific disease, when compared with a healthy sample). It is worth mentioning that DMPs are also called differentially methylated cytosines (DMCs) in the literature [36].

In order to study the changes that the methylome undergoes during physiological ageing, it is useful to identify differentially methylated positions during ageing (aDMPs) i.e. individual cytosines (normally found in a CpG context) that change their methylation status as a function of chronological age. Linear models, widely used in the context of differential RNA expression analysis [45], can also been adapted to find aDMPs [17, 36]. In the case of a continuous variable (such as chronological age) the association is performed using a linear regression modelling framework [17] (see section 2.4 for a short description of linear regression and the nomenclature used through this thesis). Briefly, for each probe in the methylation array, I fitted the following **linear regression models** to the data from the healthy individuals:

- A model with cell composition correction (CCC). As I have shown previously, the different blood cell types change their abundance with age. Therefore, in order to maximise the chances of finding aDMPs that are conserved across different cell types, it is important to include the estimated cell proportions as covariates in the model:

$$Beta \sim Age + Sex + Gran + CD4T + CD8T + B + Mono + NK + PC1 + ... + PC17 \tag{2.10}$$

  where *Beta* is the β-value for the array probe being evaluated; *Age* is the chronological age (in years) of the samples; *Sex* encodes for the sex of the samples (0/1); *Gran*, $CD4T$, $CD8T$, *B*, *Mono* and *NK* are the cell type proportions from the samples as calculated with my cell-type deconvolution strategy and *PCN* is the *N*th principal component that captures technical variance and accounts for potential batch effects (see section 2.2.3 for more details).

- A model without CCC, which can be expressed as:

$$Beta \sim Age + Sex + PC1 + ... + PC17 \tag{2.11}$$

This leads to the identification of aDMPs which will be more confounded with the proportions of the different cell types (i.e. the change in β-value with age could be entirely driven by a change in a specific cell type that is differentially methylated at that particular probe).

Furthermore, for each probe, I calculated a p-value, based on t-statistics [36], to assess whether the putative linear association between the methylation status and chronological age was significant or not (at a significance level of $\alpha = 0.01$ after applying Bonferroni correction to account for multiple testing, see section 2.4 for more details). I used a customised version of the *dmpFinder* function in the *minfi* R package [12] to identify the aDMPs, which internally uses the *limma* framework [45]. Given the big sample size ($N = 2218 \gg 10$), I did not use variance shrinkage (i.e. empirical Bayes moderated t-statistics) as part of the statistic calculations [45].

An overview of the different aDMPs (with and without CCC) identified in the healthy individuals can be found in Figure 2.7. Around 30% of the blood methylome (at least according to the 450K array) is affected by the ageing process during human lifespan. CpG sites can become both hypomethylated (i.e. lose methylation with age) or hypermethylated (i.e. gain methylation with age). Importantly, the effect sizes of the age coefficient (i.e. the observed changes in the β-values per year) are generally small. More specifically, in the model with CCC, the median age coefficient for the hypomethylated aDMPs is -0.000426 (equivalent to a -4.26% methylation change over 100 years of human life) and for hypermethylated aDMPs is 0.000437 (equivalent to a +4.37% methylation change over 100 years of human life). This is consistent with the progressive functional decline observed during ageing [46]. It is worth mentioning that around 50% of the CpG sites that constitute the Horvath epigenetic clock are blood aDMPs according to our analysis (Fig. 2.7c,d). Overall, these results are consistent with previous studies [47–50].

Next, I looked at the top 100 aDMPs that were identified (according to their p-value and t-statistic, Fig. S1.6 and Fig. 2.8). The first aDMP in the list was cg16867657, a probe that consistently gains methylation with age (Fig. 2.8a) and has been previously identified as the strongest aDMP across tissues and human populations in several studies [48, 52–56]. cg16867657 is associated with the CpG island in the promoter of the ELOVL2 gene, which encodes an enzyme that catalises one of the reactions in the elongation of polyunsaturated

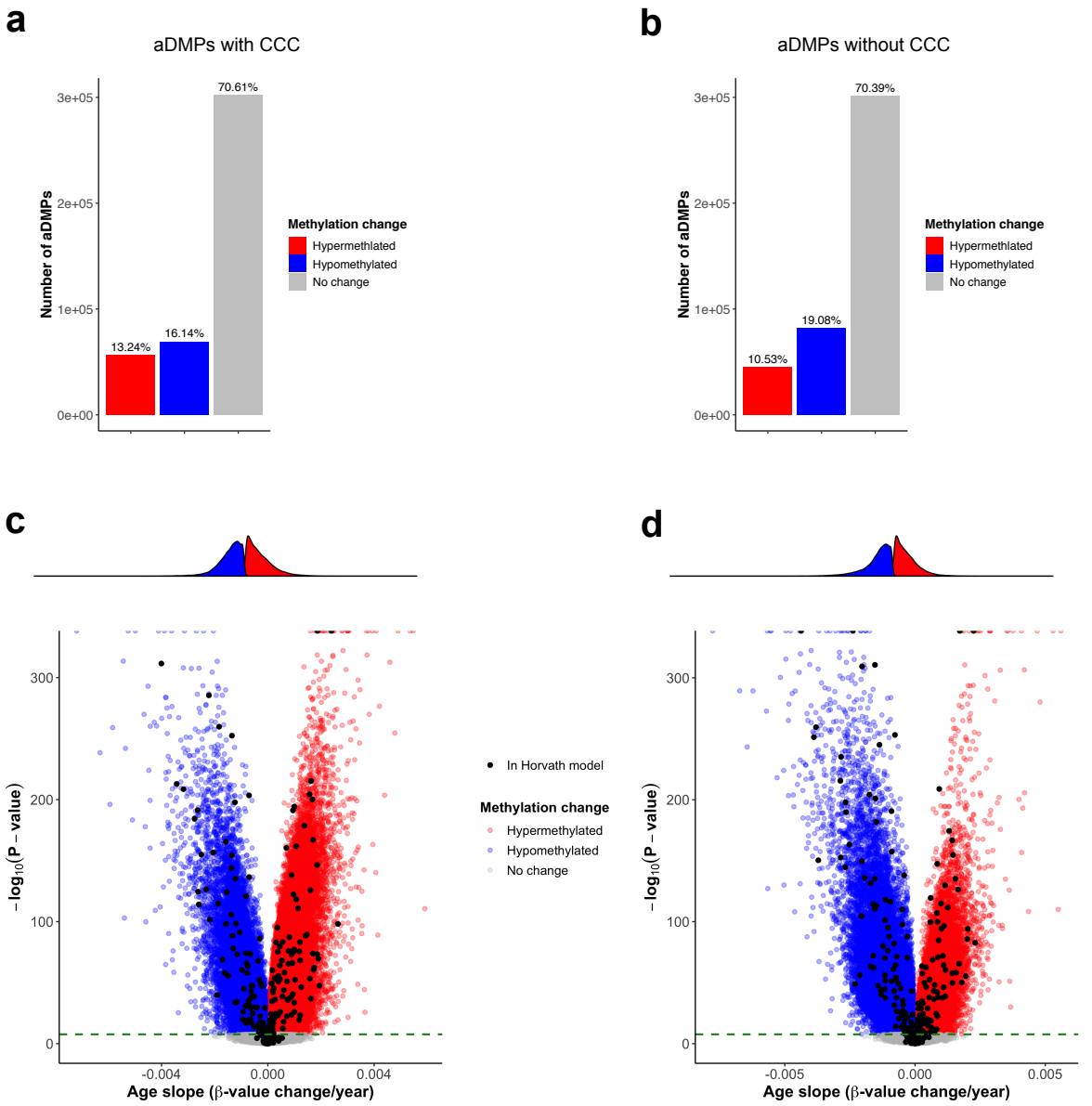


Fig. 2.7 The blood methylome changes during physiological human ageing. **a.** Barplot showing the total number of differentially methylated positions during ageing (aDMPs) that were identified (in grey: probes that did not reach statistical significance). In this case, the model with cell composition correction (CCC) was applied. **b.** As in a., but using the model without CCC. **c.** Volcano plot showing the relationship between the p-value (y-axis) and the effect size (x-axis) of the age coefficient for each one of the array probes (each point represents a probe). Those probes above the dashed green line ($\alpha = 0.01$ after Bonferroni correction) are the identified aDMPs. Above the volcano plot, a density plot captures the distributions of the age coefficient for the hypermethylated aDMPs (in red) and the hypomethylated aDMPs (in blue). In this case, the model with CCC was applied. The black points are the 353 CpG probes that constitute the Horvath epigenetic clock model [51]. **d.** As in c., but using the model without CCC.

fatty acids [55]. Furthermore, other aDMPs that were located among my top hits have previously been reported as well (such as cg06639320 in the FHL2 gene, which is the second aDMP, Fig. 2.8b) [52]. These results validate the statistical methods used so far to process the DNA methylation data and to identify aDMPs.

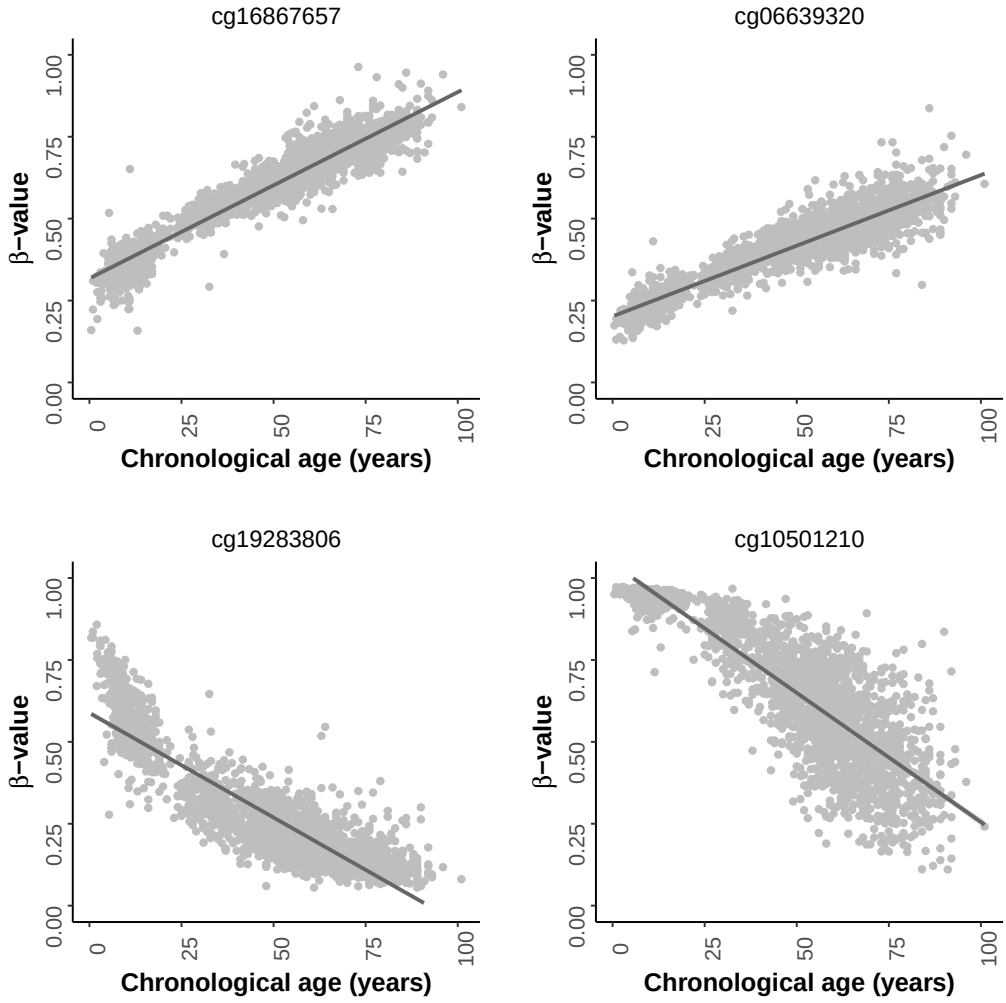


Fig. 2.8 Changes in the β-values of four different differentially methylated positions during ageing (aDMPs) in the blood of the healthy individuals. cg16867657 and cg06639320 are the top aDMPs that gain methylation with age (i.e. become hypermethylated) according to the model that accounts for cell composition correction (CCC). cg19283806 and cg10501210 are the top aDMPs that lose methylation with age (i.e. become hypomethylated) according to the model that accounts for CCC. In order to aid visualisation, the black line displays the linear model β-value $\sim$ Age.

It is important to mention that not all the CpG sites change their DNA methylation levels with age in a perfectly linear manner. For instance, the two top hypomethylated aDMPs (Fig. 2.8c,d) modify their rate at ages 20-25 years. This was already recognised

by Horvath [51] and that's why the age is transformed into a logarithmic scale before the age of 20 years in order to improve the model fit (see section 2.2.1). Furthermore, genetic background can have a significant effect on the DNA methylation patterns and interact with the ageing process to shape the epigenome [50, 56]. Unfortunately, I did not have genetic data for the healthy individuals but this could help to refine the identification of aDMPs in the future. Additionally, it would be interesting to apply methods to control for bias and inflation in the test statistics, by estimating the empirical null distribution [57]. Finally, other types of epigenetic features can be derived to understand the effects of ageing in the epigenome, such as variably methylated positions during ageing (aVMPs) [47], differentially methylated regions (DMRs, which consider several correlated CpGs at the same time) [36] or differentially methylated cytosines in individual cell types (DMCTs, which consider interactions between the phenotypic variable and the proportions of cell types) [58].

2.1.5 Shannon methylation entropy

Shannon entropy (H) can be used in the context of DNA methylation analysis to estimate the information content stored in a given set of CpG sites [47, 56, 59–61]. I calculated it using the same approach as in Hannum *et al.* [56]:

$$H = -\frac{1}{N} \cdot \sum_{i=1}^{N} [\beta_i \cdot \log_2(\beta_i) + (1-\beta_i) \cdot \log_2(1-\beta_i)] \tag{2.12}$$

where β_i represents the methylation β-value for the ith array probe (or CpG site) and $N = 428266$ if all the array probes that survived the pre-processing pipeline are considered (i.e. genome-wide, ot at least array-wide). Shannon entropy is minimised when the methylation levels of all the CpGs are either 0% or 100%. and maximised when all of them are 50% (Fig. 2.9).

Next, I calculated the genome-wide Shannon entropy for the blood samples in the healthy individuals. Consistent with previous reports [47, 56, 59, 61], the genome-wide Shannon entropy associated with the methylome increases during ageing (Fig. 2.10a; Spearman correlation coefficient $= 0.1985$; p-value $= 3.8281 \cdot 10^{-21}$), which implies that the epigenome loses information content. Finally, it is worth mentioning that I observed a remarkable effect of the batch on the Shannon entropy calculations, which can generate high entropy variability for a given age (Fig. 2.10b). However, after removing potential outlier batches (such as GSE41273, GSE59065 or GSE97362) the increase of Shannon methylation entropy during

ageing was still consistent. Thus, accounting for technical variation (see section 2.2.3) becomes crucial when assessing this type of data, even after careful pre-processing.

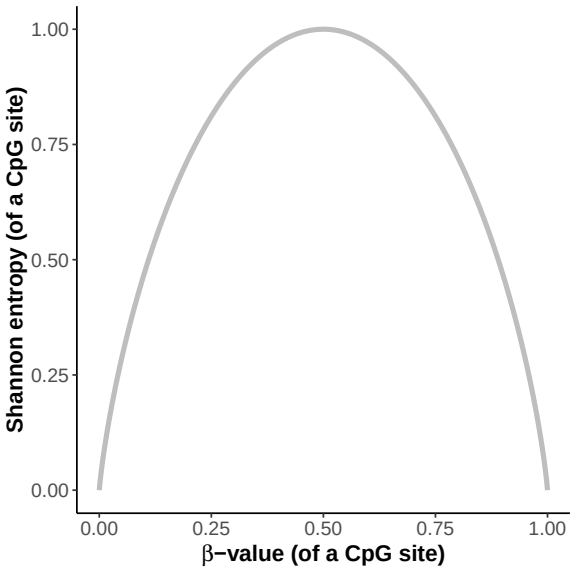


Fig. 2.9 Plot showing the relationship between the β-value and the methylation Shannon entropy at a given CpG site (in my case, at a given array probe).

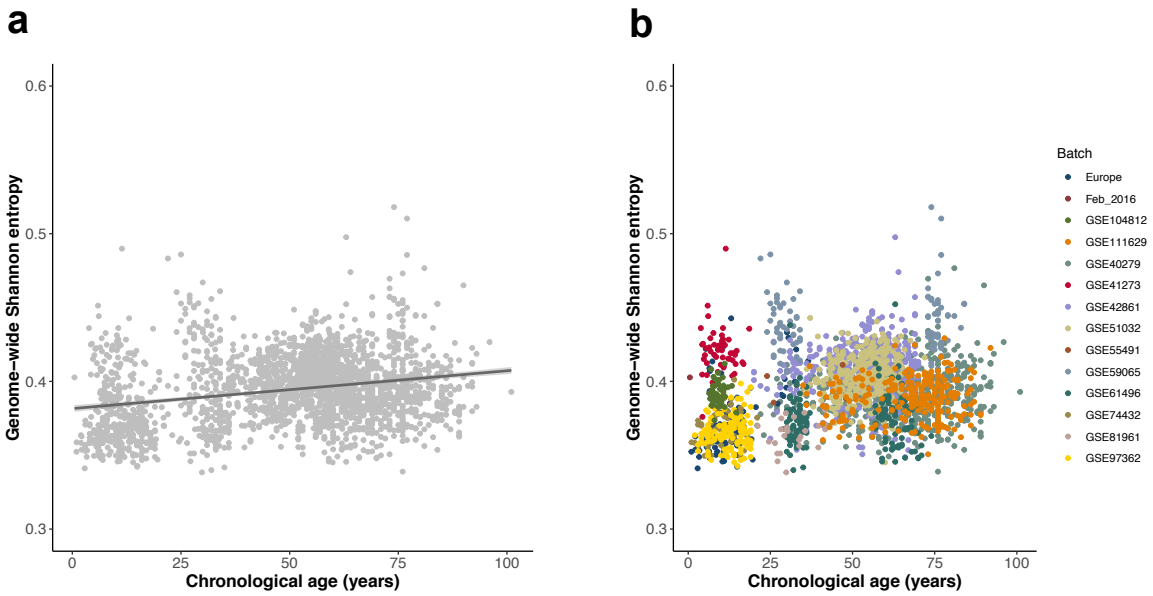


Fig. 2.10 a. Scatterplot showing the changes in genome-wide methylation Shannon entropy during ageing in the healthy individuals. Each sample is represented by one point. The black line displays the linear model Entropy $\sim$ Age. **b.** Same as in a., but colouring the samples according to the batch where they came from.

2.2 Behaviour of Horvath's epigenetic clock during ageing

2.2.1 Calculating epigenetic age using Horvath's epigenetic clock

Steve Horvath's model, originally published in 2013 [51], is without any doubt the most widely used epigenetic clock in the literature. Given that it works across tissues with high accuracy and that it has been validated in many human cohorts, I have used it as the main tool to quantify epigenetic ageing in this work.

Horvath's model measures epigenetic age (a.k.a. *DNAmAge*) by making use of the DNA methylation levels at 353 CpG sites, as quantified with the Illumina methylation arrays (27K or 450K). Previous studies have generally employed a ready-to-use online calculator for *DNAmAge* provided by Steve Horvath [62]. This has clearly simplified the computational process and helped a lot of research groups to test the behaviour of the epigenetic clock in their system of interest. However, this has also led to the treatment of the epigenetic clock as a 'black-box', without critical assessment of the statistical methodology behind it. Therefore, I decided to replicate the original code and to make it available in a GitHub repository for the scientific community to be used [63]. Furthermore, I tested the impact of different steps involved in the estimation of epigenetic age acceleration (EAA), including the presence/absence of background correction, removal of technical variation from batch effects and the importance of the age distribution when fitting the control models, which I discuss in the following sections.

The main pipeline to calculate the epigenetic age (*DNAmAge*) from a sample has the following steps (some of them are shared with the previously described pipeline for DNA methylation pre-processing in section 2.1.2):

1. **Background correction**. I implemented a pipeline that starts with the raw DNA methylation data (IDAT files) for a sample. First, I tested what was the effect of applying *noob* background correction, before calculating the β-values, on the median absolute error (MAE) of the predictions (see section 2.2.2). Background correction did not have a major impact in the final predictions as long as I also corrected for batch effects (Fig. S1.7, Fig. 2.13c, see section 2.2.3). Therefore, I decided to keep the *noob* background correction for consistency with the other pre-processing pipeline.

2. **Quality control**. I applied the same criteria as previously described in section 2.1.2.

3. **Probe filtering**. Horvath's model was originally trained starting with 21368 array probes that had the following characteristics [51]:

- They were shared between the 27K and 450K methylation arrays.
- They had ≤ 10 missing values across all the training data.

Therefore, these were the probes selected for downstream analysis.

4. **β-value calculation**. β-values were calculated as previously described in section 2.1.2. It is worth mentioning that Horvath's original code includes two alternatives for the imputation of missing β-values:

   - Slow imputation (applied when the number of missing β-values is < 3000). In this case, *k*-nearest neighbours (KNN) is used. KNN imputation borrows information from the DNA methylation profiles of the most similar probes (the neighbours) according to a metric (normally the Euclidean distance). The *impute.knn* function from the *impute* R package can be used for these purposes [64].
   - Fast imputation (applied when the number of missing β-values is ≥ 3000). In this case, the values from the blood gold-standard (see below) can be used as the imputed values.

   In the case of my dataset, no missing values were present for the 21368 probes so there was no need to perform imputation.

5. **Gold-standard normalisation**. A modified version of BMIQ normalisation is used [18]. In this case, instead of mapping the distribution of the Infinium II probes to the distribution of Infinium I probes, the mapping is done from the distribution of the 21368 probes in the sample to the distribution of a previously derived gold-standard for the same set of probes. This gold-standard was created by taking the average β-values for the 21368 probes across all the whole blood samples from [65].

6. **Calculating epigenetic age** (*DNAmAge*). As previously observed for some of the aDMPs, the rate of β-value change can be different before and after adult age (Fig. 2.8). For this reason, Horvath performed a transformation of the chronological age before training the model:

$$f(c) = c_t = \begin{cases} \ln\left(\frac{c+1}{a+1}\right) & \text{if: } c \leq a \\ \left(\frac{c-a}{a+1}\right) & \text{if: } c > a \end{cases} \tag{2.13}$$

where c_t is the transformed chronological age that was used as the dependent variable during training, c is the chronological age (in years) and a is the adult age (for humans, 20 years). This transformation allows to account for a relationship between chronological age and methylation changes that is logarithmic until adult age and linear afterwards (Fig. 2.11).

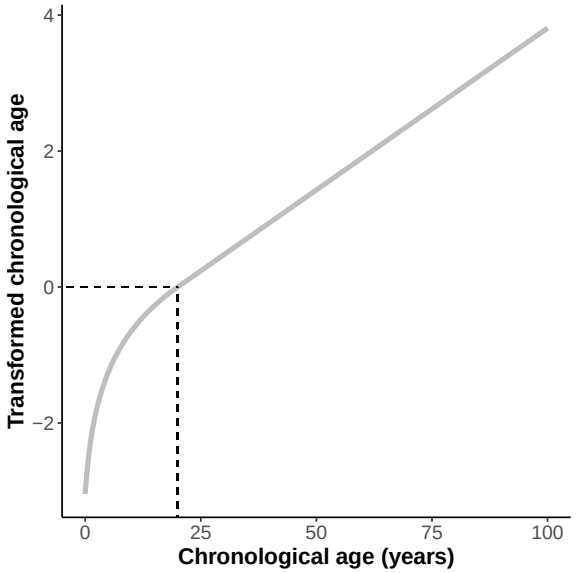


Fig. 2.11 Plot showing the relationship between the chronological age in years (c) and the transformed chronological age (c_t) in Horvath's model. This transformation allows accounting for different rates of β-value change before and after adult age (20 years in humans, as pointed out by the dashed black line).

Given a sample to predict, the epigenetic age can then be calculated as:

$$DNAmAge = g(\hat{c}_t) = g(\hat{\beta}_0 + \sum_{i=1}^{353} \hat{\beta}_i \cdot x_i) \tag{2.14}$$

where $\hat{c}_t$ is the predicted transformed age according to Horvath's model, $\hat{\beta}_0$ is the intercept in the Horvath's model, $\hat{\beta}_i$ is the coefficient (weight) for the ith probe (only 353 probes are finally used), x_i is the β-value for the ith probe after gold-standard normalisation and $g(\cdot)$ if the inverse of $f(\cdot)$, such that:

$$g(\hat{c}_t) = f^{-1}(\hat{c}_t) = \hat{c} = \begin{cases} e^{\hat{c}_t} \cdot (a+1) - 1 & \text{if: } \hat{c}_t \leq 0 \\ \hat{c}_t \cdot (a+1) + a & \text{if: } \hat{c}_t > 0 \end{cases} \tag{2.15}$$

where $\hat{c}$ is the predicted age according to Horvath's model (i.e. *DNAmAge*).

2.2.2 Horvath's epigenetic clock measures physiological ageing

Using the methodology from the previous section, I calculated the epigenetic age (*DNAmAge*) in the blood of the healthy individuals. Given that these individuals are supposed to be disease-free, Horvath's epigenetic clock should predict epigenetic ages that are similar to the chronological age of the samples, and this was indeed the case (Fig. 2.12a, Pearson's correlation coefficient (PCC) $= 0.9671$, p-value ≈ 0). This validates that Horvath's epigenetic clock does indeed measure the ageing process (at least in a cross-sectional population) and sets a foundation for the rest of the analyses presented in this thesis.

As mentioned in Chapter 1, the difference between epigenetic age and chronological age is known as **epigenetic age acceleration** (EAA), with a positive EAA (i.e. $DNAmAge > Age$) associated with several age-related health problems. In order to calculate the EAA for our healthy individuals, I fitted the following linear regression models (hereinafter referred as the *control models*):

- With cell composition correction (CCC):

$$DNAmAge \sim Age + Sex + Gran + CD4T + CD8T + B + Mono + NK + PC1 + ... + PC17 \tag{2.16}$$

where *DNAmAge* is the epigenetic age calculated with Horvath's epigenetic clock; *Age* is the chronological age (in years) of the samples; *Sex* encodes for the sex of the samples (0/1); $Gran$, $CD4T$, $CD8T$, B, $Mono$ and NK are the cell type proportions from the samples as calculated with my cell-type deconvolution strategy and *PCN* is the *N*th principal component that captures technical variance and accounts for potential batch effects (see section 2.2.3 for more details).

Horvath's epigenetic clock was trained using multiple tissues and its predictions should be robust to changes in blood cell composition. However, previous studies have highlighted that adding this correction can improve the ability to detect 'pure' ageing effects [32, 33] (i.e. epigenetic age acceleration mainly caused by DNA methylation changes that happen in the nucleus of all cell types). For a given sample, the $\text{EAA}_{\text{with CCC}}$ is the residual from the model i.e. the difference between the actual *DNAmAge* and the

prediction from the control model (which is conceptually similar to the difference between *DNAmAge* and chronological age, but accounting for the rest of covariates as well). The $\text{EAA}_{\text{with CCC}}$ that I have defined is very similar to the previously reported measure of 'intrinsic EAA' (IEAA) [32, 33].

- Without CCC:

$$DNAmAge \sim Age + Sex + PC1 + ... + PC17 \tag{2.17}$$

In this case the residuals of the model are referred as the $\text{EAA}_{\text{without CCC}}$ for the different samples.

It is possible to calculate the overall accuracy of the predictions using the median absolute error (MAE), that can be calculated as:

$$MAE = \text{median}\{|EAA_i|\} \tag{2.18}$$

where EAA_i is the epigenetic age acceleration for the ith sample calculated with one of the models (with CCC or without CCC). The MAE for all the healthy individuals (full lifespan) in the control models should be close to zero, and this was indeed what I observed ($MAE_{\text{with CCC}} = 2.7117$ years, $MAE_{\text{without CCC}} = 2.8211$ years). These results are below the original MAE reported by Horvath in his test set (3.6 years) [51]. However, it is worth mentioning that some of the samples from my healthy individuals (such as samples from batches GSE40279 and GSE42861) could have been used by Horvath as part of his training set [51], and therefore these results must be interpreted carefully.

Even though Horvath's model seems to predict epigenetic age accurately, it is also clear that some samples deviate substantially from the expected prediction. This is specially obvious for the older samples (> 55 years), that have a systematically younger epigenetic age than expected (see deviations from the diagonal in Fig. 2.12a). If a control model is fit to the full lifespan dataset (which contains around 50% samples which are > 55 years), this leads to a model with a smaller thane expected age coefficient (slope), which introduces a bias when estimating epigenetic age acceleration for different age groups (Fig. 2.12b). Although many studies do not take this problem into account, this phenomenon has been previously reported in the context of humans [66, 67] and mice [68]. However, to this date, it is unclear whether

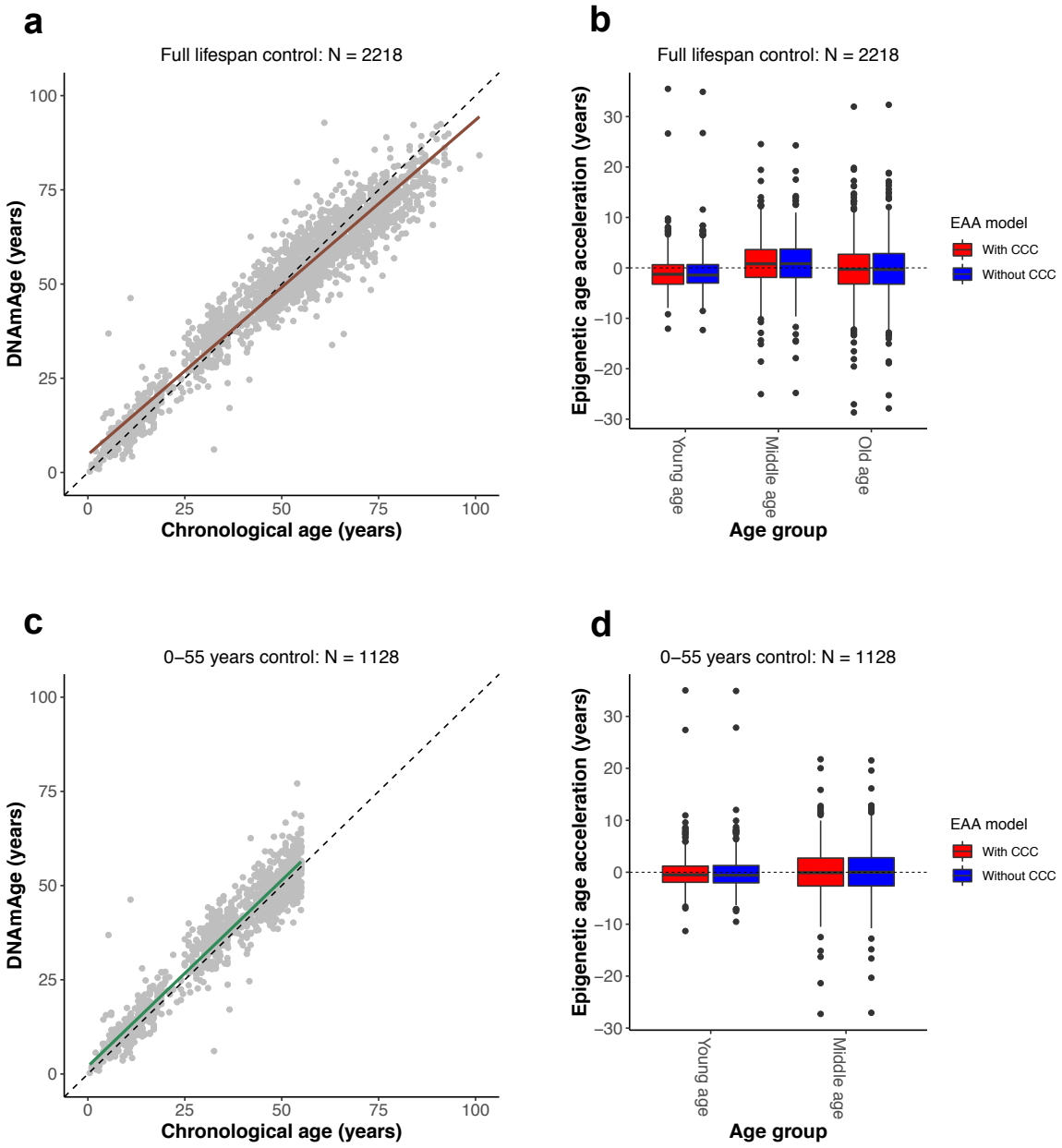


Fig. 2.12 Horvath's epigenetic clock measures physiological ageing. **a.** Scatterplot showing the relationship between epigenetic age (*DNAmAge*) according to Horvath's model [51] and chronological age of the samples for the healthy individuals. Each sample is represented by one point. The black dashed line represents the diagonal to aid visualisation. The solid brown line represents the linear model DNAmAge $\sim$ Age, which deviates from the diagonal if the full lifespan samples are used. **b.** Boxplots displaying the epigenetic age acceleration (EAA) distributions for different age ranges (young age: ≤ 20 years; middle age: $20 < Age \leq 55$ years; old age: > 55 years) after fitting the control models to the full lifespan samples. The dashed black line represents $EAA = 0$, where the distributions should be centred around. This is not the case for the samples in the young age and middle age groups. In red: EAA model with cell composition correction (CCC). In blue: EAA model without CCC. **c.** As in a., but removing the samples in the old age group (> 55 years). The solid green line represents the linear model DNAmAge $\sim$ Age, which is much more similar to the diagonal if only young and middle age samples are considered. **d.** As in b., but fitting the control models to the samples in the young and middle age groups (0-55 years). The bias in the EAA is corrected in this case (the distributions are centred around zero for the different age groups).

it represents a technical artefact or it has a biological explanation (e.g. survivor bias of the older individuals, the molecular processes that drive ageing slow down with age, ...).

This highlights the importance of having a properly age-matched control when performing analyses with the Horvath's epigenetic clock. As expected, removing the older samples (> 55 years) from the control models corrected for this bias (Fig. 2.12c,d) and reduced the *MAE* ($MAE_{\text{with CCC}} = 2.2742$ years, $MAE_{\text{without CCC}} = 2.3237$ years). This is the strategy that I used when screening for epigenetic age acceleration in the context of developmental disorders (see Chapter 3).

2.2.3 Correcting for batch effects in the context of the epigenetic clock

As mentioned in the previous section, it is expected that, after fitting the control models, the EAA distributions of the samples from the healthy individuals should be centred around zero. However, if the principal components (PCs) that capture technical variation were not included in the control models (see equations 2.16 and 2.17), this was not the case for several batches (Fig. 2.13a, Fig. S1.8a). Therefore, I hypothesised that technical variation can affect the predictions from Horvath's epigenetic clock and that batch effects need to be explicitly accounted for in this context, even after applying the internal normalisation step against the blood gold-standard [51]. This section explains how I implemented this batch effect correction (i.e. how I derived the principal components that capture technical variance across batches).

A batch effect is a systematic technical source of variation that is unrelated to the biological or scientific variables in a study [69]. They affect low- and high-throughput measurements and can be caused by a wide variety of situations: different technicians performing the experiments, different laboratories generating the data, different lots of reagents or arrays used, ... [69]. Correcting for bath effects is crucial, especially when integrating data from different studies and sources [70], as it is the case in the analyses presented in this thesis. Data generated by DNA methylation arrays is also affected by batch effects and several methods have been described in the literature to correct for them, normally at the level of probe intensities [71] or M-values [70, 72]. In the context of the epigenetic clock, previous attempts to account for technical variation have used the first 5 principal components (PCs) estimated directly from the DNA methylation data (presumably the β-values) [73]. However, this approach potentially removes meaningful biological variation, especially in studies where there are global changes in DNA methylation, such as cancer [71] or developmental disorders (see Chapter 3). Furthermore, given that Horvath's epigenetic

clock was trained with data pre-processed using different strategies, it is unclear how applying an additional batch effect correction step to the intensities or β-values would impact the predictions [74].

Thus, I decided to correct for the potential batch effects when fitting the control models (see equations 2.16 and 2.17). I make use of the control probes present on the 450K array, which have been shown to carry information about unwanted variation from a technical source (i.e. technical variance) [70, 71, 75]. These probes are designed to capture technical variance in negative controls, measure between-array differences and quantify the performance of different steps of the array protocol, such as bisulfite conversion, staining or hybridisation [71, 76]. I performed principal component analysis (PCA, with centering but not scaling using the *prcomp* function in R) on the raw intensities of the control probes (847 probes $\cdot$ 2 channels $=$ 1694 intensity values) for all the healthy individuals ($N = 2218$) and the samples with developmental disorders (cases, $N = 666$, see Chapter 3). This showed that the first two PCs capture the batch structure in both healthy individuals (Fig. 2.13b) and cases (Fig. S1.9). Including the first 17 PCs as part of the epigenetic age acceleration (EAA) modelling (see equations 2.16 and 2.17), which together accounted for 98.06% of the technical variance in all the samples (Fig. S1.10), significantly reduced the median absolute error (MAE) of the predictions in the healthy individuals ($MAE_{\text{with CCC}} = 2.7117$ years, $MAE_{\text{without CCC}} = 2.8211$ years, mean $MAE = 2.7664$, Fig. 2.13c). Notably, the reduction in the MAE provided by the batch effect correction was higher than the improvement provided by cell composition correction, a common practice in the epigenetic clock field [32, 33]. The optimal number of PCs was found by making use of the *findElbow* function from [77].

Finally, deviations from a median EAA close to zero in some of the batches after batch effect correction (Fig. 2.13d, Fig. S1.8b) could be explained by other variables, such as a small batch size or an overrepresentation of young samples (Fig. 2.14). The latter is a consequence of the fact that Horvath's model underestimates the epigenetic ages of older samples, which I have discussed in the previous section. Thus, I have shown that correcting for batch effects in the context of the epigenetic clock is important, especially when combining datasets from different sources for meta-analysis purposes. Batch effect correction is essential to remove technical variance that could affect the epigenetic age of the samples and confound biological interpretation. Furthermore, given the flexibility of this modelling approach, I have applied batch effect correction across other types of analyses in the thesis, such as DMPs identification (see equation 2.10).

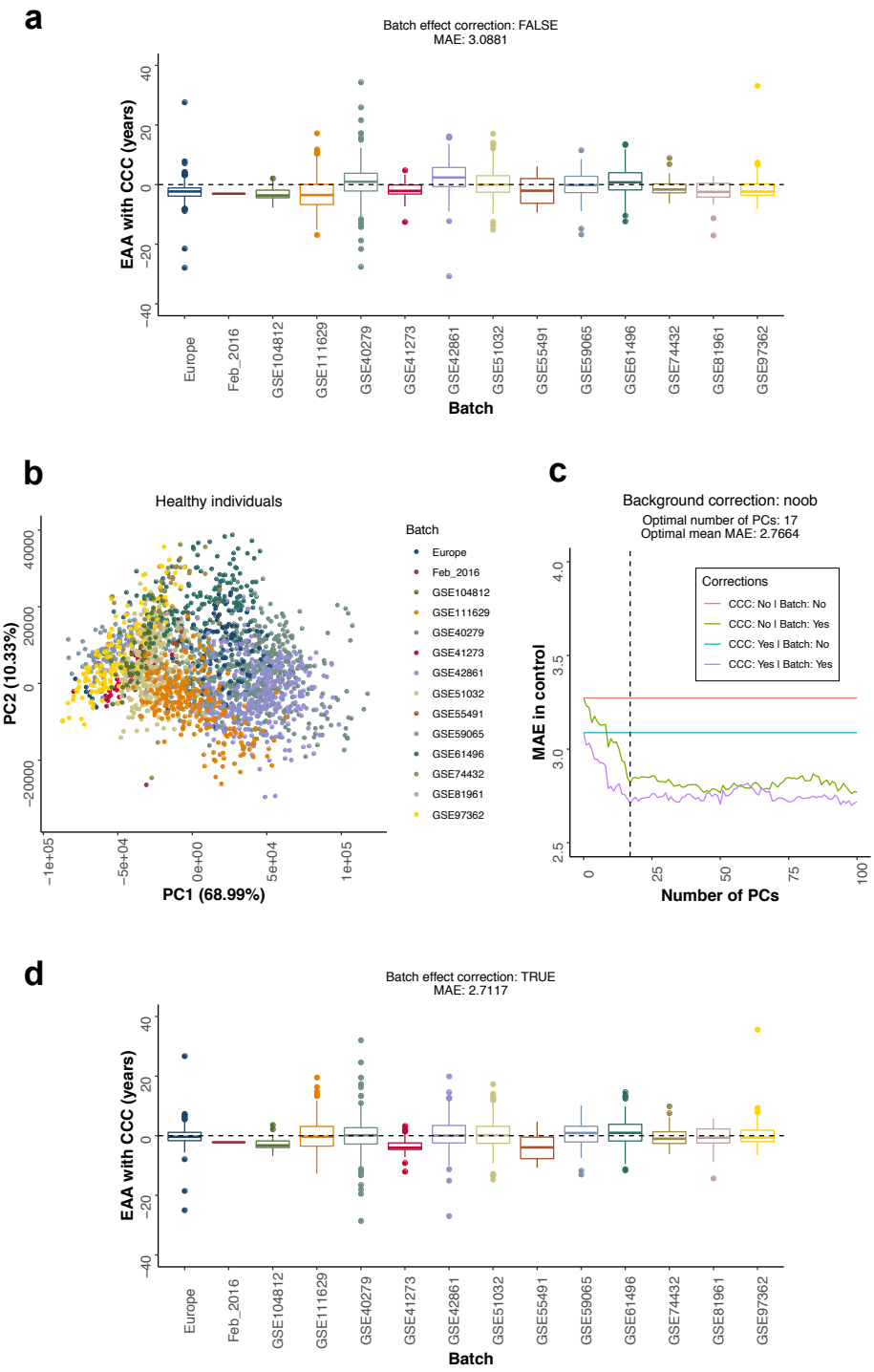


Fig. 2.13 Correcting for batch effects in the context of the epigenetic clock. **a.** Distribution of the epigenetic age acceleration (EAA) for the different batches of healthy individual samples, using the control model with cell composition correction (CCC) and before applying batch effect correction. The dashed black line represents $EAA = 0$, where the distributions should be centred around. **b.** Scatterplot showing the values of the first two principal components (PCs) for the healthy individual samples after performing PCA on the control probes of the 450K arrays. Each point corresponds to a different sample and the colours represent the different batches. The different batches cluster together in the PCA space, showing that the control probes indeed capture technical variation. Please note that all the PCA calculations were done using samples from both healthy individuals (full lifespan, $N = 2218$) and cases from developmental disorders ($N = 666$, see Chapter 3). **c.** Plot showing how the median absolute error (MAE) of the prediction in the healthy individual samples, that should tend to zero, is reduced when the PCs capturing the technical variation are included as part of the modelling strategy (see equations 2.16 and 2.17). The dashed line represents the optimal number of PCs (17) that was finally used. The optimal mean MAE is calculated as the average MAE between the green and purple lines. **d.** As in a., but after applying batch effect correction (i.e. equivalent to equation 2.16).

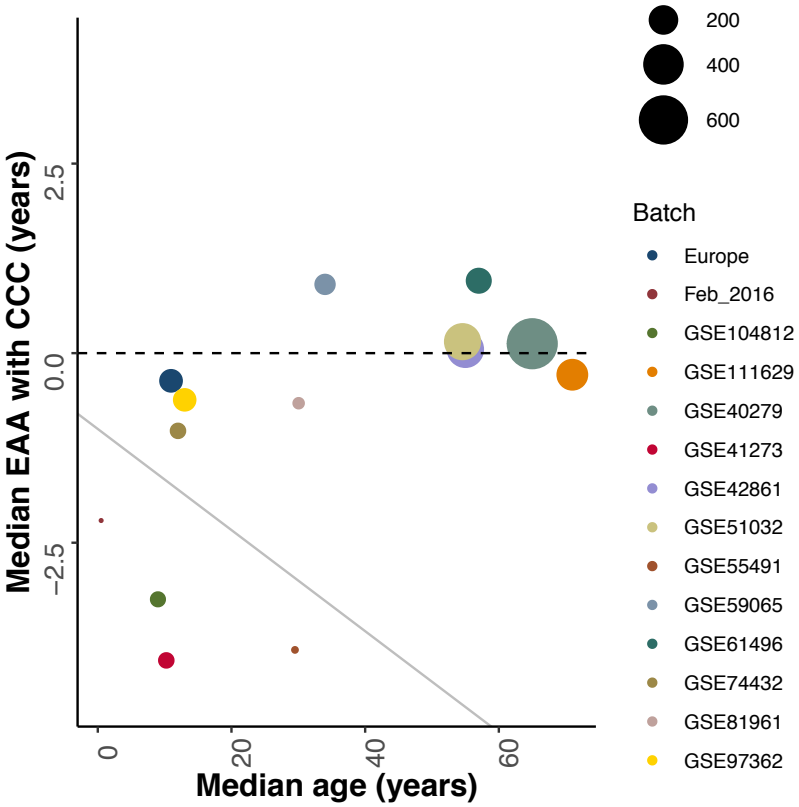


Fig. 2.14 After applying batch effect correction in the samples from the healthy individuals, deviations from a median epigenetic age acceleration (EAA) of zero (dotted black line) in some of the batches can be explained by other causes. The grey line separates in the lower left corner those weird batches (Feb_2016, GSE104812, GSE41273, GSE55491), which have a small sample size and/or a low median age.

2.3 Behaviour of other epigenetic clocks during ageing

2.3.1 Hannum's epigenetic clock

Besides Horvath's epigenetic clock, other models have been proposed in the literature to measure the ageing process using DNA methylation. Among them, Hannum's epigenetic clock has also been shown to accurately predict epigenetic age in several cohorts [32, 33, 67, 78–80]. Hannum's model was originally trained in whole blood and it makes use of a linear combination of β-values from 71 probes in the 450K array.

I calculated the epigenetic ages according to Hannum's model (*HannumAge*), although I only used 68 out of the 71 probes (the other 3 were filtered out during my pre-processing). Hannum's epigenetic clock predicted quite accurately in the dataset of healthy individuals, although with a slight overestimation of the epigenetic ages (Fig 2.15a), which has also been previously observed [78]. Furthermore, it is possible to observe the non-linear behaviour of Hannum's clock for young ages ($\leq$ 20 years), for which the authors did not correct in their original publication [56]. Horvath's and Hannum's epigenetic clocks correlate between them (Fig. 2.15b). The magnitude of this correlation (*HannumAge* vs *DNAmAge*: PCC = 0.9778) was slightly stronger than the correlation between *HannumAge* and chronological age (PCC = 0.9756), which could highlight the fact that both models indeed measure epigenetic age.

Next, I estimated the epigenetic age acceleration (EAA) according to Hannum's epigenetic clock, using similar models to the ones previously described (although in this case the dependent variable was *HannumAge*, see equations 2.16 and 2.17). The median absolute errors for Hannum's model ($MAE_{\text{with CCC}} = 2.8422$ years, $MAE_{\text{without CCC}} = 2.9484$ years) were slightly higher than the ones obtained for Horvath's clock ($MAE_{\text{with CCC}} = 2.7117$ years, $MAE_{\text{without CCC}} = 2.8211$ years), which could also be influenced by the fact that 3 of the model probes were not available. The EAAs estimated by Hannum's and Horvath's clocks showed a moderate correlation (Fig. 2.15c,d), consistent with previous estimates [80]. Including cell composition correction improved the correlation between the EAAs from both clocks, highlighting the fact that Hannum's clock seems to be confounded with the changes in blood cell composition with age [78, 80].

Overall, Hannum's epigenetic clock performed well in my dataset. However, given that it produces slightly worse predictions than Horvath's and could be partially tracking blood immunosenescence instead of multi-tissue ageing effects, I used the latter as my main proxy to measure the ageing process in this thesis. Finally, it is also worth mentioning that the data

that was used to train Hannum's model (GSE40279) is also part of the dataset of healthy individuals that I assembled and, therefore, this analysis does not constitute a completely independent assessment of the behaviour of Hannum's epigenetic clock.

2.3.2 Epigenetic mitotic clock: *epiTOC*

In 2016, Yang and colleagues conceived a novel type of epigenetic clock called *epiTOC* (epigenetic Timer Of Cancer), which measures the rate of (stem) cell division in both normal and cancerous tissues [81]. This epigenetic mitotic clock tracks the gain in methylation levels that happens in 385 CpG sites, which localise in the promoter of genes that are targeted by Polycomb Repressing Complex 2 (PRC2). Importantly, these CpG sites are unmethylated across fetal tissues and therefore this provides a ground state to measure these changes during human lifespan.

I calculated the mitotic age (*pcgtAge*) of the healthy individuals in my dataset, although I only used 378 out of the 385 probes (the other 7 were filtered out during my pre-processing). The mitotic age of the individuals correlated with both chronological age (PCC $= 0.5131$, Fig. 2.16a) and *DNAmAge* (PCC $= 0.5602$, Fig. 2.16b), which is expected given the cumulative number of divisions of the hematopoietic stem cells [82]. Furthermore, I estimated the epigenetic age acceleration (EAA) according to the epigenetic mitotic clock, using similar models to the ones previously described (although in this case the dependent variable was *pcgtAge*, see equations 2.16 and 2.17). Interestingly, the EAAs for *pcgtAge* and *DNAmAge* showed a small but highly statistically significant correlation (Fig. 2.16c,d), which was stronger in the case of the model with cell composition correction. This, together with the fact that *DNAmAge* has a stronger correlation with *pcgtAge* than chronological age, could suggest that the Horvath epigenetic clock captures methylation changes linked to cell division.

This was quite surprising given that Horvath's epigenetic clock predicts across tissues with different turnover rates [81]. Nevertheless, it has been recently demonstrated that *DNAmAge* increases linearly with cell passage *in vitro* if TERT (the catalytic subunit of telomerase) is expressed, suggesting that *DNAmAge* does seem to track cell division to a certain extent [83]. Furthermore, I also did some preliminary work where I calculated the *DNAmAge* of different healthy tissues (that came from cancer patients). I observed that tissues with a high turnover (such as breast) [51, 84] had a higher *DNAmAge* when compared with tissues with a low turnover (data not shown). Therefore, it would be interesting to further our understanding of the contribution of cell division to Horvath's epigenetic clock and its

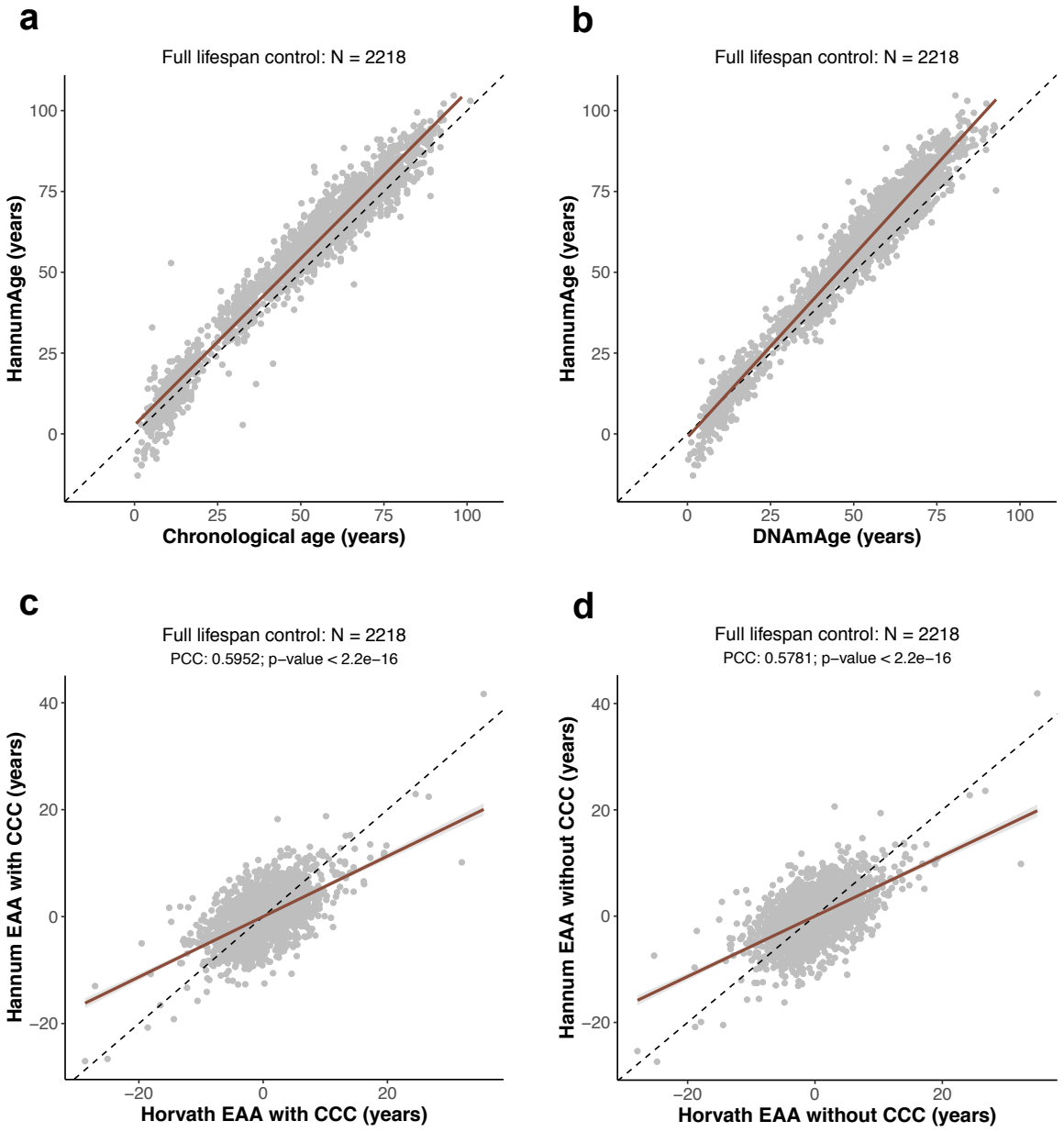


Fig. 2.15 Behaviour of Hannum's epigenetic clock in the healthy individuals. **a.** Scatterplot showing the relationship between the epigenetic age predicted with Hannum's model (*HannumAge*) [56] and chronological age of the samples for the healthy individuals. Each sample is represented by one point. The black dashed line represents the diagonal to aid visualisation. The solid brown line represents the linear model HannumAge $\sim$ Age. **b.** Relationship between the Hannum and Horvath epigenetic ages estimated for the same sample. The solid brown line represents the linear model HannumAge $\sim$ DNAmAge. **c.** Relationship between the epigenetic age acceleration (EAA) calculated with the Hannum and the Horvath's epigenetic clocks. In this case the models include cell composition correction (CCC). The solid brown line represents the linear model $\text{Hannum\_EAA}_{\text{with CCC}} \sim \text{Horvath\_EAA}_{\text{with CCC}}$. **d.** As in c., but in this case the models do not include CCC.

relation to the hypermethylation in PRC2-bound regions as measured by the epigenetic mitotic clock.

2.4 Additional methods

A short introduction to the linear regression framework

Linear models are a broad class of statistical analyses that are at the core of many bioinformatic methods, including differential RNA expression analyses [45] or genome-wide association studies (GWAS) [85]. An instance of such models is linear regression [86], a statistical approach that allows modelling of the relationship between:

- A dependent variable $\mathbb{Y}$, with observations $y_i \in R$ and $i \in \{1,...,n\}$, where n is the total number of observations (i.e. samples).
- One or more independent variables $\mathbb{X}_j$, with observations $x_{ij} \in R$ and $j \in \{1,...,k\}$, where k is the total number of independent variables (a.k.a covariates). These variables can indicate, for example, whether a specific condition or phenotype is present in a given sample, quantify the effects of a continuous variable (such as chronological age) or adjust for the effects of batch effects; which gives this statistical framework a great analysis flexibility [45].

We assume that:

$$y_i = \sum_{j=1}^{k} x_{ij}\beta_j + \varepsilon_i \tag{2.19}$$

where β_k are unknown parameters that need to be estimated from the data and ε_i is the random error. In matrix form:

$$Y = X\beta + \varepsilon \tag{2.20}$$

where $Y \in R^n$ is the vector $\{y_1,...,y_n\}$, $X \in R^{n\times k}$ is the $n \times k$ matrix of x_{ij}'s, $\beta \in R^k$ is the vector $\{\beta_1,...,\beta_k\}$ and $\varepsilon \in R^n$ is the vector $\{\varepsilon_1,...,\varepsilon_n\}$.

Assuming that $\mathbb{E}(\varepsilon) = 0$, $\mathrm{Var}(\varepsilon) = \sigma^2 > 0$ and $\mathrm{Cov}(\varepsilon) = \sigma^2 I_n$ (where I_n is the $n \times n$ identity matrix) and applying the Gauss-Markov theorem [86], it can be demonstrated that:

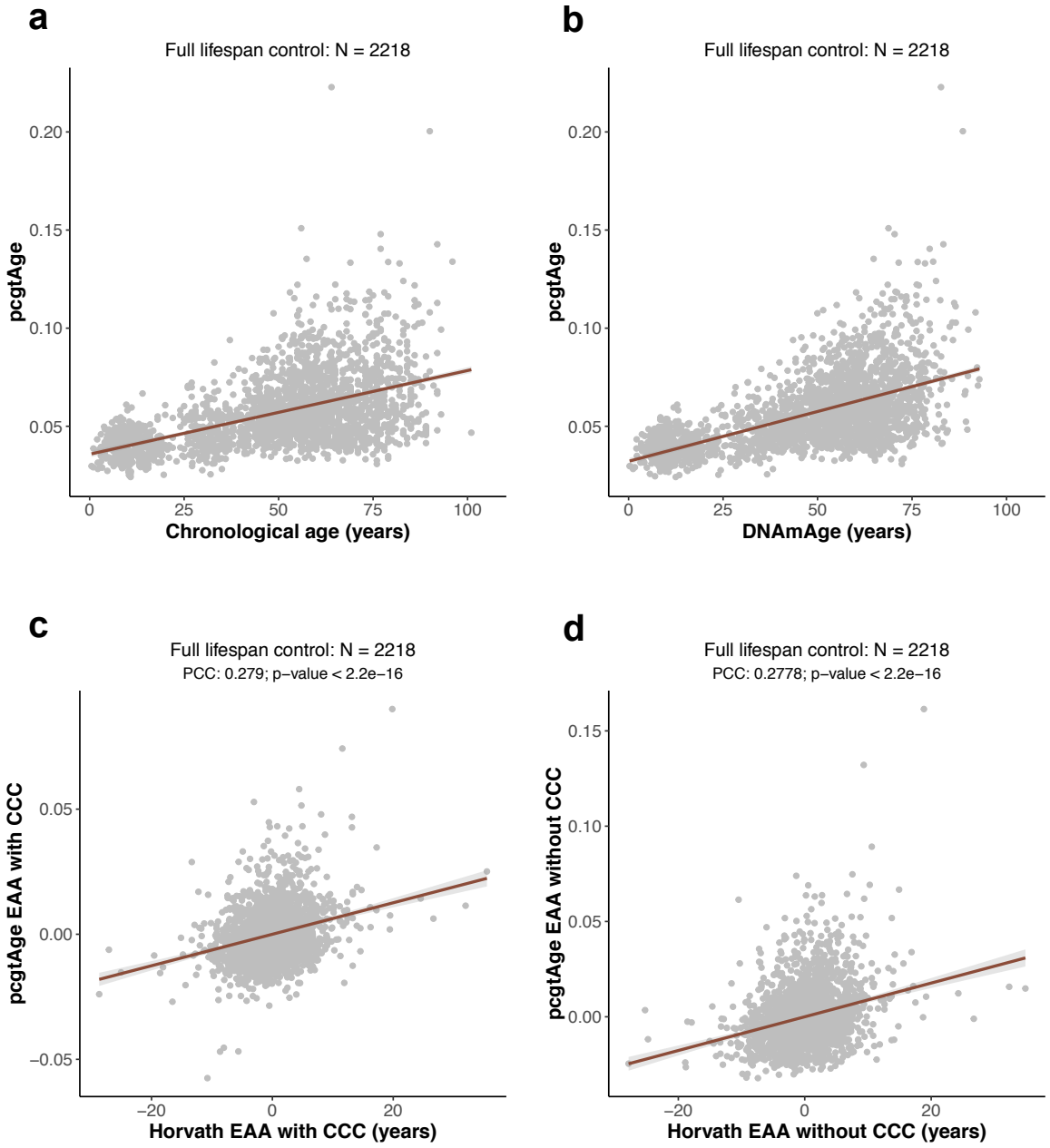


Fig. 2.16 Behaviour of the epigenetic mitotic clock (*epiTOC*) in the healthy individuals. **a.** Scatterplot showing the relationship between mitotic age (*pcgtAge*) [81] and chronological age of the samples for the healthy individuals. Each sample is represented by one point. The solid brown line represents the linear model pcgtAge $\sim$ Age. **b.** Relationship between *pcgtAge* and *DNAmAge* estimated for the same sample. The solid brown line represents the linear model pcgtAge $\sim$ DNAmAge. **c.** Relationship between the epigenetic age acceleration (EAA) calculated with the mitotic and the Horvath's epigenetic clocks. In this case the models include cell composition correction (CCC). The solid brown line represents the linear model pcgtAge_EAA$_{\text{with CCC}}$ $\sim$ Horvath_EAA$_{\text{with CCC}}$. **d.** As in c., but in this case the models do not include CCC.

$$\hat{\beta} = (X'X)^{-1}X'Y \tag{2.21}$$

where X' is the transpose of X and $\hat{\beta}$ is the least-squares estimator of β, since it minimises:

$$\sum_{i=1}^{n}(y_i - \sum_{j=1}^{k} x_{ij}\hat{\beta}_j)^2 \tag{2.22}$$

It is possible to test whether there is a statistically-significant linear association between the dependent variable ($\mathbb{Y}$) and one of the independent variables ($\mathbb{X}_j$) i.e. to test:

$$H_0 : \beta_j = 0 \quad \text{against} \quad H_A : \beta_j \neq 0 \tag{2.23}$$

where H_0 is the null hypothesis and H_A is the alternative hypothesis. A t-statistic (T) can be derived after performing the fitting of the linear regression model [87]:

$$T = \frac{\hat{\beta}_j}{se(\hat{\beta}_j)} \tag{2.24}$$

where $se(\hat{\beta}_j)$ is the standard error of $\hat{\beta}_j$. When H_0 is true, then the statistic T follows a Student's t distribution with $n-k$ degrees of freedom i.e. $T \sim t_{n-k}$. This allows to estimate the p-value for the linear association of $\mathbb{Y}$ with a given $\mathbb{X}_j$.

Finally, it is worth mentioning the nomenclature that I used for the linear regression models along this thesis. For example, the following model fits a linear association between the dependent variable (e.g. β-value at a specific CpG probe in the array) with intercept and 3 covariates (e.g. age, sex and disease status):

$$\begin{bmatrix} y_1 \\ y_2 \\ \dots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & x_{12} & x_{13} \\ 1 & x_{21} & x_{22} & x_{23} \\ \dots & \dots & \dots & \dots \\ 1 & x_{n1} & x_{n2} & x_{n3} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \dots \\ \varepsilon_n \end{bmatrix} \tag{2.25}$$

where y_i is the β-value at a certain CpG probe for the ith sample, x_{i1} is the age for the ith sample, x_{i2} is the sex (e.g. 0 for male and 1 for female) for the ith sample, x_{i3} is the

disease status (e.g. 0 for a healthy individual and 1 for an individual with a disease) for the ith sample, β_0 is the intercept coefficient, β_j are the covariate coefficients ($j = 1$ for age, $j = 2$ for sex, $j = 3$ for disease status) and ε_i is the error for the ith sample.

I would use the following nomenclature for the previous model (following 'R-style' nomenclature):

$$\text{Beta} \sim \text{Age} + \text{Sex} + \text{Disease\_status} \tag{2.26}$$

Chapter 3

Biological aspects of the epigenetic clock

3.1 Background

Synchrony and asynchrony between an epigenetic clock and developmental timing https://www.nature.com/ar
019-39919-3

3.2 Discussion

- Oscillatory amplitude is greatest at enhancers. Maybe when DNMT3A is absent they are more sensitive to changes in DNA methylation, in this case towards hypomethylation. https://www.cell.com/cell-systems/fulltext/S2405-4712(18)30279-5

Add paragraph with discussion on epigenetic mitotic clock from TAC 2.

3.3 Additional methods

[For aDMPs we use linear regression, for Sotos DMPs, we use t-statistics]

In the case of a continuous phenotype (e.g. age) the association is carried out under a linear regression model framework, while for a binary phenotype (e.g. cancer/normal status) we use t-statistics [https://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-13-59]

Experimental procedures for DNA methylation data generation

Estimating epigenetic age acceleration in a biological system

Include all the acceleration models (Horvath, pcgtAge, entropy)

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 [88]. 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) [89–91], 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 [92, 93]. 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 [94].

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 [95–99]. 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) [100], a recombinant binding protein (e.g. methyl-CpG-binding domains or MBD) [101], covalent biotin tagging [102], capture probes/baits for the sites of interest [103–105], array-based approaches (e.g. 27K, 450K and EPIC arrays in human) [5–7, 106] or PCR-based approaches [107–112]. These methods have many limitations, including enrichment biases, complex protocols and difficulties in quantification [95, 97].

Current evidence shows that depletion-based methods do not have enrichment biases, tend to be simpler and are more readily quantifiable [95, 98]. 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 [113–117]. 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) [118, 119], although enzymes such as BglII [120], XmaI [121], Taq$^{\alpha}$I [122, 123], MspJI [124] , ApeKI [125], HpyCH4IV or HpaII [126] have also been used. RRBS has proven extremely useful for cost-effective, global studies of DNA methylation [68, 118, 122, 127], 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 [128].

In the context of epigenetic clocks, most studies have used methylation arrays in humans [51, 56, 129] and MspI-based RRBS in mice, dogs and wolves [68, 130–133]. 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 [118]. Nevertheless, it is known that many age-related changes in the methylome occur in other genomic regions (such as enhancers) [47, 48, 134, 135], and current technologies could be biasing our discoveries. Furthermore, epigenetic clocks could be used in the near future to perform high-throughput screenings of anti-ageing drugs or employed as ageing biomarkers in clinical trials [136]. 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 [51, 137–142]. 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 [121, 123]. 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).

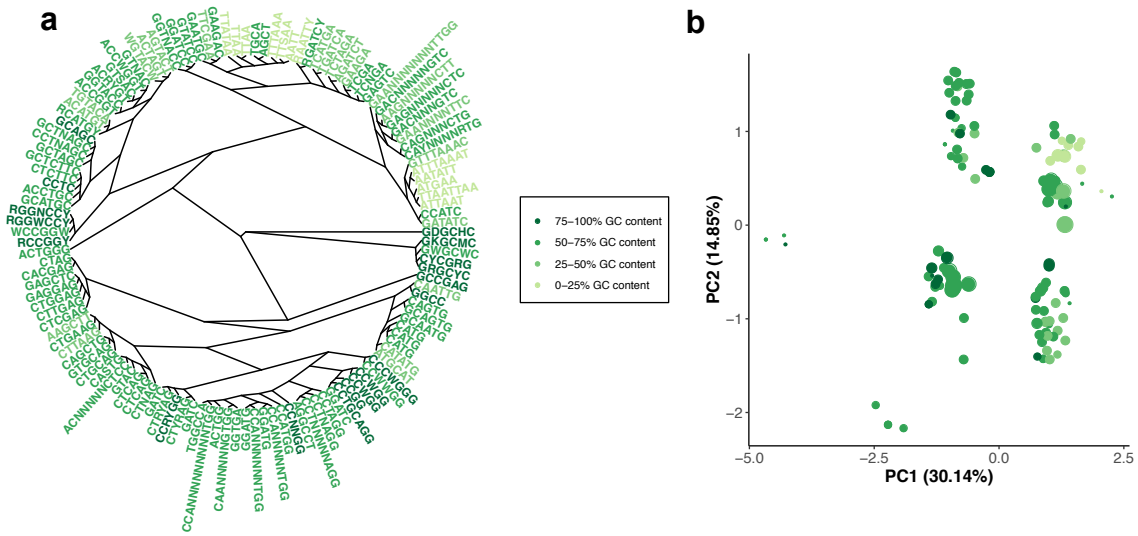


Fig. 4.1 The landscape of restriction enzyme motifs. **a.** Phylogenetic analysis of the motifs that are recognised by the different commercially-available restriction enzymes which are insensitive to CpG methylation. Each sequence represents a different isoschizomer family considered in this study. 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.

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 [118] (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 [113, 115–117, 121, 122, 125, 126]. 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:

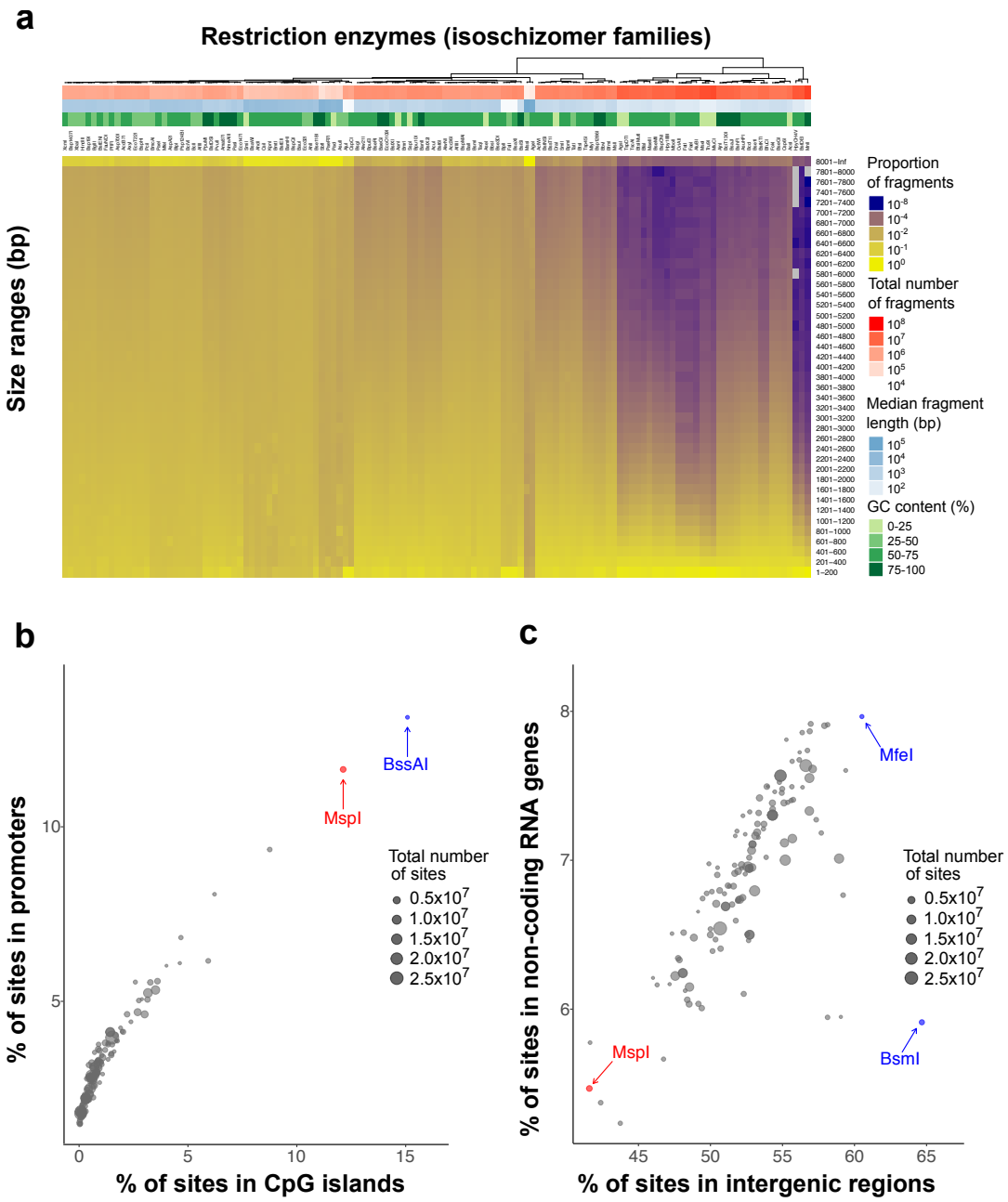


Fig. 4.2 Restriction enzyme digestion as a tool for genomic enrichment. **a.** Heatmap showing the fragment length distributions generated by different restriction enzymes in the human genome (hg38). Each column represents the distribution for an isoschizomer family of restriction enzymes that contains 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.

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

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

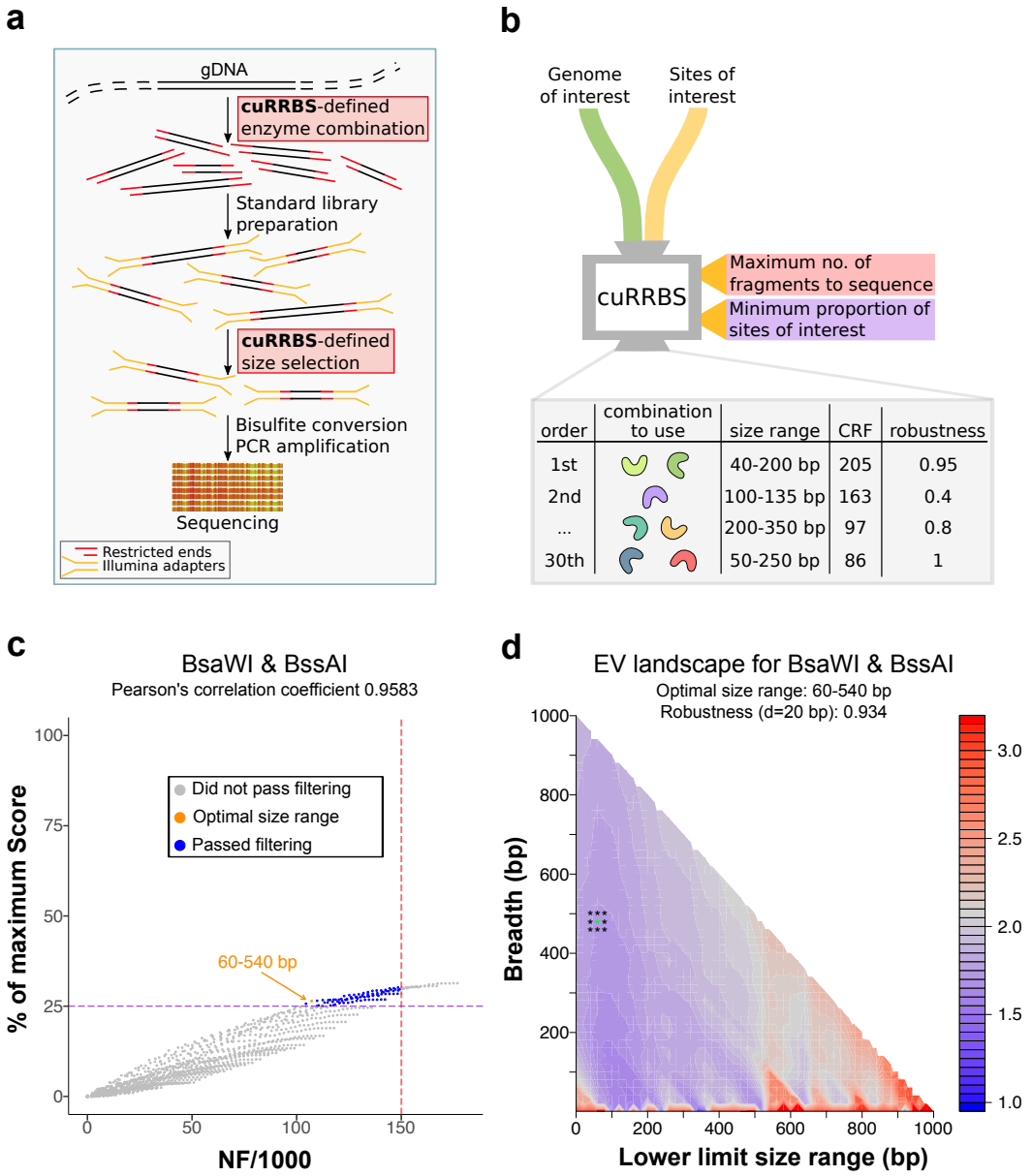


Fig. 4.3 cuRRBS overview. **a.** Outline of an RRBS protocol. Highlighted are the two steps that would be modified according to the output produced by cuRRBS (i.e. the restriction enzymes used for the genomic digestion and the size selection). Legend is displayed on the bottom left. **b.** Schematic 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 [137]. *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 [137]. *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 [143], which currently lack an MspI-based RRBS protocol, owing to the enzyme's CHG methylation sensitivity [144].

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 [138], Horvath's epigenetic clock [51], transcription factor binding sites that are affected by DNA methylation [140, 142], imprinted loci [137], CpGs found in the exon-intron boundaries [141] and CHG sites that are differentially methylated between different arabidopsis accessions [139] (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 [115]. 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) [121, 123]. By evaluating the predictive power of cuRRBS in these two

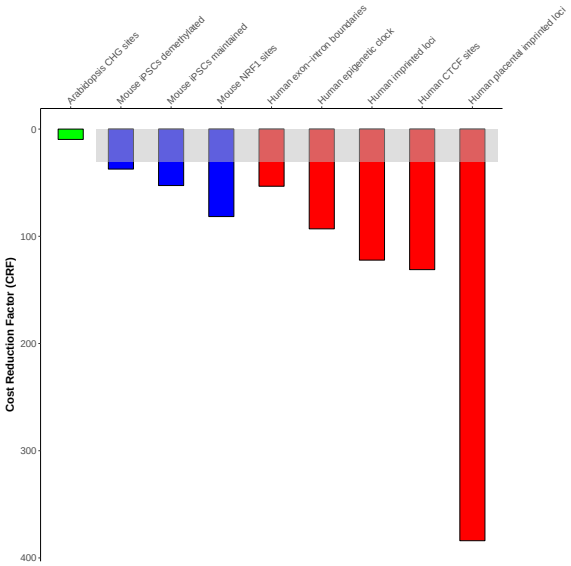


Fig. 4.4 Running cuRRBS in different biological systems. Barplot showing the values for the *Cost Reduction Factor* (*CRF*) in the different biological systems that were tested (see Fig. S3.5) [51, 137–142]. The colours in the bars represent the different species interrogated (green: *Arabidopsis thaliana*, 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 [121]. 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 [121] (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) [121]. 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 [123]. 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 [123]. 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

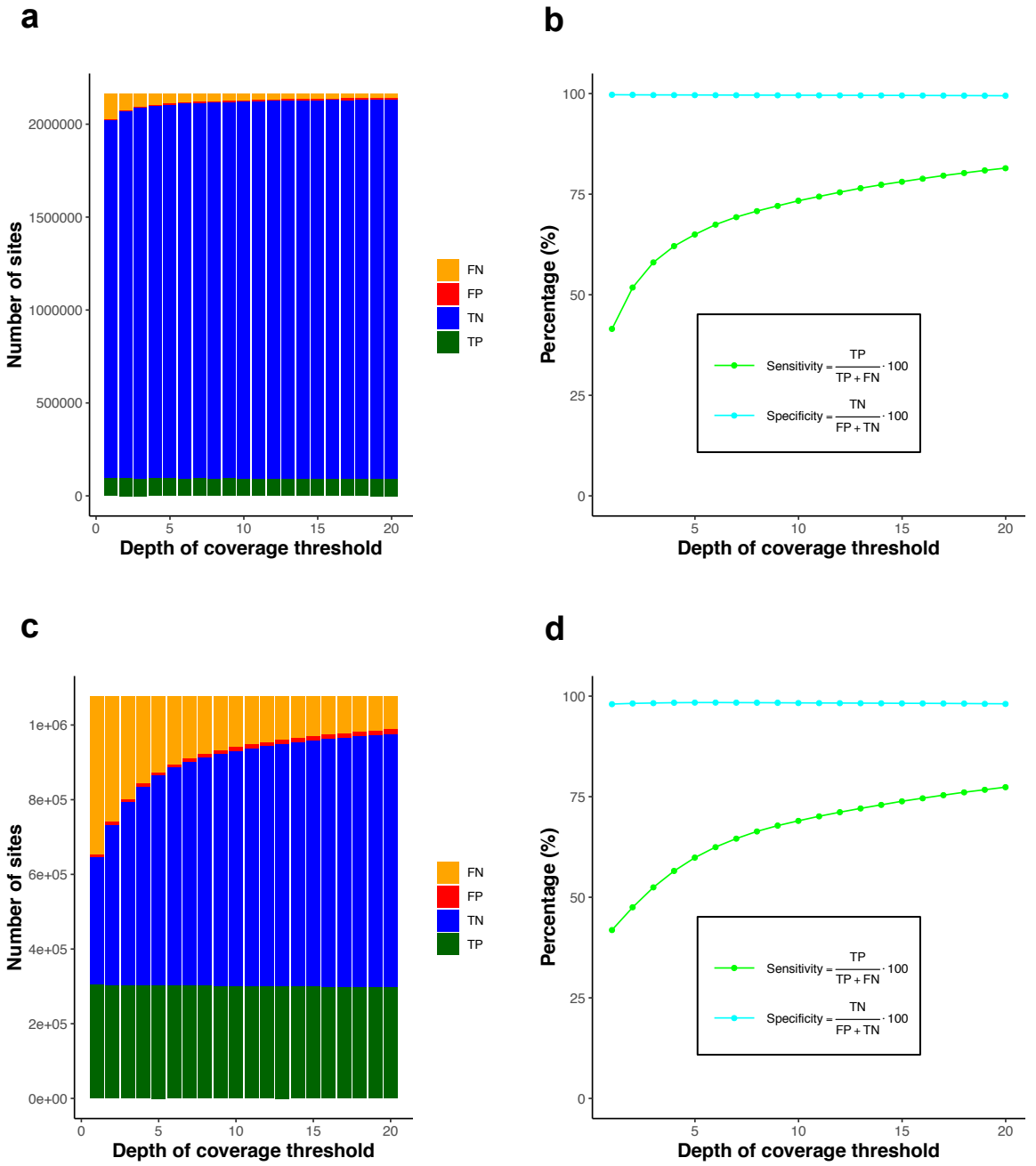


Fig. 4.5 Experimental validation of cuRRBS. **a.** Barplots showing the number of true positives (TP, in green), true negatives (TN, in blue), false positives (FP, in red) and false negatives (FN, in orange) when comparing cuRRBS theoretical prediction with the actual XmaI-RRBS experimental data [121]. 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 [121]. 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 [123]. 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 [123].

makes the study of DNA methylation more affordable across all species for which genomic sequences are available. Furthermore, it can open the door to the design of future studies in a clinical context [122], 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 [115]. Moreover, there are several methods that are able to impute DNA methylation levels in sites that are not covered experimentally [145, 146]. 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 [147, 148] or to optimise chromosome conformation capture techniques [149, 150]. 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 [151, 152]. 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) [153], with the exception of CpG islands (CGIs), which were extracted from the UCSC Genome Browser [154]. 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 [145]. 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.* [155]. Genomic annotation for the CGIs in the mouse genome (mm10) was also obtained from the UCSC Genome Browser [154]. All annotations were handled using the *pybedtools* library [156, 157].

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* [158]. 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 [159]. 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

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.

enzyme combination and size range would be the simplest and most appropriate for the given 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 [159] 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 [51] and were lifted over to hg38 [160] before running cuRRBS.

- Canonical and placental imprints in human. These loci were obtained from Hanna *et al.* [137]. The sites were lifted over to hg38 [160] 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 [140]. We scanned the peaks for high-scoring motifs according to the CTCF JASPAR model [161]. 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 [140].

- 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 [138], 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 [142], where peaks were called using Peakzilla [162]. 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 [161]. 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 [161].

- CHG sites in *Arabidopsis thaliana*. Non-CpG DMRs arising from the epigenomic diversity between *Arabidopsis thaliana* accessions were obtained from Kawakatsu *et al.* [139]. 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 [51]. All the site annotation files can be found in my GitHub repository [159]

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 [121] and MspI&Taq$^{\alpha}$I -RRBS data generated on the Illumina HiSeq platform [123] 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) [163]. 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 [121, 123] (see section 4.5). We ran cuRRBS fixing the theoretical size ranges tested to the ones reported in the publications [121, 123] and we used as our sites of interest the CpGs that overlapped with CpG islands (CGI-CpGs) in the human [121] and the mouse genomes [123] respectively. From the cuRRBS output files we recovered the IDs of the sites that should be theoretically sequenced. Moreover, using the experimental RRBS data [121, 123], 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 [159].

Appendix

Supplementary figures

S.1 Statistical aspects of the epigenetic clock

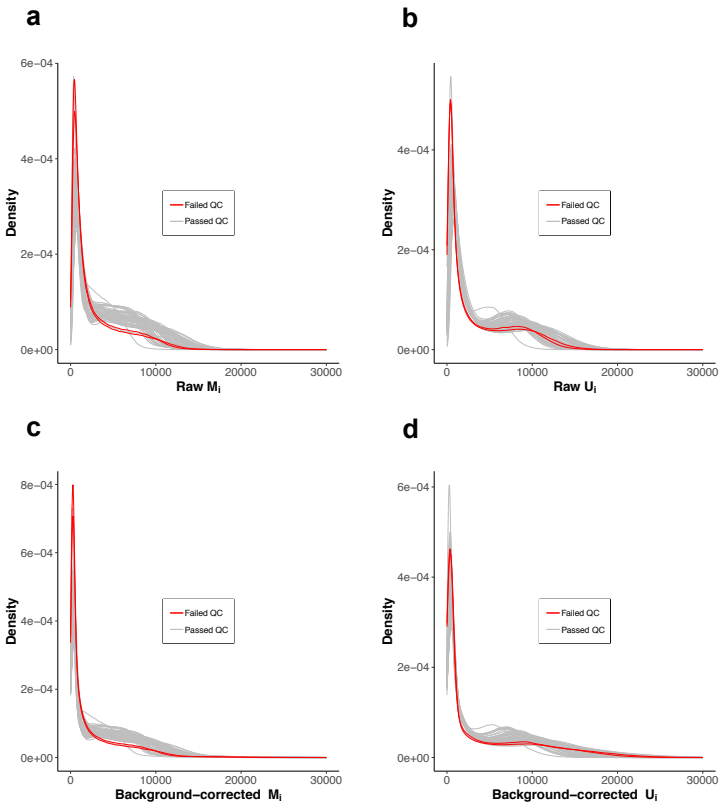


Fig. S1.1 Effects of *noob* background correction on the array flurescence intensities. Distributions of the array fluorescence intensities for the **a.** methylated signals (M_i) before background correction; **b.** unmethylated signals (U_i) before background correction; **c.** methylated signals (M_i) 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.

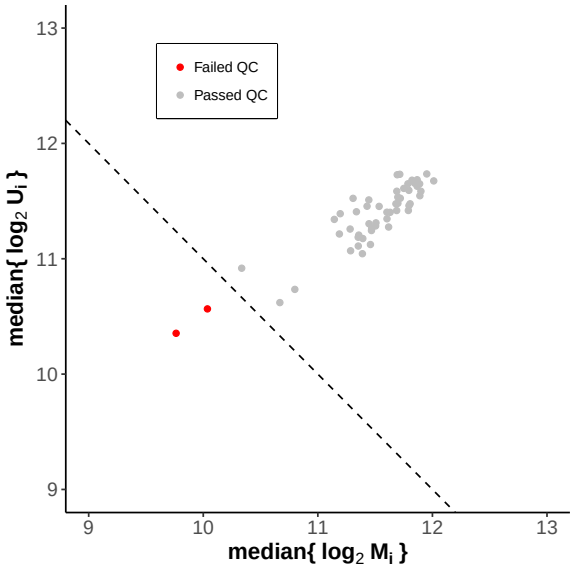


Fig. S1.2 Quality control (QC) strategy to identify outlier samples, according to their global intensity values, in the GSE41273 batch. Those samples with low median intensity values (see criteria in section 2.1.2) were discarded from downstream analyses (2/53, in red). Each sample is represented by one point. The dashed line represents the intensity threshold. M_i and U_i represent the background-corrected methylated and unmethylated intensity measurements for the different 450K array probes in a given sample.

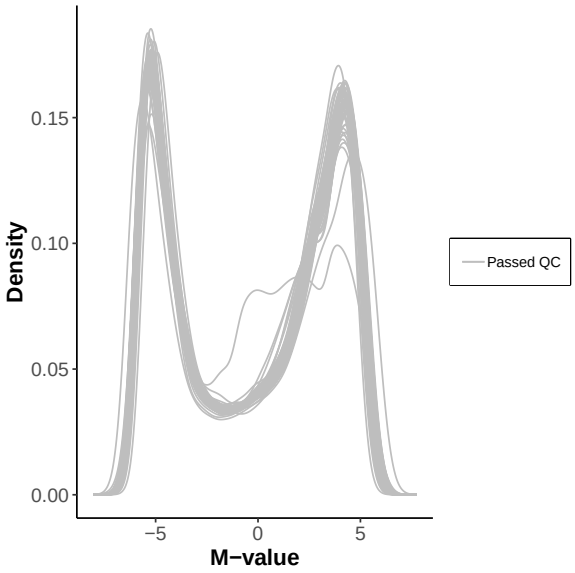


Fig. S1.3 M-value distributions in the samples of the GSE41273 batch, after all the pre-processing steps have been carried out (background correction, quality control, probe filtering and BMIQ normalisation). M-values were calculated applying the logistic transformation to the β-values, as described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

Strategy name	Reference	Gold-standard preprocessing	Reference preprocessing	Probes in reference	Algorithm	Mean RMSE	Mean MAE	Mean R^2
minfi	minfi	SQN*	SQN*	600	Houseman CP/QP	2.3246	2.0137	0.9473
dhs_dif1_houseman	DHS-DMCs	Noob+BMIQ	Default	333	Houseman CP/QP	4.8039	3.843	0.7783
dhs_NB_houseman	DHS-DMCs	Noob+BMIQ	Noob+BMIQ	333	Houseman CP/QP	4.9398	4.1559	0.8062
dhs_dif2_houseman	DHS-DMCs	Noob+Filtering+ BMIQ	Default	316	Houseman CP/QP	6.1731	5.2469	0.7779
dhs_NFB_houseman	DHS-DMCs	Noob+Filtering+ BMIQ	Noob+Filtering+ BMIQ	316	Houseman CP/QP	6.1194	5.3185	0.7816
dhs_dif1_cibersort	DHS-DMCs	Noob+BMIQ	Default	333	CIBERSORT	2.3914	1.9502	0.8702
dhs_NB_cibersort	DHS-DMCs	Noob+BMIQ	Noob+BMIQ	333	CIBERSORT	2.8578	2.3833	0.8453
dhs_dif2_cibersort	DHS-DMCs	Noob+Filtering+ BMIQ	Default	316	CIBERSORT	2.9751	2.4714	0.8552
dhs_NFB_cibersort	DHS-DMCs	Noob+Filtering+ BMIQ	Noob+Filtering+ BMIQ	316	CIBERSORT	3.0684	2.5403	0.8571
dhs_dif1_rpc	DHS-DMCs	Noob+BMIQ	Default	333	RPC	2.0421	1.7032	0.8873
dhs_NB_rpc	DHS-DMCs	Noob+BMIQ	Noob+BMIQ	333	RPC	2.5289	2.1689	0.8705
dhs_dif2_rpc	DHS-DMCs	Noob+Filtering+ BMIQ	Default	316	RPC	2.9653	2.3887	0.8722
dhs_NFB_rpc	DHS-DMCs	Noob+Filtering+ BMIQ	Noob+Filtering+ BMIQ	316	RPC	3.0755	2.5266	0.8611
idol_NB_houseman	IDOL	Noob+BMIQ	Noob+BMIQ	300	Houseman CP/QP	2.0347	1.6778	0.9632
idol_NFB_houseman	IDOL	Noob+Filtering+ BMIQ	Noob+Filtering+ BMIQ	281	Houseman CP/QP	1.927	1.5498	0.9672
idol_NB_cibersort	IDOL	Noob+BMIQ	Noob+BMIQ	300	CIBERSORT	2.1997	1.7958	0.9626
idol_NFB_cibersort	IDOL	Noob+Filtering+ BMIQ	Noob+Filtering+ BMIQ	281	CIBERSORT	1.9818	1.6216	0.9704
idol_NB_rpc	IDOL	Noob+BMIQ	Noob+BMIQ	300	RPC	2.26	1.8812	0.9679
idol_NFB_rpc	IDOL	Noob+Filtering+ BMIQ	Noob+Filtering+ BMIQ	281	RPC	2.0122	1.6288	0.9692

Fig. S1.4 Table showing the different cell-type deconvolution strategies that were benchmarked. BMIQ: beta-mixture quantile normalisation. CP/QP: constrained projection/quadratic programming. MAE: mean absolute error. Noob: *noob* background correction. R^2: coefficient of determination. RMSE: root mean squared error. RPC: robust partial correlations. SQN: stratified quantile normalisation. 'Default' refers to the pre-processing strategy employed in the original DHS-DMCs publication, as implemented in the *EpiDISH* R package (*centDHSbloodDMC.m*) [35, 39]. See section 2.1.3 in the main text for more details on what the different references refer to.

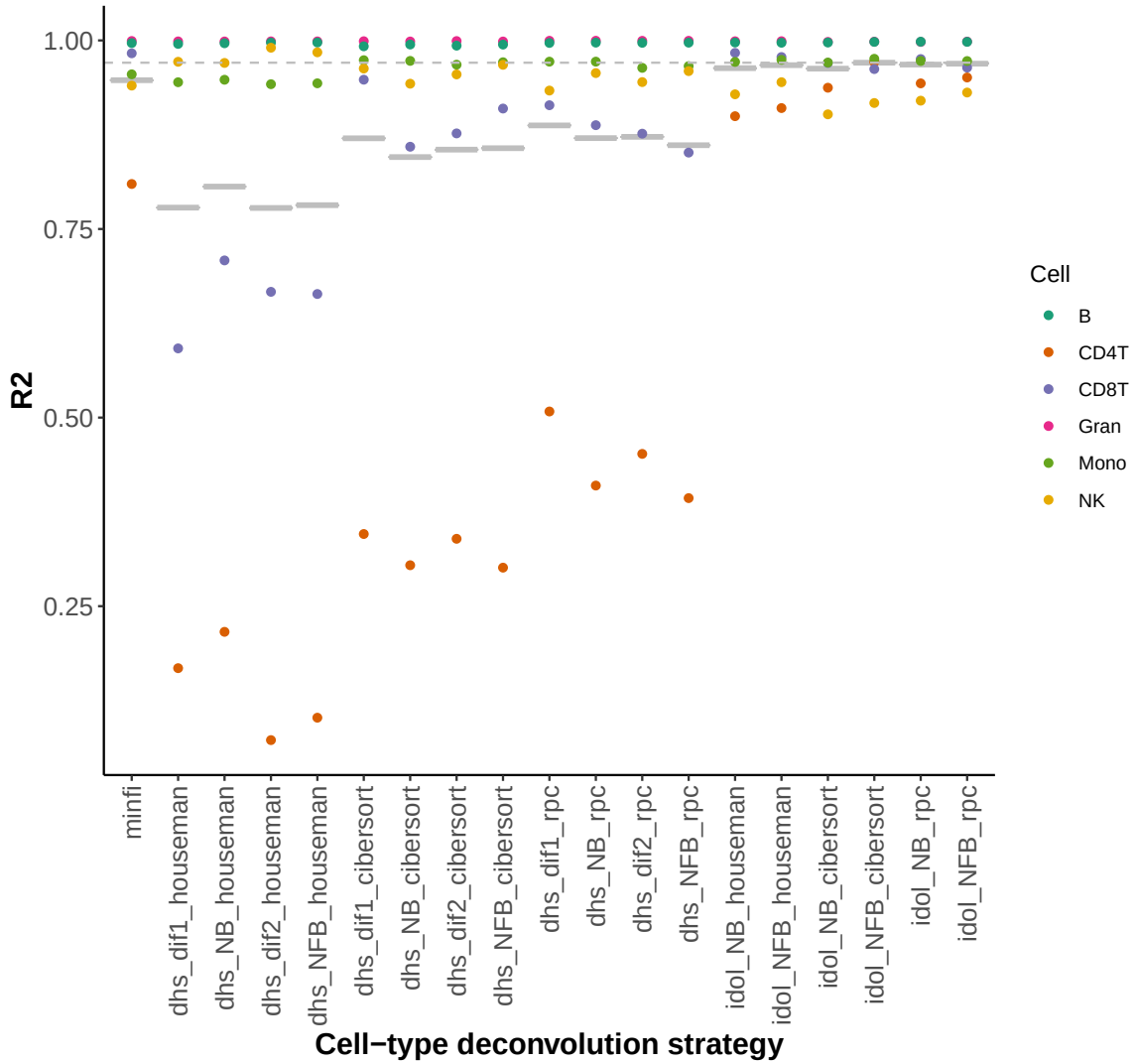


Fig. S1.5 Benchmarking of the cell-type deconvolution strategies in blood. The x-axis shows the different strategies that were tested (for a detailed description see Fig. S1.4). The y-axis shows the results for the coefficient of determination (R^2) when comparing the predictions with the real proportions of cells in a gold-standard dataset (GSE77797) [38]. The grey horizontal solid lines represent the mean for the R^2 across cell types and the grey dashed line the maximum of these values.

ProbeID	Chromosome	Coordinate	Intercept	Slope	T statistic	p-value	Methylation change	In Horvath model	Gene(s)
cg16867657	chr6	11044877	0.5458189	0.0053562	96.7079	0	Hypermethylated	No	ELOVL2
cg06639320	chr2	106015739	-0.18099	0.0040751	68.4826	0	Hypermethylated	No	FHL2
cg21572722	chr6	11044894	0.4485118	0.0029979	67.7891	0	Hypermethylated	No	ELOVL2
cg22454769	chr2	106015767	-0.37256	0.0054721	65.4459	0	Hypermethylated	No	FHL2
cg07547549	chr20	44658225	-0.109895	0.0039332	60.4444	0	Hypermethylated	No	SLC12A5
cg24724428	chr6	11044888	0.1715795	0.003787	60.3559	0	Hypermethylated	No	ELOVL2
cg17110586	chr19	36454623	-0.076933	0.0027991	59.6101	0	Hypermethylated	No	
cg19283806	chr18	66389420	1.1244081	-0.0052494	-55.5368	0	Hypomethylated	No	CCDC102B
cg10501210	chr1	207997020	-0.767615	-0.0071941	-54.848	0	Hypomethylated	No	
cg24079702	chr2	106015771	-0.239806	0.0037027	54.5055	0	Hypermethylated	No	FHL2
cg22796704	chr10	49673534	0.5923358	-0.0038938	-54.2818	0	Hypomethylated	No	ARHGAP22
cg04875128	chr15	31775895	-0.29584	0.0048949	53.8691	0	Hypermethylated	No	OTUD7A
cg23606718	chr2	131513927	-0.192302	0.0024361	53.8427	0	Hypermethylated	No	FAM123C
cg00059225	chr5	151304357	0.2564821	0.0023987	52.8361	0	Hypermethylated	No	GLRA1
cg23500537	chr5	140419819	0.2019473	0.0029768	52.4657	0	Hypermethylated	No	
cg07553761	chr3	160167977	-0.085898	0.0030009	52.1708	0	Hypermethylated	No	TRIM59
cg14674720	chr2	219827930	-0.15175	0.0022723	52.1475	0	Hypermethylated	No	
cg16419235	chr8	57360613	-0.110675	0.0021004	52.087	0	Hypermethylated	No	PENK
cg07082267	chr16	85429035	-0.234831	-0.0024153	-51.9394	0	Hypomethylated	No	
cg11970349	chr4	8582287	0.4395301	0.0024517	51.7603	0	Hypermethylated	No	GPR78
cg14556683	chr19	15342982	-0.354214	0.0030292	51.4444	0	Hypermethylated	No	EPHX3
cg06493994	chr6	25652602	-0.281467	0.0018639	51.2747	0	Hypermethylated	Yes	SCGN
cg19560758	chr1	8086721	0.123634	0.0017654	51.0739	0	Hypermethylated	No	ERRFI1
cg22736354	chr6	18122719	-0.328228	0.0023877	50.7215	0	Hypermethylated	Yes	NHLRC1
cg17885226	chr6	105388731	-0.011797	0.0030608	50.2096	0	Hypermethylated	No	
cg08262002	chr4	16575323	0.448234	-0.0036267	-50.1807	0	Hypomethylated	No	LDB2
cg18933331	chr1	110186418	0.1394501	-0.0026901	-49.3592	0	Hypomethylated	No	
cg00329615	chr3	118706648	0.3767479	-0.0049889	-49.1687	0	Hypomethylated	No	IGSF11
cg08097417	chr7	130419133	-0.212277	0.0018305	48.9874	0	Hypermethylated	No	KLF14
cg00748589	chr12	11653486	0.1822405	0.0024207	48.2695	0	Hypermethylated	No	
cg11084334	chr3	9594264	-0.022951	0.0027848	47.6682	0	Hypermethylated	No	LHFPL4
cg11071401	chr17	48637194	0.3081191	0.0023875	47.6374	0	Hypermethylated	No	CACNA1G
cg06784991	chr1	53308768	0.0728526	0.0021442	47.4979	0	Hypermethylated	No	ZYG11A
cg00439658	chr17	72848669	-0.187047	0.0019148	47.3396	0	Hypermethylated	No	GRIN2C
cg16054275	chr1	169556022	-0.308762	-0.0031404	-47.2773	0	Hypomethylated	No	F5
cg14692377	chr17	28562685	-0.319816	0.0019735	47.2725	0	Hypermethylated	No	SLC6A4
cg13649056	chr9	136474626	0.0939199	0.0018608	47.0121	0	Hypermethylated	No	
cg11693709	chr15	40542019	0.4398948	-0.0041179	-46.6849	0	Hypomethylated	No	PAK6
cg07080372	chr11	796607	-0.044385	-0.0020517	-46.5748	0	Hypomethylated	No	SLC25A22
cg19671120	chr2	98962974	0.2917162	0.0019275	46.5463	0	Hypermethylated	No	CNGA3
cg16219603	chr8	57360586	-0.243393	0.001599	46.4953	0	Hypermethylated	No	PENK
cg11705975	chr10	120354248	0.1345631	0.0025062	46.1335	0	Hypermethylated	No	PRLHR
cg15480367	chr14	93389485	0.1737257	0.0020641	46.1196	0	Hypermethylated	No	CHGA
cg24466241	chr1	53308908	-0.192473	0.0028258	45.9054	5.9288E-323	Hypermethylated	No	ZYG11A
cg02650266	chr4	147558239	-0.028284	0.0018604	45.5452	2.5444E-319	Hypermethylated	No	

cg03738025	chr6	105388694	0.1325219	0.0037303	45.5435	2.6480E-319	Hypermethylated	No	
cg08160331	chr11	75140865	0.1225186	0.0024513	45.5115	5.5982E-319	Hypermethylated	No	KLHL35
cg14361627	chr7	130419116	-0.029613	0.0024426	45.4145	5.4238E-318	Hypermethylated	No	KLF14
cg08128734	chr1	206685423	0.5891423	-0.0054386	-45.0487	2.8384E-314	Hypomethylated	No	RASSF5
cg26290632	chr8	91094847	0.2029635	0.0020152	45.0401	3.4695E-314	Hypermethylated	No	CALB1
cg01974375	chr1	151298954	0.0385361	-0.0019059	-45.0297	4.4226E-314	Hypomethylated	No	PI4KB
cg23479922	chr5	16179633	-0.5691	0.0045894	44.9595	2.2879E-313	Hypermethylated	No	MARCH11
cg09809672	chr1	236557682	0.175291	-0.0040059	-44.8504	2.9374E-312	Hypomethylated	Yes	EDARADD
cg00481951	chr3	187387650	0.1841224	0.0023342	44.6878	1.3200E-310	Hypermethylated	No	SST
cg03545227	chr2	220173100	0.0832971	0.0013552	44.5825	1.5491E-309	Hypermethylated	No	PTPRN
cg18618815	chr17	48275324	-0.292108	-0.0031805	-44.5025	1.0061E-308	Hypomethylated	No	COL1A1
cg11649376	chr12	81473234	0.1177648	-0.0025894	-44.4751	1.9099E-308	Hypomethylated	No	ACSS3
cg11436113	chr20	19191145	-0.245529	-0.0028774	-44.446	3.7798E-308	Hypomethylated	No	
cg20591472	chr1	110008990	0.2290873	0.0029438	44.3726	2.1018E-307	Hypermethylated	No	SYPL2
cg12757011	chr2	162281111	-0.036861	0.0022385	44.3402	4.4864E-307	Hypermethylated	No	TBR1
cg06570224	chr3	157812475	-0.255113	0.0021525	44.3003	1.1387E-306	Hypermethylated	No	
cg12878812	chr12	119419696	-0.152434	0.0017975	44.1946	1.3495E-305	Hypermethylated	No	SRRM4
cg07931844	chr15	72102213	-0.347225	-0.0020941	-44.1556	3.363E-305	Hypomethylated	No	NR2E3
cg15341124	chr14	102027734	0.1822515	0.0021014	43.8202	8.5279E-302	Hypermethylated	No	DIO3; MIR1247
cg12534424	chr7	127992316	-0.038607	0.0019362	43.5602	3.7086E-299	Hypermethylated	No	PRRT4
cg25410668	chr1	28241577	0.5378571	0.0033963	43.5204	9.4093E-299	Hypermethylated	No	RPA2
cg19392831	chr10	120355756	0.1002692	0.0017162	43.3469	5.4065E-297	Hypermethylated	No	PRLHR
cg16008966	chr1	114761794	0.2872323	-0.0024427	-43.054	5.0499E-294	Hypomethylated	No	
cg05308819	chr1	155959156	-0.383566	-0.0018965	-43.0379	7.3568E-294	Hypomethylated	No	
cg08468401	chr3	14303131	-0.481126	-0.0045074	-43.0226	1.0497E-293	Hypomethylated	No	
cg19855470	chr22	40060836	-0.111118	0.0015512	42.913	1.3565E-292	Hypermethylated	No	CACNA1I
cg11220950	chr16	2042693	0.0102849	0.0019377	42.8543	5.3374E-292	Hypermethylated	No	SYNGR3
cg16717122	chr15	51973920	0.3252301	0.00151	42.8415	7.1833E-292	Hypermethylated	No	SCG3
cg22156456	chr17	39844239	-0.229764	-0.0018499	-42.8279	9.8668E-292	Hypomethylated	No	EIF1
cg06335143	chr1	53308654	-0.088651	0.0022272	42.8111	1.4619E-291	Hypermethylated	No	ZYG11A
cg23746497	chr6	105388668	0.072451	0.0034686	42.7311	9.4375E-291	Hypermethylated	No	
cg08234504	chr5	139013317	-0.235634	-0.0015863	-42.72	1.2233E-290	Hypomethylated	No	
cg24436906	chr2	242498081	0.4803492	0.0019615	42.6333	9.2401E-290	Hypermethylated	No	BOK
cg13848598	chr10	115804578	-0.111233	0.0024786	42.4955	2.2983E-288	Hypermethylated	No	ADRB1
cg10804656	chr10	22623460	-0.950746	0.0028943	42.4594	5.3272E-288	Hypermethylated	No	
cg13135455	chr2	241860318	0.0059196	-0.0022231	-42.4071	1.8043E-287	Hypomethylated	No	
cg23078123	chr1	68577796	0.759047	-0.0026555	-42.3732	3.9744E-287	Hypomethylated	No	GPR177
cg13327545	chr10	22623548	-0.358846	0.0022651	42.3019	2.0954E-286	Hypermethylated	No	
cg03431918	chr17	77716367	0.1575907	-0.0017119	-42.2827	3.2734E-286	Hypomethylated	No	
cg01820374	chr12	6882083	-0.47997	-0.0022168	-42.2819	3.3323E-286	Hypomethylated	Yes	LAG3
cg20747538	chr3	137838021	-0.227794	-0.0019417	-42.2727	4.1287E-286	Hypomethylated	No	
cg27320127	chr2	47798396	0.3532211	0.0019054	42.2074	1.8912E-285	Hypermethylated	No	KCNK12
cg20273670	chr17	21356245	-0.202763	0.0032538	42.1546	6.4709E-285	Hypermethylated	No	
cg19702785	chr20	43727089	-0.307403	0.0016088	42.1542	6.5405E-285	Hypermethylated	No	KCNS1
cg14583999	chr3	10019040	0.051048	-0.0038329	-42.1149	1.6328E-284	Hypomethylated	No	TMEM111
cg01844642	chr3	51989764	-0.160677	0.0021369	42.1066	1.9788E-284	Hypermethylated	No	GPR62

cg00602811	chr2	145278564	-0.192604	-0.0038479	-42.1046	2.0743E-284	Hypomethylated	No	ZEB2
cg01770755	chr15	41914122	-0.106172	0.0017079	42.0334	1.089E-283	Hypermethylated	No	
cg00484358	chr1	110610995	0.2396367	0.0016647	42.0065	2.0361E-283	Hypermethylated	No	ALX3
cg18064714	chr7	20824556	-0.082174	0.00167	41.9065	2.0891E-282	Hypermethylated	No	SP8
cg16512661	chr5	2743620	0.2799574	0.0020114	41.717	1.7193E-280	Hypermethylated	No	
cg11741201	chr11	35638398	-0.069447	-0.0023228	-41.523	1.5688E-278	Hypomethylated	No	FJX1
cg22016779	chr2	230452311	-0.370728	-0.0023361	-41.4895	3.4156E-278	Hypomethylated	No	DNER
cg18473521	chr12	54448265	0.1111276	0.0041993	41.3931	3.2188E-277	Hypermethylated	No	HOXC4
cg01528542	chr12	81468232	-0.352352	-0.0036075	-41.3691	5.6171E-277	Hypomethylated	No	

Fig. S1.6 Table showing the characteristics of the top 100 differentially methylated positions during ageing (aDMPs) in the blood of the healthy individuals, ordered by p-value and the absolute value of the T statistic. The chromosome and coordinate refer to the *hg19* human genome assembly. The reported genes are the closest genes associated with the array probe, as specified by the 450K array annotation. In this case, cell composition correction (CCC) was applied during modelling (see section 2.1.4).

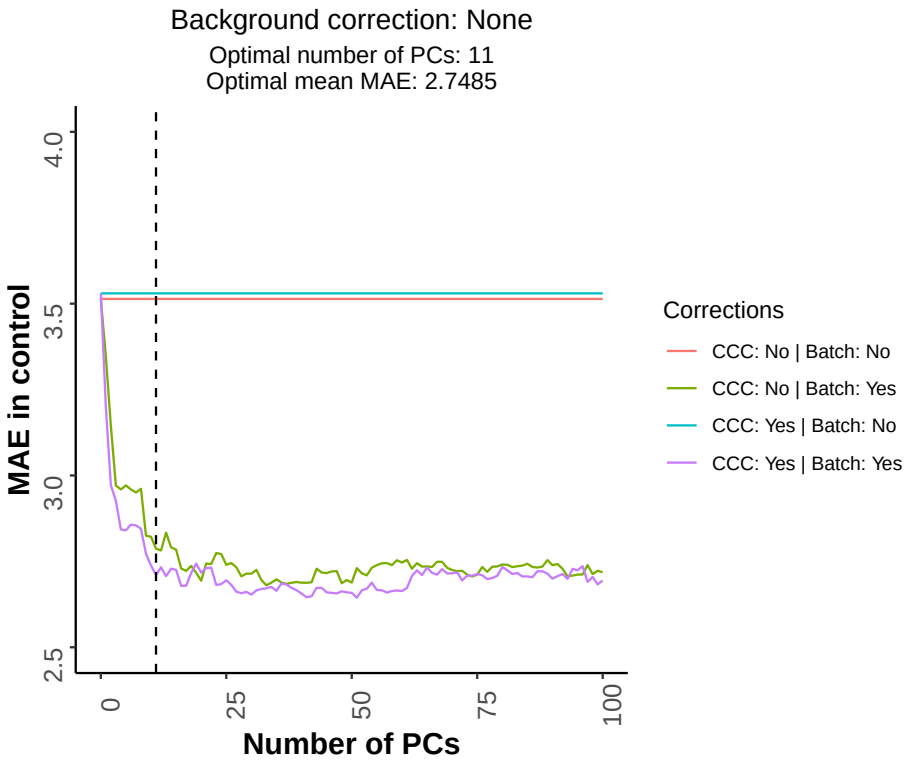


Fig. S1.7 Plot showing how the median absolute error (MAE) of the prediction in the healthy individual samples, that should tend to zero, is reduced when the PCs capturing the technical variation are included as part of the modelling strategy (see equations 2.16 and 2.17). The dashed line represents the optimal number of PCs (11) that was finally used. The optimal mean MAE is calculated as the average MAE between the green and purple lines. In this case, no background correction was applied to the methylation data before calculating the epigenetic ages according to Horvath's epigenetic clock [51].

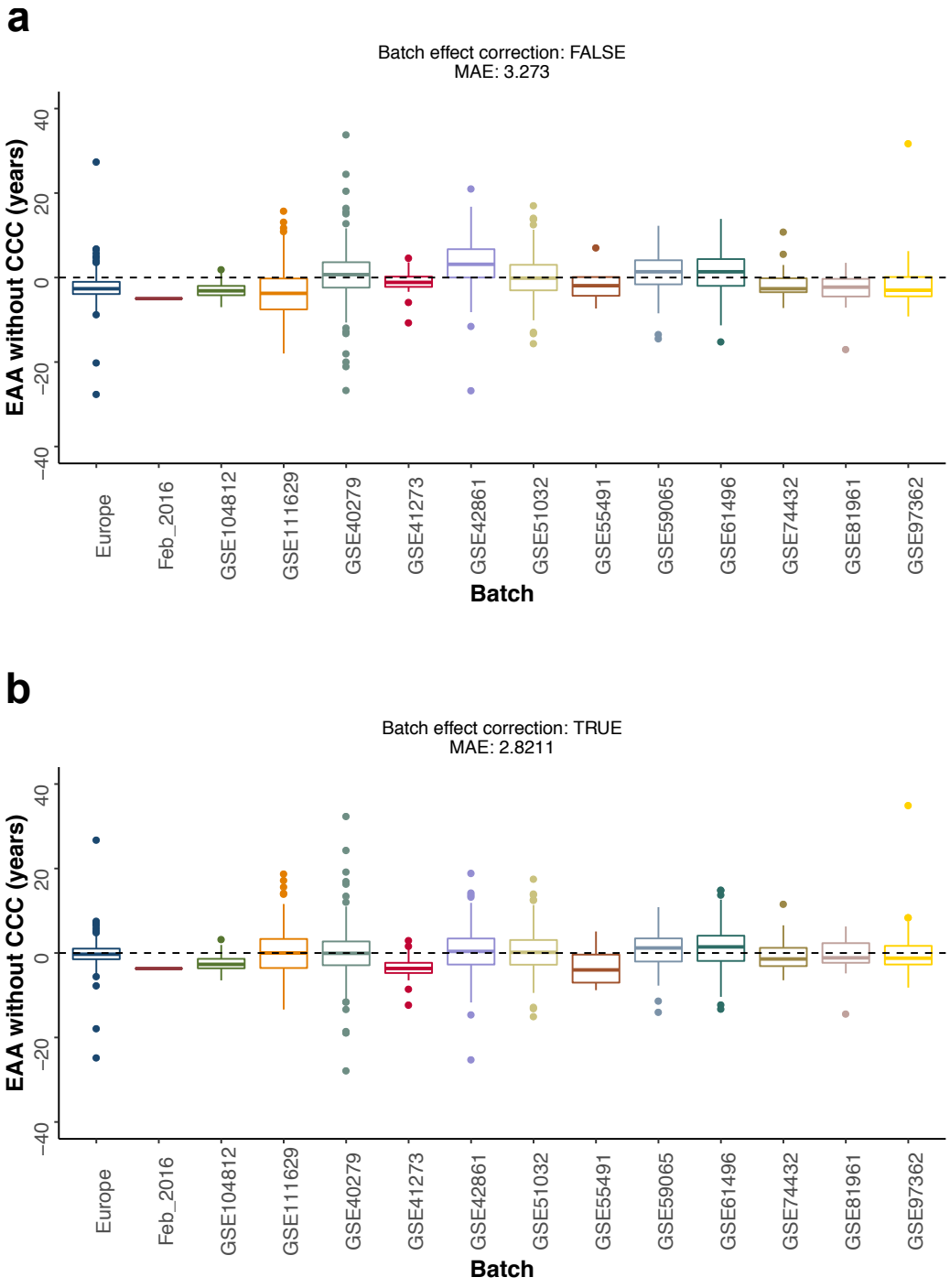


Fig. S1.8 Correcting for batch effects in the context of the epigenetic clock. **a.** Distribution of the epigenetic age acceleration (EAA) for the different batches of healthy individual samples, using the control model without cell composition correction (CCC) and before applying batch effect correction. The dashed black line represents $EAA = 0$, where the distributions should be centred around. **b.** As in a., but after applying batch effect correction (i.e. equivalent to equation 2.17).

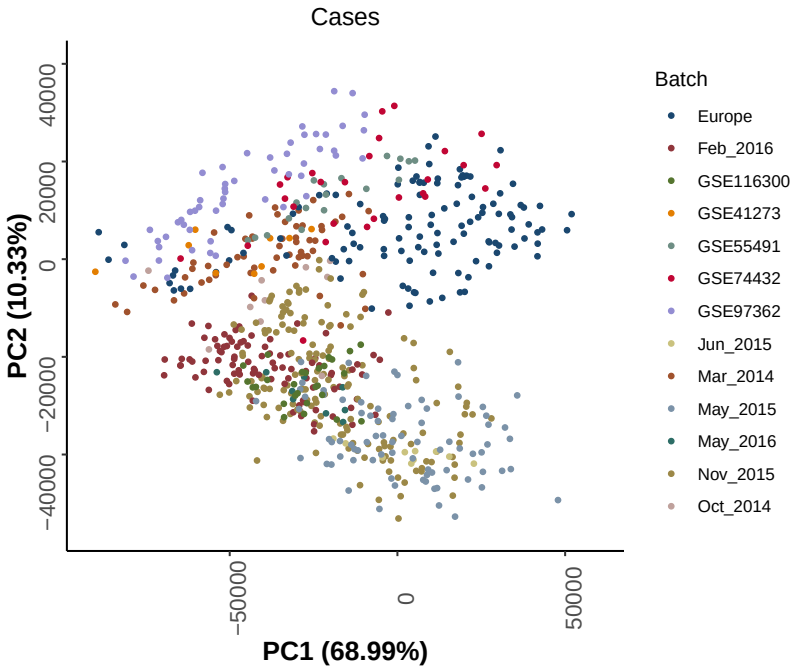


Fig. S1.9 Scatterplot showing the values of the first two principal components (PCs) for the samples with developmental disorders (cases, see Chapter 3) after performing PCA on the control probes of the 450K arrays. Each point corresponds to a different sample and the colours represent the different batches. The different batches cluster together in the PCA space, showing that the control probes indeed capture technical variation. Please note that all the PCA calculations were done using samples from both healthy individuals (full lifespan, $N = 2218$) and cases from developmental disorders ($N = 666$).

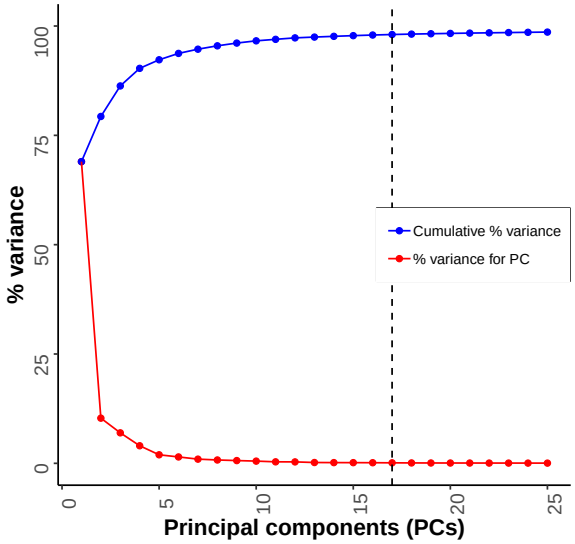


Fig. S1.10 Plot showing the percentages of technical variance explained by the different PCs from the control probes. The dashed line represents the optimal number of PCs (17) that was finally used.

S.2 Biological aspects of the epigenetic clock

S.3 Technological aspects of epigenetic clocks

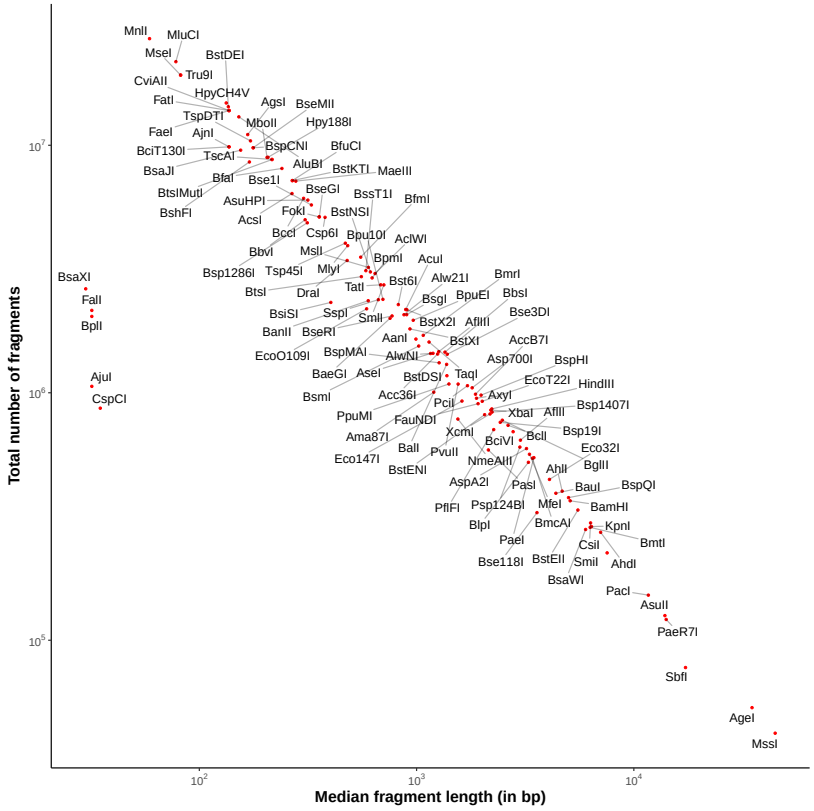


Fig. S3.1 Scatterplot which summarises the fragment length distributions for the same isoschizomer families portrayed in Fig 4.2a. The red dots represent the actual values of median fragment length and total number of fragments for each family. The black lines assign each name label to the correspondent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

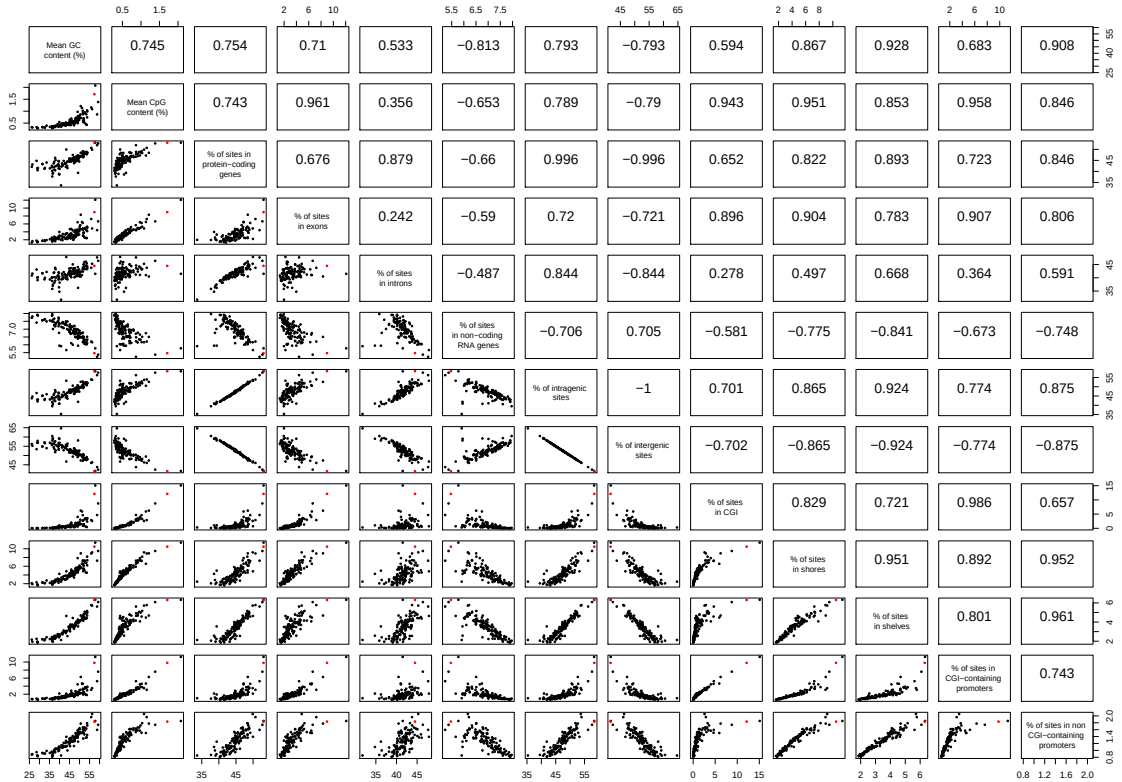


Fig. S3.2 Matrix of scatterplots showing the percentages of cleavage sites from different restriction enzymes that overlap with several genomic features (listed on the diagonal) in the human genome (hg38). The red dot in each scatterplot represents the values for MspI. The numbers above the diagonal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

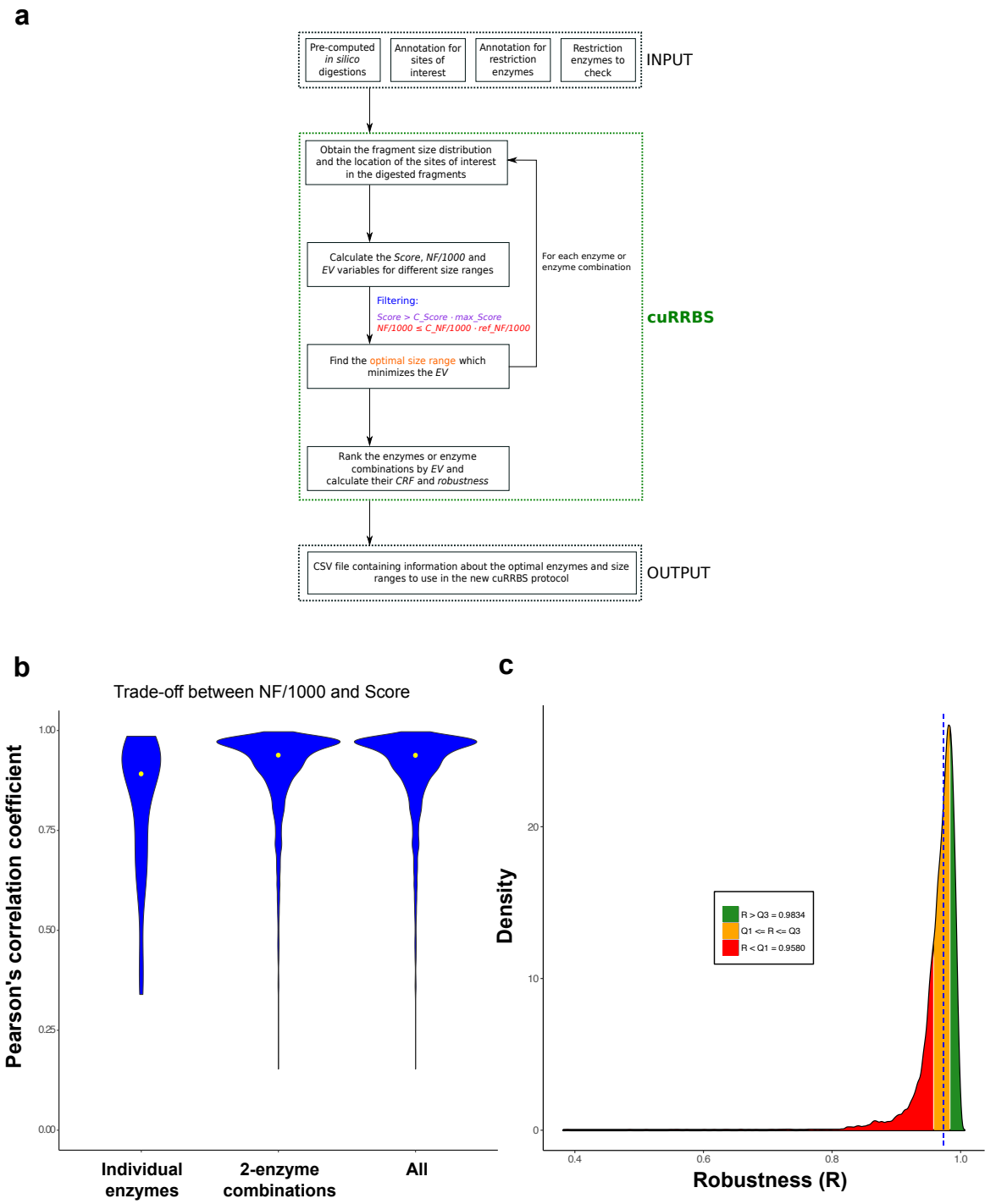


Fig. S3.4 Additional insights into cuRRBS. **a.** Detailed flowchart showing the input, main steps in cuRRBS and the output of the software. **b.** Violin plots showing the distribution of Pearson's correlation coefficients between the number of fragments (*NF*) and the *Score* for all the different enzymes tested with cuRRBS (single-enzyme, double-enzyme, all). In this example we used the Horvath epigenetic clock system [51], 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) [51, 137–142] with the same parameters as described in 'Running cuRRBS for different *in silico* systems' in section 4.7 (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 [51, 137–142] for which cuRRBS was run *in silico*. Some variables from the top hits in cuRRBS output are also reported.

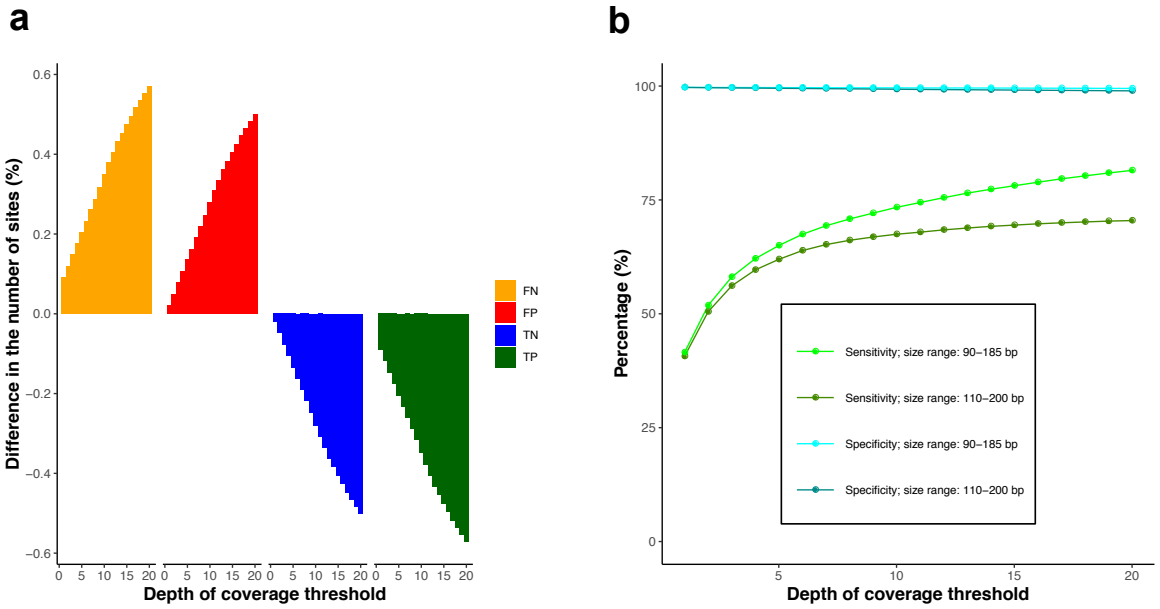


Fig. S3.6 Effect of experimental errors during size selection in cuRRBS predictions. **a.** Barplots showing the difference in the number of true positives (TP, in green), true negatives (TN, in blue), false positives (FP, in red) and false negatives (FN, in yellow) derived from cuRRBS theoretical predictions for the XmaI-RRBS data [121] 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 [121]. 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.

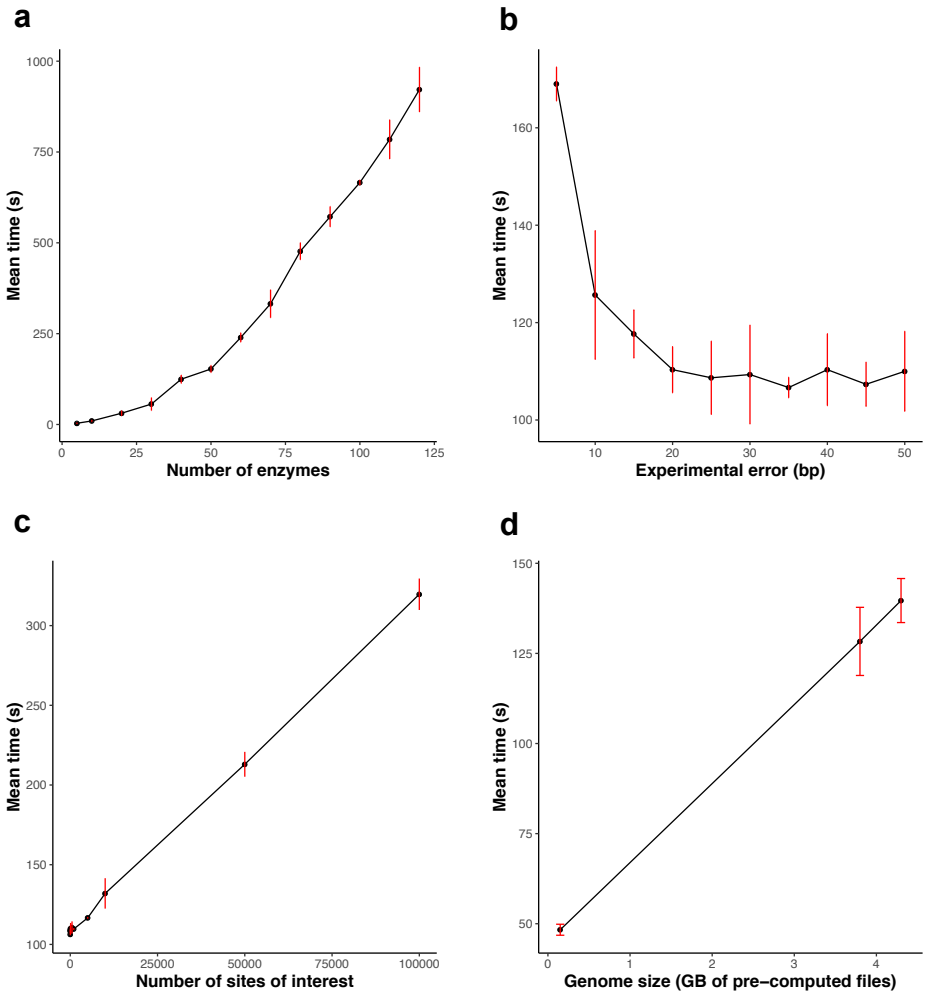


Fig. S3.7 cuRRBS computational efficiency. **a.** Plot showing the dependency between the number of enzymes checked and the computational (real) time required by the software (mean between 3 independent runs). cuRRBS was run for the Horvath epigenetic clock system [51] with a *read length* of 75 bp, 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 [51] 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] Andrew E Teschendorff and Shijie C Zheng. Cell-type deconvolution in epigenome-wide association studies: a review and recommendations. *Epigenomics*, 9(5):757–768, 2017.

[23] Lovisa E Reinius, Nathalie Acevedo, Maaike Joerink, Göran Pershagen, Sven-Erik Dahlén, Dario Greco, Cilla Söderhäll, Annika Scheynius, and Juha Kere. Differential DNA Methylation in Purified Human Blood Cells: Implications for Cell Lineage and Studies on Disease Susceptibility. *PLOS ONE*, 7(7):e41361, 2012.

[24] Yun Liu, Martin J Aryee, Leonid Padyukov, M Daniele Fallin, Espen Hesselberg, Arni Runarsson, Lovisa Reinius, Nathalie Acevedo, Margaret Taub, Marcus Ronninger, Klementy Shchetynsky, Annika Scheynius, Juha Kere, Lars Alfredsson, Lars Klareskog, Tomas J Ekström, and Andrew P Feinberg. Epigenome-wide association data implicate DNA methylation as an intermediary of genetic risk in rheumatoid arthritis. *Nature Biotechnology*, 31:142–147, 2013.

[25] Andrew E Jaffe and Rafael A Irizarry. Accounting for cellular heterogeneity is critical in epigenome-wide association studies. *Genome Biology*, 15(2):R31, 2014.

[26] Kevin McGregor, Sasha Bernatsky, Ines Colmegna, Marie Hudson, Tomi Pastinen, Aurélie Labbe, and Celia M T Greenwood. An evaluation of methods correcting for cell-type heterogeneity in DNA methylation studies. *Genome Biology*, 17(1):84, 2016.

[27] Marta Czesnikiewicz-Guzik, Won-Woo Lee, Dapeng Cui, Yuko Hiruma, David L Lamar, Zhi-Zhang Yang, Joseph G Ouslander, Cornelia M Weyand, and Jörg J Goronzy. T cell subset-specific susceptibility to aging. *Clinical Immunology*, 127(1):107–118, 2008.

[28] Klaudia Kuranda, Jacques Vargaftig, Philippe de la Rochere, Christine Dosquet, Dominique Charron, Florence Bardin, Cecile Tonnelle, Dominique Bonnet, and Michele Goodhardt. Age-related changes in human hematopoietic stem/progenitor cells. *Aging Cell*, 10(3):542–546, 2011.

[29] Yequn Chen, Yanhong Zhang, Guojun Zhao, Chang Chen, Peixuan Yang, Shu Ye, and Xuerui Tan. Difference in Leukocyte Composition between Women before and after Menopausal Age, and Distinct Sexual Dimorphism. *PLOS ONE*, 11(9):e0162953, 2016.

[30] Sebastian Seidler, Henning W Zimmermann, Matthias Bartneck, Christian Trautwein, and Frank Tacke. Age-dependent alterations of monocyte subsets and monocyte-related chemokine pathways in healthy adults. *BMC Immunology*, 11(1):30, 2010.

[31] Angela R Manser and Markus Uhrberg. Age-related changes in natural killer cell repertoires: impact on NK cell function and immune surveillance. *Cancer Immunology, Immunotherapy*, 65(4):417–426, 2016.

[32] Steve Horvath, Michael Gurven, Morgan E. Levine, Benjamin C. Trumble, Hillard Kaplan, Hooman Allayee, Beate R. Ritz, Brian Chen, Ake T. Lu, Tammy M. Rickabaugh, Beth D. Jamieson, Dianjianyi Sun, Shengxu Li, Wei Chen, Lluis Quintana-Murci, Maud Fagny, Michael S. Kobor, Philip S. Tsao, Alexander P. Reiner, Kerstin L. Edlefsen, Devin Absher, and Themistocles L. Assimes. An epigenetic clock analysis of race/ethnicity, sex, and coronary heart disease. *Genome Biology*, 17(1):171, 2016.

[33] Brian H. Chen, Riccardo E. Marioni, Elena Colicino, Marjolein J. Peters, Cavin K. Ward-Caviness, Pei Chien Tsai, Nicholas S. Roetker, Allan C. Just, Ellen W. Demerath, Weihua Guan, Jan Bressler, Myriam Fornage, Stephanie Studenski, Amy R. Vandiver, Ann Zenobia Moore, Toshiko Tanaka, Douglas P. Kiel, Liming Liang, Pantel Vokonas, Joel Schwartz, Kathryn L. Lunetta, Joanne M. Murabito, Stefania Bandinelli, Dena G. Hernandez, David Melzer, Michael Nalls, Luke C. Pilling, Timothy R. Price, Andrew B. Singleton, Christian Gieger, Rolf Holle, Anja Kretschmer, Florian Kronenberg, Sonja Kunze, Jakob Linseisen, Christine Meisinger, Wolfgang Rathmann, Melanie Waldenberger, Peter M. Visscher, Sonia Shah, Naomi R. Wray, Allan F. McRae, Oscar H. Franco, Albert Hofman, Andrï¿½ G. Uitterlinden, Devin Absher, Themistocles Assimes, Morgan E. Levine, Ake T. Lu, Philip S. Tsao, Lifang Hou, Jo Ann E. Manson, Cara L. Carty, Andrea Z. LaCroix, Alexander P. Reiner, Tim D. Spector, Andrew P. Feinberg, Daniel Levy, Andrea Baccarelli, Joyce van Meurs, Jordana T. Bell, Annette Peters, Ian J. Deary, James S. Pankow, Luigi Ferrucci, and Steve Horvath. DNA methylation-based measures of biological age: Meta-analysis predicting time to death. *Aging*, 8(9):1844–1865, 2016.

[34] Alexander J Titus, Rachel M Gallimore, Lucas A Salas, and Brock C Christensen. Cell-type deconvolution from DNA methylation: a review of recent applications. *Human Molecular Genetics*, 26(R2):R216–R224, 2017.

[35] Andrew E Teschendorff, Charles E Breeze, Shijie C Zheng, and Stephan Beck. A comparison of reference-based algorithms for correcting cell-type heterogeneity in Epigenome-Wide Association Studies. *BMC Bioinformatics*, 18(1):105, 2017.

[36] Andrew E Teschendorff and Caroline L Relton. Statistical and integrative system-level analysis of DNA methylation data. *Nature Reviews Genetics*, 19:129–147, 2018.

[37] Eugene Andres Houseman, William P Accomando, Devin C Koestler, Brock C Christensen, Carmen J Marsit, Heather H Nelson, John K Wiencke, and Karl T Kelsey. DNA methylation arrays as surrogate measures of cell mixture distribution. *BMC Bioinformatics*, 13:86, 2012.

[38] Devin C Koestler, Meaghan J Jones, Joseph Usset, Brock C Christensen, Rondi A Butler, Michael S Kobor, John K Wiencke, and Karl T Kelsey. Improving cell mixture deconvolution by identifying optimal DNA methylation libraries (IDOL). *BMC Bioinformatics*, 17:120, 2016.

[39] Andrew E. Teschendorff and Shijie C. Zheng. EpiDISH Bioconductor Package, 2017.

[40] Andrew E. Jaffe. FlowSorted.Blood.450k Bioconductor Package, 2018.

[41] Aaron M Newman, Chih Long Liu, Michael R Green, Andrew J Gentles, Weiguo Feng, Yue Xu, Chuong D Hoang, Maximilian Diehn, and Ash A Alizadeh. Robust enumeration of cell subsets from tissue expression profiles. *Nature Methods*, 12:453–457, 2015.

[42] Janko Nikolich-Žugich. The twilight of immunity: emerging concepts in aging of the immune system. *Nature Immunology*, 19(1):10–19, 2018.

[43] Claudio Franceschi. Inflammaging as a Major Characteristic of Old People: Can It Be Prevented or Cured? *Nutrition Reviews*, 65(s3):S173–S176, 2007.

[44] Ivan K Chinn, Clare C Blackburn, Nancy R Manley, and Gregory D Sempowski. Changes in primary lymphoid organs with aging. *Seminars in Immunology*, 24(5):309–320, 2012.

[45] Matthew E Ritchie, Belinda Phipson, Di Wu, Yifang Hu, Charity W Law, Wei Shi, and Gordon K Smyth. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Research*, 43(7):e47, 2015.

[46] Carlos Lopez-Otin, Maria A Blasco, Linda Partridge, Manuel Serrano, and Guido Kroemer. The hallmarks of aging. *Cell*, 153(6):1194–1217, 2013.

[47] 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.

[48] 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.

[49] Tianyu Zhu, Shijie C Zheng, Dirk S Paul, Steve Horvath, and Andrew E Teschendorff. Cell and tissue type independent age-associated DNA methylation changes are not rare but common. *Aging*, 10(11):3541–3557, 2018.

[50] Jenny van Dongen, Michel G Nivard, Gonneke Willemsen, Jouke-Jan Hottenga, Quinta Helmer, Conor V Dolan, Erik A Ehli, Gareth E Davies, Maarten van Iterson, Charles E Breeze, Stephan Beck, BIOS Consortium, Peter A.C.'t Hoen, René Pool, Marleen M J van Greevenbroek, Coen D A Stehouwer, Carla J H van der Kallen, Casper G Schalkwijk, Cisca Wijmenga, Sasha Zhernakova, Ettje F Tigchelaar, Marian Beekman, Joris Deelen, Diana van Heemst, Jan H Veldink, Leonard H van den Berg, Cornelia M van Duijn, Bert A Hofman, André G Uitterlinden, P Mila Jhamai, Michael Verbiest, Marijn Verkerk, Ruud van der Breggen, Jeroen van Rooij, Nico Lakenberg, Hailiang Mei, Jan Bot, Dasha V Zhernakova, Peter van't Hof, Patrick Deelen, Irene Nooren, Matthijs Moed, Martijn Vermaat, René Luijk, Marc Jan Bonder, Freerk van Dijk, Michiel van Galen, Wibowo Arindrarto, Szymon M Kielbasa, Morris A Swertz, Erik W van Zwet, Aaron Isaacs, Lude Franke, H Eka Suchiman, Rick Jansen, Joyce B van Meurs, Bastiaan T Heijmans, P Eline Slagboom, and Dorret I Boomsma. Genetic and environmental influences interact with age and sex in shaping the human methylome. *Nature Communications*, 7:11115, 2016.

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

[52] Paolo Garagnani, Maria G Bacalini, Chiara Pirazzini, Davide Gori, Cristina Giuliani, Daniela Mari, Anna M Di Blasio, Davide Gentilini, Giovanni Vitale, Sebastiano Collino, Serge Rezzi, Gastone Castellani, Miriam Capri, Stefano Salvioli, and Claudio Franceschi. Methylation of ELOVL2 gene as a new epigenetic marker of age. *Aging Cell*, 11(6):1132–1134, 2012.

[53] Renata Zbieć-Piekarska, Magdalena Spólnicka, Tomasz Kupiec, Żanetta Makowska, Anna Spas, Agnieszka Parys-Proszek, Krzysztof Kucharczyk, Rafał Płoski, and Wojciech Branicki. Examination of DNA methylation status of the ELOVL2 marker may be useful for human age prediction in forensic science. *Forensic Science International: Genetics*, 14:161–167, 2015.

[54] Maria Giulia Bacalini, Joris Deelen, Chiara Pirazzini, Marco De Cecco, Cristina Giuliani, Catia Lanzarini, Francesco Ravaioli, Elena Marasco, Diana Van Heemst, H. Eka D. Suchiman, Roderick Slieker, Enrico Giampieri, Rina Recchioni, Fiorella Marcheselli, Stefano Salvioli, Giovanni Vitale, Fabiola Olivieri, Annemieke M.W. Spijkerman, Martijn E.T. DollCrossed, John M. Sedivy, Gastone Castellani, Claudio Franceschi, Pieternella E. Slagboom, and Paolo Garagnani. Systemic Age-Associated DNA Hypermethylation of ELOVL2 Gene: In Vivo and in Vitro Evidences of a Cell Replication Process. *Journals of Gerontology - Series A Biological Sciences and Medical Sciences*, 72(8):1015–1023, 2017.

[55] Shyamalika Gopalan, Oana Carja, Maud Fagny, Etienne Patin, Justin W Myrick, Lisa M McEwen, Sarah M Mah, Michael S Kobor, Alain Froment, Marcus W Feldman, Lluis Quintana-Murci, and Brenna M Henn. Trends in DNA Methylation with Age Replicate Across Diverse Human Populations. *Genetics*, 206(3):1659–1674, 2017.

[56] 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.

[57] Maarten van Iterson, Erik W van Zwet, Bastiaan T Heijmans, and the BIOS Consortium. Controlling bias and inflation in epigenome- and transcriptome-wide association studies using the empirical null distribution. *Genome Biology*, 18(1):19, 2017.

[58] Shijie C Zheng, Charles E Breeze, Stephan Beck, and Andrew E Teschendorff. Identification of differentially methylated cell types in epigenome-wide association studies. *Nature Methods*, 15(12):1059–1066, 2018.

[59] Tina Wang, Brian Tsui, Jason F Kreisberg, Neil A Robertson, Andrew M Gross, Michael Ku Yu, Hannah Carter, Holly M Brown-Borg, Peter D Adams, and Trey Ideker. Epigenetic aging signatures in mice livers are slowed by dwarfism, calorie restriction and rapamycin treatment. *Genome Biology*, 18(1):57, 2017.

[60] Hehuang Xie, Min Wang, Alexandre De Andrade, Maria De F. Bonaldo, Vasil Galat, Kelly Arndt, Veena Rajaram, Stewart Goldman, Tadanori Tomita, and Marcelo B. Soares. Genome-wide quantitative assessment of variation in DNA methylation patterns. *Nucleic Acids Research*, 39(10):4099–4108, 2011.

[61] Garrett Jenkinson, Elisabet Pujadas, John Goutsias, and Andrew P Feinberg. Potential energy landscapes identify the information-theoretic nature of the epigenome. *Nature Genetics*, 49:719–729, 2017.

[62] Steve Horvath. DNAmAge online calculator: https://dnamage.genetics.ucla.edu/home, 2013.

[63] Daniel E Martin-Herranz. demh/epigenetic_ageing_clock: Epigenetic ageing clock v1.0.0. GitHub repository: https://github.com/demh/epigenetic_ageing_clock/, 2019.

[64] Olga Troyanskaya, Michael Cantor, Gavin Sherlock, Pat Brown, Trevor Hastie, Robert Tibshirani, David Botstein, and Russ B. Altman. Missing value estimation methods for DNA microarrays. *Bioinformatics*, 17(6):520–525, 2001.

[65] S Horvath, Y Zhang, P Langfelder, R S Kahn, M P Boks, and K Van Eijk. Aging effects on DNA methylation modules in human brain and blood tissue. *Genome Biol*, 13, 2012.

[66] Louis Y El Khoury, Tyler Gorrie-Stone, Melissa Smart, Amanda Hughes, Yanchun Bao, Alexandria Andrayas, Joe Burrage, Eilis Hannon, Meena Kumari, Jonathan Mill, and Leonard C Schalkwyk. Properties of the epigenetic clock and age acceleration. *bioRxiv*, page 363143, 2018.

[67] Riccardo E Marioni, Ian J Deary, Caroline L Relton, Matthew Suderman, Luigi Ferrucci, Brian H Chen, Steve Horvath, Stefania Bandinelli, Stephan Beck, Tiffany Morris, Nancy L Pedersen, and Sara Hägg. Tracking the Epigenetic Clock Across the Human Life Course: A Meta-analysis of Longitudinal Cohort Data. *The Journals of Gerontology: Series A*, 74(1):57–61, 2018.

[68] 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.

[69] Jeffrey T Leek, Robert B Scharpf, Héctor Corrada Bravo, David Simcha, Benjamin Langmead, W Evan Johnson, Donald Geman, Keith Baggerly, and Rafael A Irizarry. Tackling the widespread and critical impact of batch effects in high-throughput data. *Nature Reviews Genetics*, 11:733–739, 2010.

[70] Jovana Maksimovic, Alicia Oshlack, Johann A Gagnon-Bartsch, and Terence P Speed. Removing unwanted variation in a differential methylation analysis of Illumina HumanMethylation450 array data. *Nucleic Acids Research*, 43(16):e106–e106, 2015.

[71] Jean-Philippe Fortin, Aurélie Labbe, Mathieu Lemire, Brent W Zanke, Thomas J Hudson, Elana J Fertig, Celia M T Greenwood, and Kasper D Hansen. Functional normalization of 450k methylation array data improves replication in large cancer studies. *Genome Biology*, 15(11):503, 2014.

[72] E M Price and Wendy P Robinson. Adjusting for Batch Effects in DNA Methylation Microarray Data, a Lesson Learned , 2018.

[73] Steve Horvath, Peter Langfelder, Seung Kwak, Jeff Aaronson, Jim Rosinski, Thomas F. Vogt, Marika Eszes, Richard L.M. Faull, Maurice A. Curtis, Henry J. Waldvogel, Oi Wa Choi, Spencer Tung, Harry V. Vinters, Giovanni Coppola, and X. William Yang. Huntington's disease accelerates epigenetic aging of human brain and disrupts DNA methylation levels. *Aging*, 8(7):1485–1512, 2016.

[74] Steve Horvath. FAQs DNAmAge online calculator: https://horvath.genetics.ucla.edu/html/dnamage/faq.htm#_Toc385147421, 2013.

[75] Johann A Gagnon-Bartsch and Terence P Speed. Using control genes to correct for unwanted variation in microarray data. *Biostatistics*, 13(3):539–552, 2012.

[76] Illumina. GenomeStudio® Methylation Module v1.8 User Guide. Technical report, 2010.

[77] Altuna Akalin. AmpliconBiSeq GitHub repository: findElbow function, 2014.

[78] Riccardo E Marioni, Sonia Shah, Allan F McRae, Brian H Chen, Elena Colicino, Sarah E Harris, Jude Gibson, Anjali K Henders, Paul Redmond, Simon R Cox, Alison Pattie, Janie Corley, Lee Murphy, Nicholas G Martin, Grant W Montgomery, Andrew P Feinberg, M Daniele Fallin, Michael L Multhaup, Andrew E Jaffe, Roby Joehanes, Joel Schwartz, Allan C Just, Kathryn L Lunetta, Joanne M Murabito, John M Starr, Steve Horvath, Andrea A Baccarelli, Daniel Levy, Peter M Visscher, Naomi R Wray, and Ian J Deary. DNA methylation age of blood predicts all-cause mortality in later life. *Genome Biology*, 16(1):25, 2015.

[79] Laura Perna, Yan Zhang, Ute Mons, Bernd Holleczek, Kai-Uwe Saum, and Hermann Brenner. Epigenetic age acceleration predicts cancer, cardiovascular, and all-cause mortality in a German case cohort. *Clinical Epigenetics*, 8(1):64, 2016.

[80] Marguerite R Irvin, Stella Aslibekyan, Anh Do, Degui Zhi, Bertha Hidalgo, Steven A Claas, Vinodh Srinivasasainagendra, Steve Horvath, Hemant K Tiwari, Devin M Absher, and Donna K Arnett. Metabolic and inflammatory biomarkers are associated with epigenetic aging acceleration estimates in the GOLDN study. *Clinical Epigenetics*, 10(1):56, 2018.

[81] Zhen Yang, Andrew Wong, Diana Kuh, Dirk S. Paul, Vardhman K. Rakyan, R. David Leslie, Shijie C. Zheng, Martin Widschwendter, Stephan Beck, and Andrew E. Teschendorff. Correlation of an epigenetic mitotic clock with cancer risk. *Genome Biology*, 17(1):205, 2016.

[82] Isabel Beerman, Christoph Bock, Brian S. Garrison, Zachary D. Smith, Hongcang Gu, Alexander Meissner, and Derrick J. Rossi. Proliferation-dependent alterations of the DNA methylation landscape underlie hematopoietic stem cell aging. *Cell Stem Cell*, 12(4):413–425, 2013.

[83] Ake T Lu, Luting Xue, Elias L Salfati, Brian H Chen, Luigi Ferrucci, Daniel Levy, Roby Joehanes, Joanne M Murabito, Douglas P Kiel, Pei-Chien Tsai, Idil Yet, Jordana T Bell, Massimo Mangino, Toshiko Tanaka, Allan F McRae, Riccardo E Marioni, Peter M Visscher, Naomi R Wray, Ian J Deary, Morgan E Levine, Austin Quach, Themistocles Assimes, Philip S Tsao, Devin Absher, James D Stewart, Yun Li, Alex P

Reiner, Lifang Hou, Andrea A Baccarelli, Eric A Whitsel, Abraham Aviv, Alexia Cardona, Felix R Day, Nicholas J Wareham, John R B Perry, Ken K Ong, Kenneth Raj, Kathryn L Lunetta, and Steve Horvath. GWAS of epigenetic aging rates in blood reveals a critical role for TERT. *Nature Communications*, 9(1):387, 2018.

[84] Mary E. Sehl, Jill E. Henry, Anna Maria Storniolo, Patricia A. Ganz, and Steve Horvath. DNA methylation age is elevated in breast tissue of healthy women. *Breast Cancer Research and Treatment*, 164(1):209–219, 2017.

[85] Peter M. Visscher, Naomi R. Wray, Qian Zhang, Pamela Sklar, Mark I. McCarthy, Matthew A. Brown, and Jian Yang. 10 Years of GWAS Discovery: Biology, Function, and Translation, 2017.

[86] Morris L. Eaton. Linear Statistical Models. In *Multivariate Statistics: A Vector Space Approach*, pages 132–158. 2007.

[87] Simon J. Sheather. *A Modern Approach to Regression with R*. 2009.

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

[89] 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.

[90] 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.

[91] 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.

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

[93] 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.

[94] 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.

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

[96] 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.

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

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

[99] 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.

[100] 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.

[101] 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.

[102] 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.

[103] 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.

[104] 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.

[105] 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.

[106] 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.

[107] 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.

[108] 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.

[109] 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.

[110] 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.

[111] 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.

[112] 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.

[113] 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.

[114] 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.

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

[116] 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.

[117] 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.

[118] 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.

[119] 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.

[120] 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.

[121] 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.

[122] 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.

[123] 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.

[124] 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.

[125] 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.

[126] 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.

[127] 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.

[128] 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.

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

[130] 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.

[131] 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.

[132] 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.

[133] 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.

[134] 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.

[135] Daniel E Martin-Herranz, Erfan Aref-Eshghi, Marc Jan Bonder, Thomas M Stubbs, Oliver Stegle, Bekim Sadikovic, Wolf Reik, and Janet M Thornton. Screening for genes that accelerate the epigenetic ageing clock in humans reveals a role for the H3K36 methyltransferase NSD1. *bioRxiv*, page 545830, 2019.

[136] 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.

[137] 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.

[138] 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.

[139] 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.

[140] 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.

[141] 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.

[142] 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.

[143] 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.

[144] 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.

[145] 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.

[146] 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.

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

[148] 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.

[149] 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.

[150] 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.

[151] 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.

[152] 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.

[153] 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.

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

[155] 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.

[156] 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.

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

[158] 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.

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

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

[161] 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.

[162] 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.

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