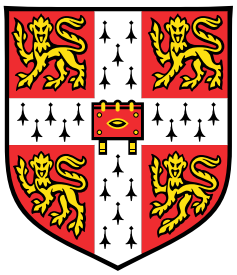
# On the epigenetic ageing clock in humans



**Daniel Elías Martín Herranz**

European Molecular Biology Laboratory,
European Bioinformatics Institute
University of Cambridge

This dissertation is submitted for the degree of
*Doctor of Philosophy*

Churchill College April 2019

A mi familia capicúa, Andrés, Pilar y Andrés.
Porque estas páginas de ciencia son un reflejo de su arte.

# Declaration

This dissertation is the result of my own work and includes nothing which is the outcome of work done in collaboration with others, except when specified in the declarations at the beginning of the chapters. I further specify this by using the pronoun 'we' when others were substantially involved in the work and 'I' for those parts that are purely my own work.

It is not substantially the same as any that I have submitted, or, is being concurrently submitted for a degree or diploma or other qualification at the University of Cambridge or any other University or similar institution. I further state that no substantial part of my dissertation has already been submitted, or, is being concurrently submitted for any such degree, diploma or other qualification at the University of Cambridge or any other University or similar institution. This dissertation contains fewer than 60,000 words exclusive of tables, footnotes, bibliography, and appendices and has fewer than 150 figures.

Daniel Elías Martín Herranz
April 2019

# Acknowledgements

This thesis has made use of a great amount of chronological time (hopefully not too much biological time) of a lot of people. This is also their work. I am deeply thankful ...

... to Janet Thornton, for opening the doors of the EBI to me, showing me the true nature of critical thinking, science and proper discussion, and for supporting my (sometimes) wild ideas and plans;

... to Wolf Reik, who is responsible for my scientific crush on epigenetics, for accepting me as an unofficial student, providing always stimulating ideas and inviting me to his garden parties;

... to Tom Stubbs, for his scientific creativity, friendship and burrito evenings;

... to the rest of my collaborators, especially Marc Jan Bonder, Antonio Ribeiro and Erfan Aref-Eshghi, for their contributions;

... to the rest of my TAC members, Oliver Stegle, Judith Zaugg and Gos Micklem, for their guidance;

... to Nils Eling, Hannah Meyer, Jack Monahan and Max Stammnitz; for taking the time to read through these pages and send me their thoughts and comments;

... to the incredible people in the Thornton and Reik labs, for their input and many shared lunches (and some beers);

... a mi familia, por su amor y apoyo siempre incondicional (y por alimentarme tan bien);

... a Parvathi 'Ale' Subbiah, porque su 'efecto' me ha dado fuerza todos los días desde que la conocí (y por ayudarme con el diseño de las figuras);

... to the EMBL-EBI crowd, especially to Jack, Nils, Lara, Omar, Hannah and Julia, for many good times at the Blue Moon and the Wiggle Mansion;

... to the rest of the Cambridge crowd, including members of Los del Cam (Max, Vlad, Gogi, Ale), Churchill College (Barbora, basketball team), the CompBio MPhil (Daniel, Elias, Dalia, Andy) and becari@s La Caixa; for keeping me sane in this bubble;

... a mis amigos de Salamanca (Salón del Té) y de Soria (Club de Bebedores Mercadona); por su eterna amistad;

... to La Caixa and EMBL, for funding me and giving me the opportunity to be writing these words;

... to all those people that I forgot to include because of my procrastination, they know who they are.

# Abstract

Epigenetic clocks are mathematical models that predict the biological age of an organism using DNA methylation data, and which have emerged in the last few years as the most accurate biomarkers of the ageing process. However, little is known about the molecular mechanisms that control the rate of such clocks. In this thesis I focus on the study of the epigenetic ageing clock in humans. First, I review and benchmark statistical and computational tools required for the analysis of DNA methylation data in the context of human ageing. Next, I validate the performance of the Horvath epigenetic clock, the most widely used multi-tissue epigenetic clock in humans, in a control blood dataset and test its behaviour in patients with a variety of developmental disorders, which harbour mutations in proteins of the epigenetic machinery. I demonstrate that loss-of-function mutations in the H3K36 methyltransferase NSD1, which cause Sotos syndrome, substantially accelerate epigenetic ageing. Furthermore, I show that the normal ageing process and Sotos syndrome share methylation changes and the genomic context in which they happen. These results suggest that the H3K36 methylation machinery is a key component of the epigenetic maintenance system in humans, which controls the rate of epigenetic ageing, and this role seems to be conserved in model organisms. Finally, I provide a technological strategy to make epigenetic clocks (or any DNA methylation-based mathematical models) more cost-effective by exploiting the ability of restriction enzymes to perform genomic enrichment. This thesis provides novel insights (statistical, biological, technological) into the epigenetic ageing clock in humans, which will help to shed light on the different processes that erode the human epigenetic landscape during ageing.

# Table of contents

# List of figures

# List of tables

# Abbreviations and acronyms

| | |
|---|---|
| 27K | Illumina Infinium HumanMethylation27 array |
| 450K | Illumina Infinium HumanMethylation450 array |
| 5caC | 5-carboxylcytosine |
| 5fC | 5-formylcytosine |
| 5hmC | 5-hydroxymethylcytosine |
| 5mC | 5-methylcytosine |
| | |
| a.k.a. | Also known as |
| aDMPs | Differentially methylated positions during ageing |
| AMP | Adenosine monophosphate |
| AMPK | Adenosine monophosphate-activated kinase |
| ASD | Autism spectrum disorder |
| ATP | Adenosine triphosphate |
| ATR-X | Alpha thalassemia/mental retardation X-linked syndrome |
| aVMPs | Variably methylated positions during ageing |
| | |
| B | $CD19^+$ B cells |
| BER | Base excision repair |
| 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 |
| cSEA | Shannon entropy acceleration for the Horvath's epigenetic clock sites |
| 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 |
| DMV | DNA methylation valley |
| 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) |
| ESCs | Embryonic stem cells |
| etc. | *Et cetera* |
| EV | Enrichment Value in cuRRBS |
| EWAS | Epigenome-wide association studies |

| | |
|---|---|
| FDR | False discovery rate |
| FN | False negatives |
| FP | False positives |
| FXS | Fragile X syndrome |
| GB | Gigabytes |
| Gbp | Giga base pairs |
| GC content | Guanine + cytosine content |
| GEO | Gene Expression Omnibus repository |
| Gran | Granulocytes |
| gSEA | Genome-wide Shannon entropy acceleration |
| GWAS | Genome-wide association studies |
| H3K27me3 | Histone H3 lysine 27 trimethylation |
| H3K36 | Histone H3 lysine 36 |
| H3K36me3 | Histone H3 lysine 36 trimethylation |
| H3K4me3 | Histone H3 lysine 4 trimethylation |
| hg19 | Reference human genome assembly 19 |
| hg38 | Reference human genome assembly 38 |
| hQTLs | Histone quantitative trait loci |
| HSCs | Haematopoietic stem cells |
| i.e. | *Id est* |
| IDOL | IDentifying Optimal DNA methylation Libraries, a strategy to build cell-type deconvolution references |
| IEAA | Intrinsic epigenetic age acceleration |
| IGF-1 | Insulin-like growth factor 1 |
| iPSCs | Induced pluripotent stem cells |
| kb | Kilo base pairs |
| KNN | *k*-nearest neighbours |

| | |
|---|---|
| $m^6A$ | $N^6$-methyladenosine |
| 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 |
| meQTLs | Methylation quantitative trait loci |
| Mono | $CD14^+$ monocytes |
| mRNA | Messenger RNA |
| $NAD^+$ | Nicotinamide adenine dinucleotide |
| NF | Theoretical number of fragments sequenced in cuRRBS |
| NFC | Normalised fold change |
| NK | $CD56^+$ natural killer cells |
| NRC | Normalised read counts |
| NRE | Normalised RNA expression |
| NRF1 | Nuclear respiratory factor 1 |
| OOB | Out-of-band fluorescence intensities in the Infinium I probes of Illumina arrays |
| OR | Odds ratio |
| 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 |
| PGCs | Primordial germ cells |
| 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 |
| Repli-seq | genome-wide analysis of replication timing by sequencing |
| RMSE | Root mean squared error |
| RNA | Ribonucleic acid |
| RNA-seq | RNA sequencing |
| ROS | Reactive oxigen species |
| RPC | Robust partial correlations |
| RRBS | Reduced Representation Bisulfite Sequencing |
| rRNA | Ribosomal RNA |
| SASP | Senescence-associated secretory phenotype |
| SCC | Spearman's correlation coefficient |
| SD | Standard deviation |
| $\text{Sex}_p$ | Sex predicted for a sample using DNA methylation data |
| SNP | Single-nucleotide polymorphism |
| SQN | Stratified quantile normalisation |
| sur | Signal of unique reads |
| TDG | Thymine DNA glycosylase |
| TKO | Triple knockout |
| TN | True negatives |
| TOR | Target of rapamycin |
| TP | True positives |
| TSS | Transcription start site |
| UTR | Untranslated region |
| WGBS | Whole Genome Bisulfite Sequencing |
| WTS | Wavelet-transformed signals |

# Chapter 1

# Introduction

'[...] there are as many theories of ag[e]ing as there are biogerontologists.'

L. Hayflick, 2007 [1]

## 1.1 The biology of ageing

### 1.1.1 A brief introduction to ageing theory

The ageing process is one of the most mysterious, complex and fascinating biological problems to be solved in the 21st century. Ageing and immortality have probably fascinated mankind since we have a conception of time and death [2].

**Biological ageing** (a.k.a. the ageing process) can be broadly defined as the time-dependent functional decline which increases vulnerability to death in most organisms [3]. The revolution taking place in genetics and molecular biology during the 20th century gave rise to more than 300 theories that attempt to explain the mechanisms behind biological ageing [4]. Any valid modern theory of ageing would need to explain at least two things [4]:

- The molecular basis for the increase in **mortality rate** (a.k.a. death rate) over time in the population of a given species. Mortality rate can be broadly defined as the number of deaths in a population per unit of time and scaled by the size of the population. More formally, by quantifying the deaths of individuals in a population over time (and assuming that there are no increases in the population number due to reproduction, migration, etc.), the survival fraction at a given time $t$, $S(t)$, is [5]:

$$S(t) = \frac{N(t)}{N_0} \tag{1.1}$$

where $N(t)$ is the number of individuals alive at a given time $t$ and $N_0$ is the initial number of individuals in the population. It can be demonstrated that the mortality rate, $\lambda(t)$, can be expressed as [5]:

$$\lambda(t) = -\frac{1}{S(t)} \cdot \frac{dS(t)}{dt} \tag{1.2}$$

- The **evolutionary variations in lifespan between different species** [6]; where lifespan is defined as the time passed between birth and death of an organism. For example, the maximum lifespan in the case of the roundworm (*Caenorhabditis elegans*) is 0.16 years (58.4 days, in captivity); in the case of the fruit fly (*Drosophila melanogaster*) is 0.3 years (109.5 days, in captivity); in the case of the house mouse (*Mus musculus*) is 4 years (in captivity); in the case of humans (*Homo sapiens*) is 122.5 years and in the case of the bowhead whale (*Balaena mysticetus*) is 211 years (in the wild) according to the database AnAge [7]. Furthermore, some species (such as certain turtles, certain species of rockfish or the bristlecone pine) seem to have negligible senescence i.e. negligible changes in adult mortality rates over extended periods of time at advanced adult ages [8].

Nowadays, there are at least **two main paradigms**, complementary to each other, that try to conceptualise the problem and that are a topic of intense discussion among biogerontologists:

- Ageing as a consequence of *molecular infidelity*. In this case, stochastic chemical modifications of biomolecules, such as DNA or proteins, exceed the capacity of the repair and turnover systems of the organism and accumulate over time, which increases the entropy of the system. This leads to changes in molecular structure and, finally, changes in function, which increase vulnerability to age-related diseases [1, 9]. From an evolutionary point of view, this fits into the *disposable soma theory*, originally proposed by Thomas Kirkwood in 1977. This theory suggests that organisms have evolved to optimise the amount of energy dedicated to repair errors in somatic cells in order to maximise reproductive success (at the expense of indefinite survival) [10, 11].

- Ageing as a consequence of *hyperfunction*. In this case, the primary cause of ageing is an excessive activity of certain growth or development-related genes and pathways in later life [12–15]. In other words, ageing originates from developmental programmes that have not been turned off [12]. This idea is rooted in the concept of *antagonistic pleiotropy*, an important pillar of the evolutionary theory of ageing originally proposed by George C. Williams in 1957 [16]. It implies that certain genes have opposite effects on fitness at different ages, which is a consequence of the decrease in selection forces after reproductive age. A strong candidate is the TOR (target of rapamycin) pathway, which promotes development in early life but also the advancement of several late-life pathologies [13].

It has become clear that no single molecular mechanism will be able to explain ageing across all kingdoms of life. Different species have different life histories that are subjected to evolutionary trade-offs (e.g. regarding reproduction strategies, developmental schedules, etc.) and that can affect the rate of ageing [6, 17]. Nevertheless, it is possible to integrate all the ideas presented so far into a **theoretical framework** that can help to unify definitions across studies and set the foundations for mechanistic advancements on the biology of ageing (Fig. 1.1, inspired by ideas from [1, 15, 18–20]). Under this theoretical framework:

- The ageing process is composed of different molecular mechanisms (subprocesses) that operate at different stages of life and contribute, in variable proportions, to the appearance of different age-related diseases i.e. the risk of developing an age-related disease is the 'integral of its ageing subprocesses operating over time'. Furthermore, the development of different diseases affects the mortality rate and, thus, the probability of dying. The different ageing processes can also be understood as the sources of ageing-associated molecular damage [3].

- If the ageing subprocesses can be altered through different genetic, lifestyle or pharmacological interventions, it is possible to reduce the likelihood of several age-related diseases at the same time. This makes ageing research incredibly relevant to the biomedical sciences, since it changes the current paradigm away from developing interventions for a specific already-existing disease towards the prevention of several diseases simultaneously.

- Differences in the average lifespan between different species should be explained by different combinations of ageing subprocesses and their rates.

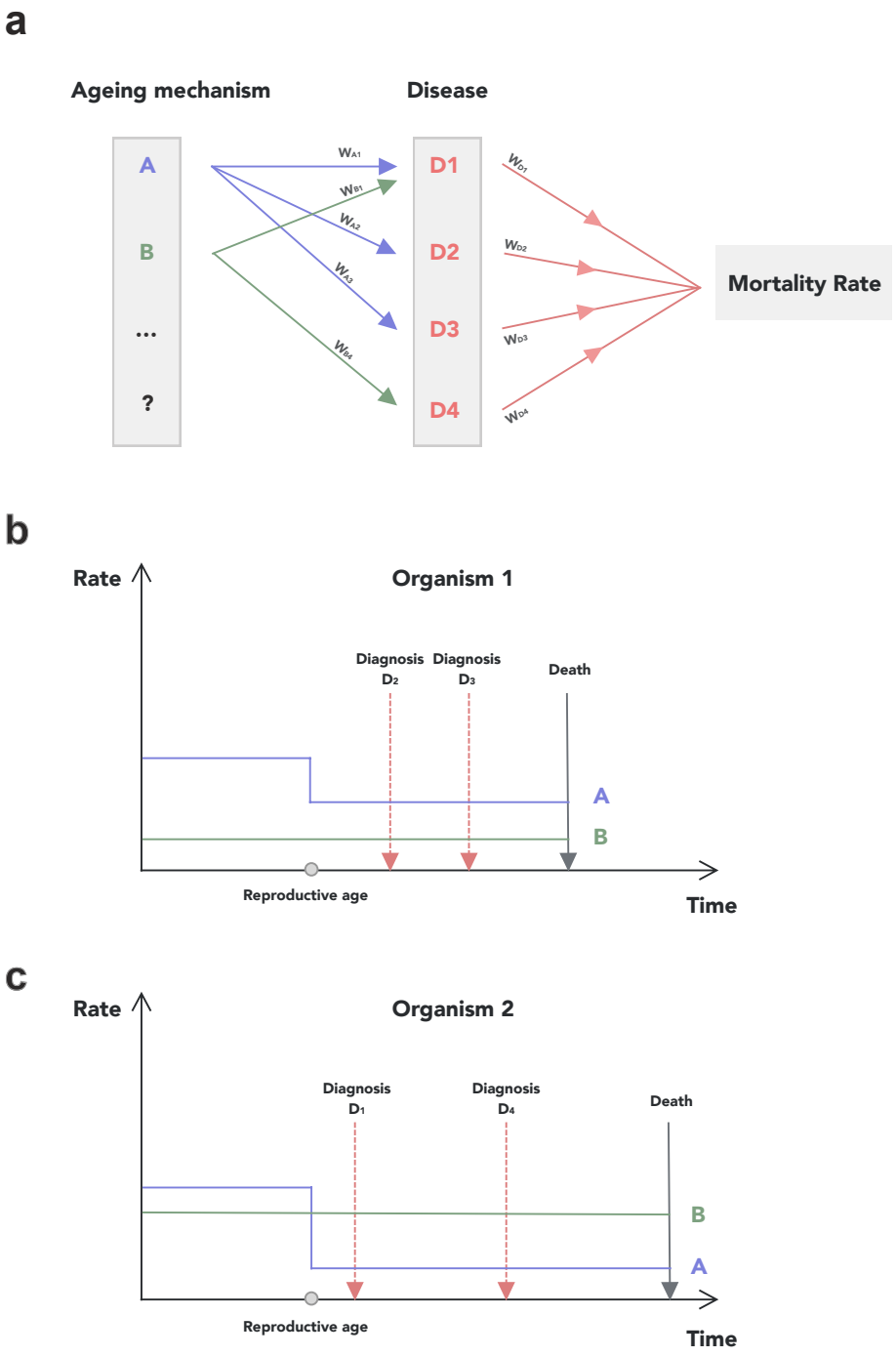



**Fig. 1.1** Theoretical framework to conceptualise the ageing process. **a.** The ageing process is composed of different molecular mechanisms (subprocesses) that operate at different stages of life and contribute, in variable proportions (specified by the weights), to the appearance of different age-related diseases. Furthermore, the development of different diseases affects the mortality rate and, thus, the probability of dying. **b.** and **c.** Examples of the life histories of two organisms. In these examples, two ageing mechanisms operate: A (which changes its rate after reproductive age e.g. activated growth-related pathways) and B (with a constant rate over time e.g. some type of (epi)mutational process). Differences in the mechanisms' profiles lead to differences in the age-related diseases that manifest over the lifespan of the organisms, even though the molecular mechanisms are the same. This, affects the mortality rate and, ultimately, the time-to-death. This figure is inspired by ideas from [1, 15, 18–20].

Consequently, systems biology approaches become fundamental to understand the ageing process [20]. In the next sections, I will provide an overview of the ageing mechanisms that may operate in different species, with a special focus on mammalian species.

### 1.1.2 The genetic basis of ageing

Given the large variability in lifespan between species [6], it is nowadays clear that the ageing process must have a genetic basis. However, for a long time, the ageing process was thought to be a 'haphazard process driven solely by entropy' [21]. Furthermore, in 1935 Clive Maine McCay had shown that caloric restriction (a reduction in calories intake without malnutrition) could extend mean and maximal lifespan in rats [22, 23], which probably shifted the focus towards environmental or external causes as the main driver forces of the ageing process. Since then, dietary restriction (which includes different types of dietary interventions that reduce food intake without malnutrition) has been established as the most successful non-genetic intervention to slow down the ageing process across species [24].

The establishment of the nematode *Caenorhabditis elegans* as a model organism in the 70s triggered its adoption in the ageing field [25], since it allowed well-controlled experiments in a much shorter period of time than rodents [26]. This lead to the discovery of the first mutants that dramatically extended lifespan, which mapped to genes in the insulin/IGF-1 signalling pathway [27, 28]. Since then, many genes have been found to significantly affect the lifespan of other model organisms as well, such as in budding yeast (*Saccharomyces cerevisiae*), in fruit flies (*Drosophila melanogaster*) and in mice (*Mus musculus*) [21, 29, 30].

Interestingly, the effects of many of these genetic mutations and their pathways are shared by distantly-related species. This suggests that at least part of the molecular mechanisms that drive the ageing process could be evolutionarily conserved. Among these ageing-related signalling pathways it is worth highlighting (Fig. 1.2) [21, 29–31]:

- **Insulin/IGF-1 pathway**. This underscores the central role of the endocrine system on the biology of ageing. Mutations that lower the level of *daf-2*, encoding an insulin/IGF-1 receptor, were originally found to double the lifespan of *C. elegans* [27, 32]. Activation of the insulin/IGF-1 pathway, a PI3K pathway, leads to the phosphorylation of a transcription factor of the FOXO family, encoded by *daf-16* in *C. elegans*, which prevents it to reach the nucleus [33]. FOXO transcription factors, of which there are several members in mammals, activate the expression of longevity-promoting genes involved in processes such as autophagy (which clears protein aggregates and damaged

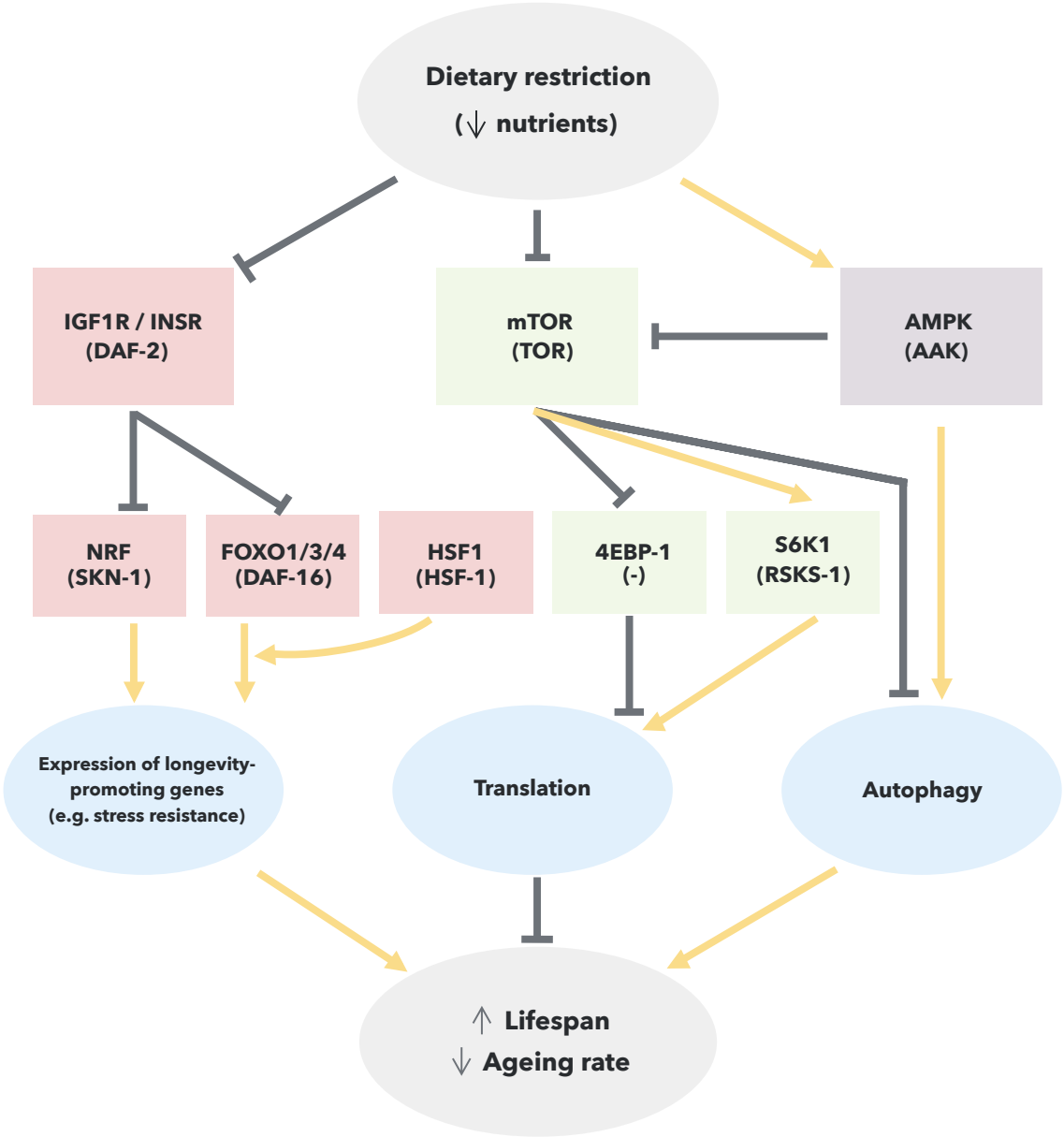



**Fig. 1.2** Main signalling pathways that affect the ageing process. These pathways sense nutrient and stress inputs (such as a dietary restriction regime) to ultimate impact the rate of ageing. The grey lines represent inhibition (negative regulation) while the yellow arrows represent activation (positive regulation). As such, dietary restriction inhibits the insulin/IGF-1 pathway (in red), inhibits the TOR pathway (in green) and activates AMPK signalling (in purple), ultimately extending lifespan. For simplification, I have only included the main proteins that transduce the signal (e.g. there are more intermediate kinases in the insulin/IGF-1 pathway). Protein names are provided for both the mammalian (top) and the *C. elegans* orthologs if available (bottom, in parenthesis).

organelles in the cell) [30], resistance to oxidative stress or stem cell maintenance [34]. This partially explains why the inhibition of the insulin/IGF-1 pathway can increase organismal lifespan. However, other downstream targets that regulate gene expression have also been identified, such as *hsf-1* (a transcription factor that regulates heat-shock response) [35] or *skn-1* (a transcription factor that coordinates a response to oxidative stress) [36] in *C. elegans*.

- **TOR pathway**. TOR (target of rapamycin) is a kinase that acts as a major amino-acid and nutrient sensor by stimulating growth (including protein translation) and blocking autophagy [29]. The effects of TOR are partly mediated by activating the ribosomal subunit S6 kinase (which promotes protein translation) and by inhibiting 4EBP (a translation inhibitor) [29, 37]. Reductions in TOR activity (via genetic or pharmacological mechanisms) increase lifespan across many species [29]. Importantly, rapamycin, a drug that inhibits TOR, can increase the mean lifespan of mice when fed late in life, which showed for the first time that pharmacological interventions targeting mammalian ageing are possible [38]. Interestingly, the increase in lifespan differed in males (9%) and females (13%) [38], highlighting the sex-specific effects of some ageing mechanisms.

- **AMPK pathway**. The AMP-activated kinase (AMPK) controls the balance between catabolic and anabolic processes depending on the cellular levels of AMP/ATP (i.e. when ATP levels decrease, AMPK is activated to promote catabolic pathways) [29, 39]. Furthermore, AMPK activation promotes autophagy, partially by inhibiting TOR [39]. The anti-diabetic drug metformin, which activates AMPK among other targets, has been shown to extend lifespan in mice [40, 41] and has been included as the first drug to target the human ageing process in a clinical trial [42].

- **Sirtuins**. Sirtuins are a family of nicotinamide adenine dinucleotide ($NAD^+$)-dependent deacetylases i.e. they generally catalyse the removal of an acetyl group from lysine residues using $NAD^+$ as a cofactor [43]. Sirtuins have been shown to play complex roles in the biology of ageing and age-related diseases, in general by cross-talking with other nutrient-sensing pathways and promoting longevity [29, 43]. Several authors have shown that increasing $NAD^+$ levels enhances the activity of sirtuins, which could constitute an additional anti-ageing pharmacological avenue in mammals. Additionally, intensive research is being carried out to identify other molecules that activate sirtuins [43].

- **Other pathways**. Mitochondrial respiration (and its production of reactive oxygen species or ROS), genome surveillance pathways (such as those involved in DNA repair or telomere maintenance), signals from the reproductive system or Wnt signalling have also been implicated in different ways in the ageing process [29, 31, 44].

These pathways seem to have a dual role depending on the environmental context that the organism is facing, behaving as **nutrient and stress sensors**. Under abundant nutrient availability and low stress (oxidative, temperature), they tend to promote growth and reproduction. While in contrast, under harsh conditions (such as those posed by dietary restriction), they favour cell protection and maintenance [21, 29]. It is worth mentioning that the responses of the different pathways to dietary restriction deeply depend on the characteristics of the diet and its timing [29]. This model also relates to the disposable soma theory, where more resources are allocated either to reproduction or somatic maintenance depending on the context [10, 11]. This is further mechanistically supported by experiments showing that decreased insulin/IGF-1 signalling (e.g. via *daf-2* mutation) produces the acquisition of germline characteristics (e.g. higher genomic stability) in *C. elegans* somatic cells [45]. Even though this model is a clear oversimplification, it becomes useful when thinking about the way in which the ageing process might have evolved and how the same biological pathways can be repurposed to activate genetic programs with completely different goals.

There are many more **complexities associated with these pathways** that would require an entire thesis on its own. For example, the insulin/IGF-1 signalling pathway can work in a cell non-autonomous manner (i.e. the activity of the pathway in one tissue can affect lifespan by influencing cells in a different tissue), which could help to coordinate ageing rates in the organism, and the effects are many times tissue-specific [21, 29]. Additionally, the pathways can have different effects depending on the life stage of the animal (e.g. development, adulthood, etc.) [46]. Furthermore, cross-talk between the pathways has previously been reported [43, 47]. Therefore, the inner workings of these signalling pathways is still an area of intense research.

The discovery of genetic pathways that can dramatically extend the lifespan of model organisms has demonstrated that **the ageing process has a genetic basis and it is possible to alter its rate**. More importantly, the appearance of age-related disease seems to be delayed in many of these long-lived organisms [29, 48], suggesting that these interventions indeed reduce the rate of some the operating ageing mechanisms (Fig. 1.1).

### 1.1.3 Hallmarks of mammalian ageing

Most studies on the biology of mammalian ageing have been conducted in mice. Many genetic mutations in conserved pathways (mainly nutrient-sensing pathways) have been shown to significantly extend the lifespan of mice. Among them, those that affect growth hormone signalling (which in mammals in turn controls the secretion of IGF-1 by the liver and therefore the insulin/IGF-1 signalling pathway) produce the longest lifespan improvements (in the order of 40-60%) [30]. Even though this is a remarkable result, it is far off the lifespan extensions achieved with 'simpler' model organisms such as *C. elegans* (where extensions of almost 1000% have been achieved with a mutation in a single gene of the insulin/IGF-1 pathway; the equivalent of a human living up to $\approx$ 1200 years!) [49]. This highlights a trend where translating lifespan interventions discovered in worms and flies yields generally less spectacular results in mice and potentially in humans.

Evolution has been experimenting with lifespan extension for a long time. Consequently, some species of mammals, such as the naked mole rat (*Heterocephalus glaber*) or some species of bats (Chiroptera), are exceptionally long-lived for their body size. Recent reports point towards the possibility that these species do not increase their mortality rate with age (i.e. they may have negligible senescence) [50, 51], which makes them incredibly interesting systems to study the biology of ageing in mammals.

In 2013, López-Otín *et al.* reviewed the main common denominators of the ageing process across organisms [3]. They defined several **hallmarks of ageing**, which can be understood as the measurable consequences of the ageing mechanisms that I proposed in Fig 1.1. I will briefly discuss some of them, with a special focus on those that directly affect the genome during mammalian ageing [3, 30]:

- **Genomic instability**. Somatic DNA mutations (single nucleotide variants, copy number changes, structural rearrangements, etc.) accumulate over time in mammalian cells (both in the nuclear genome and the mitochondrial genome) [52, 53]. Different mutational processes (good candidates for ageing mechanisms) create specific patterns of mutations (a.k.a. mutational signatures) in the genome, which have been widely studied in the context of human cancer [54]. It is possible to assign specific endogenous (e.g. DNA replication errors) and exogenous factors (e.g. smoke exposure) that contribute to the different processes. In the context of ageing, deamination of 5-methylcytosine (5mC) in a CpG context leads to C>T (cytosine to thymine) mutations, which accumulate in a clock-like manner with a rate that correlates with the proliferative activity of the tissue [55]. Furthermore, nuclear architecture and the 3-dimensional organisation

of the genome both seem to change with age, which can distort nuclear homeostasis. Interestingly, several human diseases that are considered to display premature ageing, such as Werner syndrome or Hutchinson–Gilford progeria, have mutations in proteins that lead to genomic instability [56]. Finally, it is possible that an increase in the mobilisation of transposable elements with age further contributes to destabilise the genome [57].

- **Telomere attrition**. The repetitive DNA sequences at the linear ends of mammalian chromosomes are capped with a protein complex (shelterin) to form structures known as telomeres. Due to the nature of the standard DNA replication machinery, the chromosomal DNA ends of somatic cells are eroded after each cell division (net loss of 100-200 bp of telomeric sequence per cell division). After a certain number of doublings (and therefore telomeres shortening) cells stop diving and they induce cellular senescence (see below) or cell death (apoptosis) [58]. For many years, this replicative limit, known as the Hayflick limit, was understood as the manifestation of the ageing process at the cellular level [59, 60]. Telomere shortening has indeed been shown to occur with age in most humans tissues [61]. Importantly, stem cells and germ cells express telomerase, an enzymatic complex that synthesises new telomeric repeats, avoiding telomere shortening. This way organisms can regenerate their tissues if needed, which makes it unlikely that telomere attrition is the only mechanism behind ageing. In addition to telomere length shortening, other mechanisms may contribute to replicative senescence in mammals [58]. Nevertheless, telomere biology plays a critical role in many fundamental processes, such as DNA repair and genomic stability, and non-telomeric functions for telomerase have also been suggested (such as global chromatin regulation and transcription of developmentally-regulated genes) [58]. As such, telomeres have been implicated in age-related diseases, such as cancer and cardiovascular disease [58, 61]. Interestingly, ectopic expression of the catalytic subunit of telomerase (TERT) extends the lifespan of mice that are cancer-resistant [62].

- **Cellular senescence**. Cellular senescence is a cellular state characterised by a stable cell cycle arrest. There are different types of senescence induced by different stress stimuli, including telomere shortening (replicative senescence, previously mentioned), sustained DNA damage (e.g. via irradiation) or derepression of the *INK4/ARF* locus (which encodes three tumour suppressor genes)[3, 63]. Under normal circumstances, cellular senescence carries out physiological functions such as preventing pre-malignant cells from dividing, participating in wound healing and tissue remod-

elling. Furthermore, senescent cells also secrete a cocktail of factors (termed the senescence-associated secretory phenotype, or SASP) with pleiotropic effects (pro-inflammatory, matrix remodelling, inducing growth, etc.) [63]. Senescent cells accumulate in mammalian tissues during ageing. If this happens in excess, the SASP can perturb the homeostasis of the tissue. Consequently, the removal of senescent cells in mice increases lifespan and reduces the appearance of age-related phenotypes [64–66]. Drugs that selectively induce apoptosis in senescent cells (known as senolytics) [67] are currently undergoing clinical trials in humans.

- **Epigenetic alterations**. This hallmark is reviewed in further detail in section 1.2.3, since it is the main focus of this thesis.

- **Other hallmarks of ageing**. These include loss of proteostasis (appropriate quality control of the proteome, which is mechanistically connected with autophagy pathways; strongly implicated in neurodegenerative diseases); deregulated nutrient sensing (mediated by the pathways discussed in section 1.1.2); mitochondrial dysfunction (including a reduction in the efficacy of the respiratory chain with age); stem cell exhaustion (which is thought to contribute to the decline of regenerative potential of the tissues with ageing, such as in the case of the haematopoietic system) and altered intercellular communication (including an increase in inflammation, known as inflammageing, or alterations in the neuroendocrine system) [3].

Importantly, complex interactions and interdependencies emerge between the different hallmarks of ageing. For example, in senescent cells in the mouse, a type of transposable element (LINE-1) becomes derepressed and activates type-I interferon response, which in turn causes inflammageing [68]. Furthermore, understanding the role of the environment in modulating the ageing process and the different hallmarks in mammals is becoming increasingly important.

Assuming that molecular damage is the main cause of biological ageing, the mechanisms that lead to genomic instability, telomere attrition, epigenetic alterations and loss of proteostasis are very likely the main drivers of the ageing process, with the rest of the hallmarks being a consequence of them [3]. Nevertheless, interventions targeting some of the more 'integrative hallmarks' (such as removing senescent cells, optimising dietary restriction or stem cell therapies) will probably arrive earlier in the clinic.

### 1.1.4 Studying the ageing process in humans

Average human lifespan has nearly doubled in most developed countries during the last 200 years. This has been the consequence of external factors, such as improvements in quality of water, nutrition, hygiene, housing and lifestyle, immunisation against infectious disease, antibiotics and medical care [69]. One of the most debated questions in the human ageing field is whether there is a limit to maximal human lifespan [70]. Since Benjamin Gompertz's pioneering work in 1825, it is known that the mortality rate in humans increases exponentially with age [71]. However, a recent study on Italian centenarians suggests that mortality rate, which increases exponentially up to about age 80, decelerates thereafter and reaches or closely approaches a plateau after age 105 [72]. This implies that human lifespan may continue to increase in the next decades and that **we have probably not reached our lifespan evolutionary limit as a species yet** [72, 73].

Thus, in order to avoid a massive socioeconomic burden on our societies [74], biomedical research should focus on **extending human healthspan** (i.e. the amount of time that we live free of disease) and not only lifespan. This goal, known as the 'compression of morbidity' [69], is theoretically possible if we target the core mechanisms that drive the ageing process (Fig 1.1); which is assumed to be the biggest contributor to the development of most age-related diseases, such as cancer, diabetes, cardiovascular disorders and neurodegenerative diseases [3]. Indeed, genetic and pharmacological interventions that increase lifespan in model organisms also seem to extend healthspan [75] and the compression of morbidity is a characteristic of human centenarians [76].

Most of our understanding of human ageing comes from studies carried out in **population cohorts**. Furthermore, during the last years different datasets of high-throughput molecular data (broadly know as 'omics') have been generated for many of these cohorts, including genetic data, epigenetic data, metabolomic data, imaging data or even the microbiome. These data (sometimes referred as 'deep phenotypes') complement the more traditional phenotypic measurements and health records and allow, for the first time, characterising the human ageing process with unprecedented resolution and scale. An example of such a cohort is the UK Biobank, which has enrolled > 500,000 participants [77]. Importantly, there is a trend in many of these cohorts to collect more longitudinal data (i.e. data over time for the same individual), which will likely increase the power to discover causal ageing mechanisms (as opposed to cross-sectional data, when data from different individuals at different ages is used) [78]. As Nobel laureate Sydney Brenner (known for establishing *C. elegans* as a model organism) remarked 10 years ago: "We don't have to search for a model organism anymore.

Because we are the model organisms" [79] (as a disclosure, I still believe we need model organisms to gain definitive mechanistic insights, and probably so did the late Prof. Brenner).

**The ageing process is an extremely polygenic trait**, probably one of the most complex phenotypes to be studied (as one would expect if it is composed of many different molecular processes). Candidate gene studies (with biased hypotheses) and genome-wide association studies (GWAS, unbiased) have found genetic variants that may affect the rate of ageing in humans. Many of them are associated with the function of genes that are part of nutrient-sensing pathways (such as *FOXO3* or *IGF1R*), that increase the risk of Alzheimer's (such as *APOE*), that are involved in cellular senescence (such as *CDKN2A*) or that are related to the immune system and inflammation (such as *HLA-DQA1*, *HLA-DRB1* or *IL6*) [30, 69]. Additionally, biological sex has a major impact on the ageing process and the incidence of age-related diseases. In the case of humans, females consistently live longer than males (females make around 90% of the supercentenarians i.e. individuals that live 110 years or more). However, they also seem to suffer greater morbidity in later life, which is known as the 'mortality-morbidity paradox' [80].

It is clear that human longevity has a genetic component. However, the latest estimates of heritability are quite low (ranging between 10-15%) [81, 82]. Furthermore, GWAS have yielded relatively few genetic variants compared with other complex phenotypes [30]. This could be due to the sample sizes required or methodological limitations (such as the way that the ageing phenotype is defined). Nevertheless, it is more likely that **the environment (and its interaction with the genetic background) accounts for most of the phenotypic variation in human ageing populations** [30, 69]. As such, there is evidence that diet (not only the content but also the timing) and exercise can act through nutrient-sensing pathways to regulate human healthspan and potentially lifespan [30, 69, 83–86]. Interestingly, social relationships are hypothesised to have a causal role in mortality rate, with lower levels of social integration associated with higher levels of inflammation, blood pressure or waist circumference across all human lifespan [87]. A fascinating example of the impact of environmental and lifestyle factors on human lifespan are the so called 'blue zones'. These are geographical areas (such as Ogliastra in Sardinia, Okinawa in Japan, the Nicoya peninsula in Costa Rica and the island of Ikaria in Greece) that have unusual proportions of long-lived individuals. However, the genetics of these populations are similar to their neighbours and therefore differences in the rate of ageing must be attributed to environmental and lifestyle factors [69, 88]. As such, targeted lifestyle interventions will likely complement pharmacological interventions (some of them mentioned in section 1.1.2) in order to slow down the human ageing process. Finally, epigenetic mechanisms constitute an interesting

layer of biological information that could mediate the interactions between genetics and environment to affect the ageing process and it will be the topic of discussion in the next sections.

## 1.2 Epigenetics of ageing

### 1.2.1 A brief introduction to epigenetics

The **coining of the term epigenetics** is normally attributed to Conrad H. Waddington, when in 1942 he defined it as the studies that deal with the causal mechanisms behind embryonic development (i.e. the processes by which the genotype of a single cell brings about the phenotype of an organism) [89]. This led to the unification of two apparently distinct fields (genetics and embryology), today known as the field of developmental genetics [90]. Furthermore, Waddington is also known for introducing the concept of the *epigenetic landscape*, which depicted developmental trajectories and the theory behind them in an incredibly compelling way [91]. Later work by Nanney, Riggs, Holliday and others evolved the definition of epigenetics towards the concept of *cellular memory*, that was materialised at the molecular level through DNA methylation (since it could affect transcription and be inherited after each cell division) [92]. The next decades were characterised by the discovery of a great variety of 'molecular routes' to affect gene expression (such as chromatin modifications or non-coding RNAs), which in humans culminated with consortia such as ENCODE [93] or Roadmap Epigenomics [94] and created a broader concept of epigenetics [92, 95].

Nowadays there is a debate in the scientific community about the appropriate definition of epigenetics [95, 96]. For the purpose of this thesis I will define epigenetics as **the study of molecular variation that is beyond changes in the DNA sequence, that is inherited after cell division and that regulates gene expression** (in line with the definition by Wu and Morris) [97]. However, it is important to mention that, in the context of the epigenetic clock, we are still not sure whether these molecular changes have direct functional consequences (e.g. by affecting RNA expression) and/or whether they help to define a new metastable cellular state in the cells that they occur (see section 1.3.3).

There are different types of molecular mechanisms that are normally considered 'epigenetic'. These include:

- **DNA methylation**. This will be discussed in detail in section 1.2.2.

- **Histone modifications**. The basic unit of chromatin is the nucleosome. It is composed of ∼147 bp of DNA wrapped around an octamer of histones (generally two copies of each one of the four core histones: H2A, H2B, H3 and H4; although histone variants such as H3.3 or H2A.Z have also been characterised). In order to fit around 2 meters of DNA into the nucleus of a human cell, chromatin needs to be further compacted with the help of scaffold proteins (with the furthest level of compaction achieved in the mitotic chromosome) [98]. Histones possess N-terminal regions (a.k.a histone tails) that project towards the outside of the nucleosome and are positively charged. By default, this helps to compact the chromatin by interacting with the negative charges of the DNA. However, many different types of post-translational modifications (acetylation, methylation, phosphorylation, ubiquitinylation, sumoylation, etc.) in the residues of the histone tails have been identified across the eukaryotic tree of life (although modifications have been also found in the globular domains) [99]. These histone modifications can affect the chemical properties of chromatin, its degree of compaction and ultimately contribute to the regulation of transcription (e.g. through the recruitment of downstream effector proteins). The sequence and combinations of these modifications that modulate chromatin activity was named the *histone code* [100] and its complexity is slowly being characterised thanks to technologies such as ChIP-seq [93, 94]. Finally, it is worth mentioning the nomenclature that is used to refer to histone modifications. For instance, for the histone modification 'H3K36me3', the information about the histone ('H3'), the residue where the modification happens ('K36' is lysine 36) and the type and number of modification(s) ('me3' refers to three methyl groups) is provided.

- **Other 'epigenetic' players**. Non-coding RNAs (such as long non-coding RNAs, PIWI-associated RNAs or short-interfering RNAs) have been shown to affect the epigenetic landscape through different mechanisms. Additionally, many RNA modifications (known as the *epitranscriptome*), are currently being elucidated. However, whether they are considered truly 'epigenetic' is debatable [101, 102]. Furthermore, prions (misfolded proteins that accumulate in cells and act as templates to further misfold more protein molecules) have been proposed as an epigenetic mechanism that is not based on heritable changes in nucleic acid [103].

The different epigenetic marks present complex patterns of correlation and cross-talk, which are mechanistically linked to the way that its addition and removal is regulated. This helps to define **chromatin states** (i.e. combination of different epigenetic marks) that affect

gene regulation in different ways. Historically, chromatin has been broadly classified in two categories [104–106]:

- **Euchromatin**. It presents active gene activity and it is more accessible to the transcription machinery. It is generally characterised by histone modifications such as H4K16ac, H3K4me3 or H3K36me3.
- **Heterochromatin**. It is normally subdivided in constitutive (highly condensed and transcriptionally repressed; mostly found in pericentromeric regions, telomeres and other regions that contain repetitive elements; it is generally marked by H3K9me3 and high levels of 5mC) and facultative (normally transcriptionally silent but it has the potential to adopt open conformations depending on the temporal and spatial context; it is generally marked by H3K27me3).

Consortia that have mapped many epigenetic marks (collectively known as the epigenome) in humans [93, 94] and advances in chromatin segmentation algorithms [107] have led to more a fine-grained definition of chromatin states. This has helped to identify functional elements in the genome in a high-throughput way, such as active transcription start sites (TSS, enriched in H3K4me3), enhancers (enriched in H3K4me1) or bivalent chromatin (enriched in H3K4me3 and H3K27me3) [93, 94].

Epigenetic marks contribute to define (or in Waddingtonian terms, 'canalise') **different cell types and cellular states** from the same genomic sequence. Cellular identity is normally established by master regulators (initiators), generally transcription factors that activate the expression of a genetic program (i.e. coordinated gene expression) [105]. However, in order for this cellular state to survive once the initiator is no longer present, the patterns of epigenetic marks need to be inherited after cell division. This is clearly the case for 5-methylcytosine (5mC, see section 1.2.2). In the case of histone modifications, there is evidence for the propagation of some of the repressing histone modifications (such as H3K9me3 and H3K27me3). This is possible because the machinery in charge of catalysing the addition of these chemical modifications (i.e. the *writers*, SUV39H1 and Polycomb Repressive Complex 2) also have the ability to recognise it (i.e. they are also *readers*), therefore creating a positive feedback. However, it is not clear whether many other histone modifications are copied after DNA replication in the newly synthesised DNA strand and therefore whether they are truly epigenetic [105]. Additionally, it is important to mention that enzymatic activities to reverse most (if not all) epigenetic marks (i.e. *erasers*) have been identified [104].

Besides regulating transcription and/or defining cellular states, **epigenetic mechanisms play a fundamental role in other important biological processes**. These include genomic imprinting (monoallelic expression according to parental origin) [108], X-chromosome inactivation (silencing of one of the two X chromosomes in female therian mammals) [109] or cast differentiation in eusocial insects (such as queen and worker differentiation in honeybees, where there is a 10-fold difference in lifespan) [110, 111].

One of the big questions in the field of epigenetics is **to which extent epigenetic patterns are genetically programmed** and to which extent they change in response to environmental/stochastic influences. In the case of human populations, genetic variants that affect the levels of DNA methylation (meQTLs) and histone modifications (hQTLs) at specific loci have been identified [112]. Interestingly, it is possible to predict different epigenetic marks from the raw DNA sequence, mainly by identifying transcription factor binding sites that guide different parts of the epigenetic machinery [113]. Furthermore, monozygotic twins allow to control for the genetic background and study the epigenetic variation derived from environmental and stochastic factors, which is particularly interesting in the context of complex diseases [114]. Nevertheless, the debate is far from being finished.

### 1.2.2 Fundamentals of DNA methylation in mammals

Different types of DNA modifications have been described across the tree of life. DNA methylation enzymes evolved in bacterial species to protect them from the infection of bacteriophages, although roles in bacterial transcriptional regulation have also been described [115]. In mammals, the most common DNA modification is the **addition of a methyl group in the carbon at the $5^{th}$ position of cytosines (5mC)**, which has been called the $5^{th}$ base of DNA. The traditional functions assigned to 5mC include the mediation of genomic imprinting and X-chromosome inactivation, repressing transposable elements and regulating transcription [116]. In the latter case, 5mC has been commonly associated with the repression of transcription (e.g. by altering the ability of transcription factors to bind or by attracting methyl-CpG binding domain proteins) [117]. However, it is becoming clearer over time that the picture is more complex. For example, gene bodies of highly expressed genes are methylated in order to avoid cryptic transcription [118].

5mC generally happens when the cytosine is followed by a guanine in the DNA strand (commonly known as a CG dinucleotide or CpG site) [117, 119]. In the human genome there are around 28 million CpG sites, of which approximately 60-80% are normally methylated [119]. The density of CpG sites in the genome is variable. **CpG islands** (CGIs) are CpG-

enriched genomic regions (200-2000 bp long, ∼30,000 CGIs in the human genome which account for ∼10% CpG sites) and are frequently associated with promoters (although ∼9,000 of them are found inside gene bodies) [119–121]. Promoter-associated CGIs are normally unmethylated across cell types, which contrasts with the high methylation levels in the rest of the genome. The mechanism by which these CGIs remain resistant to DNA methylation is starting to be elucidated. Recent reports suggest that active transcription together with the binding of proteins that block methylation are required for the resistance. Among these proteins (which bind non-methylated CpG sites in the CGI via their zinc-finger CXXC domain) it is worth mentioning CFP1 (which recruits an H3K4 methyltransferase that increases H3K4me3 levels, which in turns inhibits *de novo* methylation) and TET1 (see below) [122]. On the contrary, if a promoter-associated CGI is methylated, this commonly leads to transcriptional repression of the correspondent gene; something that is observed in the promoters of certain tumour suppressor genes in cancer [123].

Different **enzymes contribute to the establishment, maintenance and removal of DNA modifications** in mammals. *De novo* methyltransferases DNMT3A and DNMT3B are capable of catalysing the addition of 5mC in those CpG sites that originally lack the modification in any of the two DNA strands. Maintenance methyltransferase DNMT1 is able to add 5mC to hemimethylated DNA (i.e. when only one of the strands in the CpG site has 5mC) thanks to the symmetry of CpG sites (and its recruitment via UHRF1). This provides a mechanism for the inheritance of DNA methylation patterns after cell division, therefore making it a true epigenetic mark capable of generating cellular memory (Fig. 1.3) [117, 119]; as originally hypothesised in 1975 by Holliday, Pugh and Riggs [124, 125]. It is worth mentioning that 5mC in a non-CpG context (i.e. in CHG or CHH, where H corresponds to adenine, thymine or cytosine) has also been detected in human tissues [126]. However, its abundance is generally very low with the exception of embryonic stem cells (ESCs), induced pluripotent stem cells (iPSCs) and some brain cells; probably due to the levels of DNMT3A and/or DNMT3B in these cell types [127, 128]. A third *de novo* DNA methyltransferase that lacks the N-terminal catalytic domain, DNMT3L, has also been identified in mammals. DNMT3L cooperates mainly with DNMT3A to add 5mC in maternal genomic imprints during gametogenesis [129, 130].

It was a long-standing question whether the loss of 5mC (a.k.a demethylation) can only occur by replication-coupled passive loss (i.e. preventing DNMT1 maintenance activity and diluting 5mC content by cell division), due to methyltransferase errors or as a result of DNA repair after DNA damage [131]. In 2009, two groups conclusively identified the presence of a different type of DNA modification in mouse and human DNA, 5-hydroxymethylcytosine

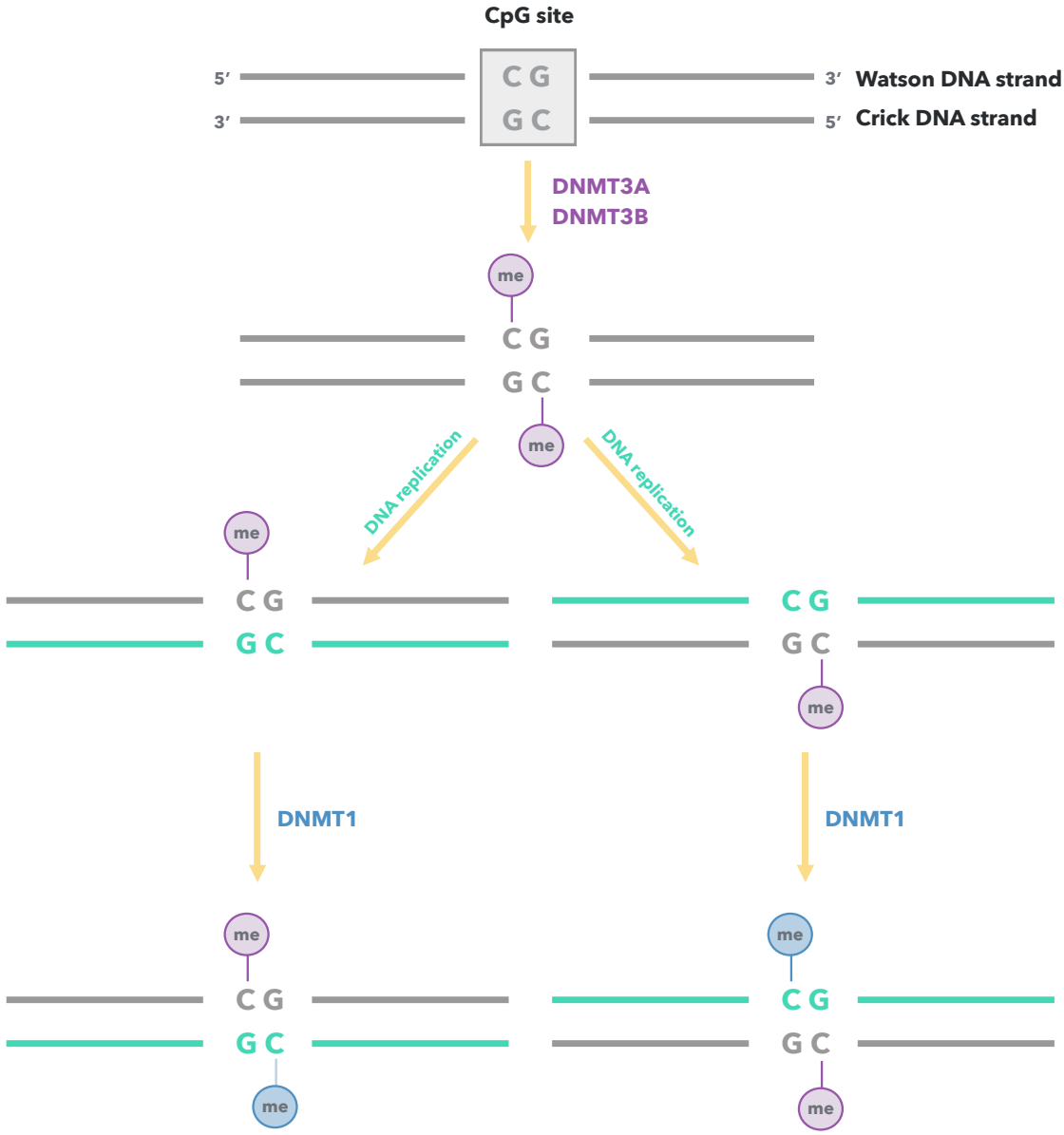



**Fig. 1.3** Establishment and maintenance of 5-methylcytosine (5mC) in mammalian genomes. Unmethylated cytosines in symmetric CpG sites are originally methylated *de novo* by DNA methyltransferases DNMT3A and DNMT3B to form 5mC. After cell division, the newly synthesised DNA strands lack the methylation mark. Maintenance DNA methyltransferase DNMT1 recognises this hemimethylated DNA and adds the missing methyl groups, therefore ensuring the inheritance of DNA methylation patterns and cellular memory.

(5hmC) [132, 133] (although, surprisingly, its presence in rat tissue had been detected almost 40 years before) [134]. Furthermore, one of them demonstrated that the enzyme TET1 is capable of oxidising 5mC to 5hmC [133]. Since then, other enzymes from the TET family (TET2, TET3) have also been shown to catalyse this reaction [135]. Further products of oxidation, 5-formylcytosine (5fC) and 5-carboxylcytosine (5caC), can also be generated by TET enzymes, although their abundance is incredibly low in the genome. Replication-dependent dilution of oxidised products or thymine DNA glycosylase (TDG)-mediated excision of 5fC and 5caC coupled with BER have been shown to complete the demethylation process (Fig. 1.4). Altogether, this shows that **active enzymatic DNA demethylation is a feature of mammalian epigenomes** [116]. Finally, it is worth mentioning that another type of DNA modification, $N^6$-methyladenine, has recently been identified in both mouse and human cells, thus further expanding the DNA alphabet [136, 137].

**DNA methylation patterns change drastically during mammalian embryonic development**. After fertilisation, mouse and human zygotes undergo epigenetic reprogramming in order to reset naive pluripotency. This is mainly characterised by a global loss of 5mC (i.e DNA hypomethylation) with different demethylation processes affecting the paternal and maternal genome. Nevertheless, some genomic regions, such as imprints, survive epigenetic reprogramming. *De novo* DNA methylation occurs after implantation of the blastocyst, which will restore DNA methylation levels for most somatic cells and eventually generate cell-type specific DNA methylation patterns [131, 138, 139].

In the case of the cells that will give rise to the germline (**primordial germ cells** or PGCs), further genome-wide DNA demethylation occurs, which makes PGCs the most hypomethylated cells found during mammalian development thus far (global methylation levels are $\sim$4%). This ensures imprint erasure and that parental epigenetic memories are removed, therefore posing a barrier for transgenerational epigenetic inheritance. Nevertheless, some regions that escape epigenetic reprogramming in mice and humans have been described (mainly evolutionarily young and potentially hazardous retrotransposons). Methylation patterns of the germline will then be re-established in a sex-specific manner [138].

Over the years, many **technologies have been developed to measure 5mC and its oxidative products** (see section 4.1 for an overview). Many assays rely on a chemical procedure called bisulfite conversion [140]. Genomic DNA is denatured and incubated with sodium bisulfite. This leaves 5mC residues intact, but unmethylated cytosines are deaminated and converted to uracil. Therefore, after PCR amplification, 5mCs are substituted by cytosines while unmethylated cytosines become thymines. This information can then be read at base-pair resolution through DNA sequencing or by hybridisation to a methylation

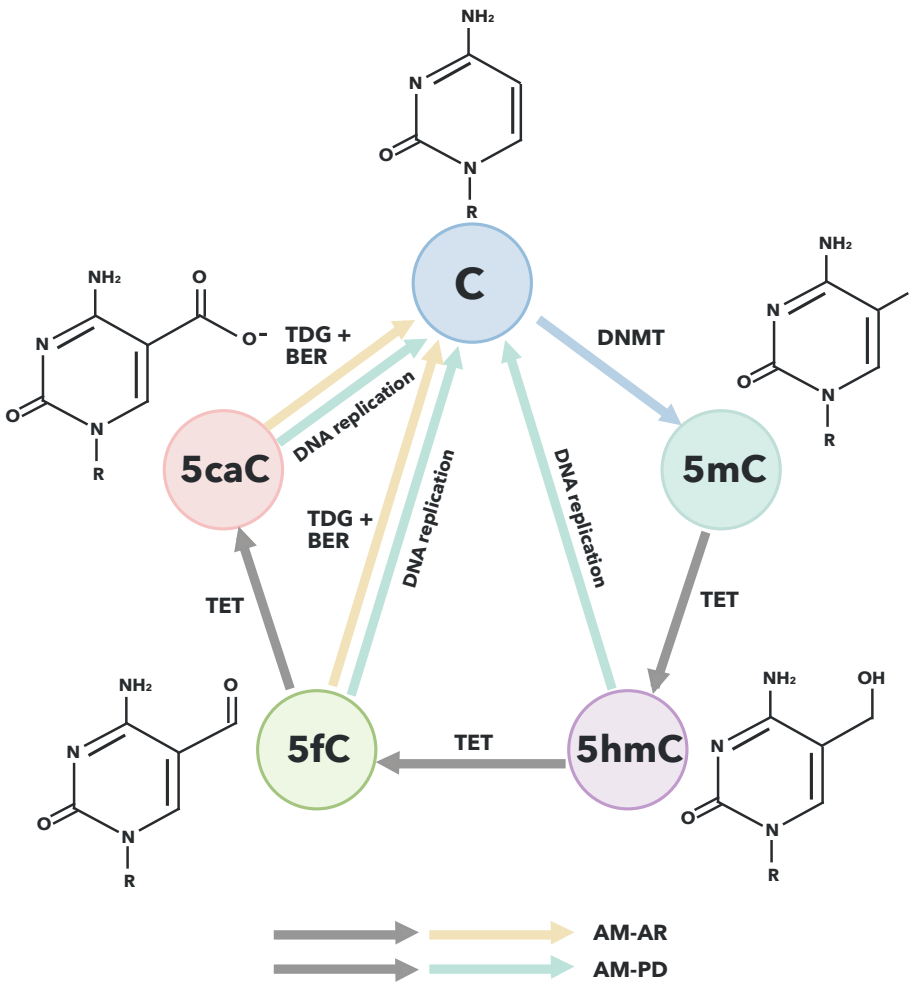



**Fig. 1.4** Oxidation of 5-methylcytosine (5mC) and the cycle of demethylation. 5mC can be oxidised to different DNA modifications (5hmC, 5fC, 5caC) by TET enzymes. The maintenance DNA methyltransferase DNMT1 can only recognise 5mC. As a consequence, after DNA replication, the rest of the modifications would be eventually lost (active modification–passive dilution or AM-PD). Alternatively, thymine DNA glycosylase (TDG)-mediated excision of 5fC and 5caC coupled with base excision repair (BER) can lead to the same outcome (active modification–active removal or AM–AR). This figure was adapted from [116].

array (such as the Illumina Infinium BeadChips, which are the platform used to generate the data analysed in this thesis) [141]. It is important to keep in mind that 5hmC is also resistant to bisulfite conversion, and therefore it is confounded with the 5mC signal [116]. Furthermore, C>T mutations (a common mutation during ageing, see section 1.1.3) can be confounded with hypomethylation events. Another caveat of bisulfite treatment is that it degrades DNA to a great degree and generates sequencing libraries of low complexity, which leads to reduced mapping rates and higher costs. A few months ago, Liu *et al.* published a bisulfite-free protocol for 5mC sequencing at base-pair resolution, which could potentially solve some of these issues and start a new generation of bisulfite-free methods [142].

Exposure to certain **environmental factors is associated with changes in the methylome** that can potentially modulate disease risk. In mice, *in utero* undernourishment leads to weight and metabolic defects in the $F_1$ offspring. Furthermore, this metabolic phenotype is inherited in the $F_2$ offspring through the paternal line. Interestingly, this could be caused because genomic regions that become hypomethylated during paternal germline specification survive epigenetic reprogramming in the $F_2$ zygote and lead to further chromatin alterations in adult tissues [143]. In a different example, smoking exposure in humans changes DNA methylation patterns of blood [144] or buccal cells [145] in a consistent and reproducible manner. However, the mechanisms behind these changes and whether they are functional or mere passenger epimutations remain obscure. Finally, many complex diseases, such as rheumatoid arthritis or many cancer types, are characterised by altered DNA methylation patterns, which suggests that epigenetic mechanisms integrate genetic and environmental aetiologies of disease [146, 147].

### 1.2.3 Links between the epigenetic machinery and ageing

As previously mentioned in section 1.1.3, **epigenetic alterations are one of the hallmarks of mammalian ageing** [3]. Given the role of genetic pathways and environmental factors in the regulation of organismal lifespan, the epigenetic layer of biological information has attracted a lot of interest in the ageing field (to the point that some authors have suggested that it is the hub that connects all the hallmarks of ageing) [148]. Indeed, many life-extending interventions (such as dietary restriction, exercise or a robust circadian rhythm) modulate the epigenetic machinery and induce chromatin changes [149]. Furthermore, since many epigenetic marks are stable over time, they could behave as cellular memories that store past environmental exposures. Taking into account the vast literature available on this topic, in this section I will try to extract the pieces of information that are more relevant for this work.

Several authors have reviewed the wide variety of chromatin changes that occur during ageing in different organisms [148–151]. These include changes in histone numbers, histone variants, histone modifications, DNA modifications, non-coding RNAs or nucleosome positioning; which eventually lead to transcriptional deregulation. Certain **mutations in proteins of the epigenetic machinery affect the lifespan of organisms from yeast to mouse**, thus proving a causal role of some of these changes in the ageing process. Furthermore, I highlight a few interesting insights from these studies:

- Global heterochromatin loss and redistribution has been suggested as one of the mechanisms behind the ageing process [152, 153]. Indeed, mutations in proteins that cause premature ageing in humans (such as nuclear lamins or the WRN helicase) have a major impact in heterochromatin structure and the genomic distribution of its characteristic repressive chromatin marks (such as H3K9me3 or H3K27me3) [154]. Cellular senescence is also associated with remodelling of heterochromatin [155]. Furthermore, heterochromatin deregulation can lead to the activation and mobilisation of transposable elements during mammalian ageing [156].

- Mutations that alter the levels of H3K27me3 and H3K4me3 can have contradictory effects in the lifespan of model organisms, probably depending on the loci and cell types that they affect. However, it appears that mutations that increase the levels of H3K36me3 consistently increase lifespan (at least in yeast, worm and fly) [148–151]. This will be of interest for Chapter 3.

- Increased levels of SIRT6, an H3K9ac and H3K56ac histone deacetylase from the sirtuin family, can extend lifespan of male mice [157]. On the contrary SIRT6-deficient mice die at about 4 weeks, have a progeroid phenotype and have increased genomic instability (due to problems in the base excision repair pathway) [158]. The role of SIRT6 in human ageing is still not clear.

- Histone chaperone ASF1, which promotes histone deposition and stability, is required for normal replicative lifespan in yeast [159]. Intriguingly, the mouse ortholog ASF1A is important to resolve bivalent chromatin upon differentiation of embryonic stem cells [160] (see discussion regarding the importance of bivalent domains during mammalian ageing in the ‘Hypermethylated regions’ section below).

- The naked mole rat, an incredibly long-lived rodent with very low cancer incidence, presents a stable epigenome that is resistant to *in vitro* reprogramming. Furthermore, higher levels of repressive chromatin marks (such as H3K27me3) are observed relative to the mouse [161].

Importantly, the DNA methylation landscape also seems to be affected during ageing in mammals. Certain CpG sites or genomic regions gain methylation with age (i.e. they become hypermethylated) while others sites lose methylation (i.e. they become hypomethylated). Furthermore, some of these age-associated methylation changes are shared across tissues, while others are tissue-specific. Notably, even though they have an stochastic component, the genomic context where these changes occur seems to be conserved in mice [162–166] and humans [167–179]:

- **Hypermethylated regions**. They are generally enriched for bivalent chromatin, regions repressed by PRC2 (Polycomb Repressing Complex 2) and CpG islands (CGIs, many of which overlap with bivalent promoters). Bivalent domains are populated with numerous transcription factor binding sites and are marked simultaneously by histone marks H3K27me3 (established by EZH2, which is part of the PRC2 complex; associated with transcriptional repression) and H3K4me3 (established by Trithorax-group proteins; associated with transcriptional activation). The two histone marks seem to co-occur on the same loci of the same cell in a majority of the bivalent domains (as opposed to an heterogeneous population of cells with different histone marks), and sometimes even on the different histone copies of the same nucleosome [180]. This opposing duality is thought to silence developmental genes in embryonic stem cells (and pluripotent stem cells in the embryo) while keeping them poised for activation (by developmental and/or environmental cues) [180]. Developmental genes (many of them lowly expressed transcription factors) are indeed highly enriched in these regions and this seems to be a feature of most gene ontology analysis performed in hypermethylated CpGs during ageing. Many of the bivalent domains disappear after differentiation, leaving only one of the two marks [181], but specific nonpluripotent bivalent domains can also be generated after differentiation [180]. Besides differentiation, the physiological ageing process also seems to change the landscape of bivalent domains, as observed in aged haematopoietic stem cells or HSCs (where around 335 bivalent domains disappear in old mouse HSCs, whereas 1,245 emerge) [182]. This process is apparently linked to the proliferative history of HSCs [183] and could contribute to the myeloid skewing observed during ageing [182, 183]. Interestingly, bivalent domain loses occur in cancer cells as well, which seems to correlate with the hypermethylation of the regions [184]. It is possible that the ageing- or cancer-related hypermethylation destroys the ability to create a bivalent equilibrium in these regions. If this happens in the stem cells, it could impair adequate differentiation and propagate the methylation change in the tissue. Overall, this provides an interesting mechanistic

link between embryonic development, lineage-specific cellular identity and the ageing process that should be further explored.

- **Hypomethylated regions**. They are generally enriched for tissue-specific enhancers (generally marked with H3K4me1) and depleted for CGIs (which makes sense, given the low methylation levels of CGIs). For example, in wild-type mouse liver, 8230 liver-specific enhancers are hypomethylated during ageing. On the contrary, only 4702 of those enhancers suffer the same fate in Ames dwarf mice (which have decreased insulin/IGF-1 signalling and a longer lifespan), which highlights that the epigenome from Ames dwarf mice appears more stable [165].

Importantly, some of these ageing-associated DNA methylation changes also seem to happen in dogs and wolves [185]. Furthermore, the rate of change of many of these age-associated regions is negatively correlated with lifespan in six different mammals [186]. Altogether, this suggests that conserved epigenetic mechanisms may operate during ageing to shape the mammalian methylome.

Hence, it is clear that the epigenome is eroded over time. In humans, this inter-individual divergence of DNA methylation patterns created upon ageing has been termed 'epigenetic drift' [187]. Interestingly, this phenomenon is found even in monozygotic twins [188, 189], again highlighting the role of environmental and stochastic factors. This is also observed at the single cell level, where cells from old organisms become more heterogenous at the epigenomic and transcriptomic level [190, 191].

## 1.3 The epigenetic ageing clock

### 1.3.1 Measuring the ageing process

In order to study any phenomena one needs to be able to measure it. Using **survival curves** (a.k.a lifespan curves, i.e. plotting the survival fraction over time, see equation 1.1) we have been able to quantify the ageing process at a population level (where the assumption is that life extension in a significant proportion of the population is a surrogate marker of slowed ageing) [26]. The adoption of this methodology in model organisms (that we can manipulate genetically and/or pharmacologically) triggered the discovery of the first genes impacting upon the ageing process (i.e. the mutants showed 'shifts' of the survival curve when compared with a control). Since then, this has been the main paradigm in ageing research, with efforts being made to automate the process and increase its throughput [192].

Nevertheless, measuring the ageing process at the organismal level has proven more difficult. Due to environmental and stochastic factors, there are significant differences in the lifespan of even isogenic organisms. Therefore, there is a real need to develop accurate **biomarkers of ageing** i.e. measurements of 'age-related change(s) in body function(s) or composition that can predict the future onset of age-related disease(s) and/or the residual lifetime left (i.e. predict the rate of ageing) more accurately than chronological age' [193]. Furthermore, according to the American Federation of Aging Research, any valid biomarker of ageing must also monitor a basic (sub)process underlying ageing, it must be able to be tested repeatedly without harming the organism (i.e. it has the potential to become a longitudinal biomarker) and be reproducible in both humans and laboratory animals (such as mice) [193].

The derivation of a biomarker of ageing leads to the definition of two types of age:

- **Chronological age**. It is the time elapsed since the birth of an individual.
- **Biological age**. It is the result derived from a specific biomarker. Each biomarker is trained using a set of biological parameters (independent variables) to predict a dependent variable (e.g. chronological age) that captures the probability of dying at a given time. The training takes place using several individuals, ideally from multiple populations. Afterwards, given a new individual and the biological parameters, the biological age can be predicted (and it should capture the risk of death more accurately than chronological age). Younger biological ages should be linked to high fitness and health whereas older biological ages should correlate with age-related disease onset and morbidity [149]. For example, if chronological age is used as the dependent variable (which is the case for most biomarkers), the biological age of an individual represents the chronological age of the average population that is most similar to the individual (according to the set of biological parameters). In this case, if the biological age of an individual is smaller than his chronological age, this could be interpreted as his probability of death being smaller than the probability of death for the average population (i.e. potentially the rate of ageing of the individual is slower than the average).

In the case of humans, the initial ageing biomarkers included traditional biological parameters such as body mass index, waist and hip circumference, blood pressure or heart rate. Over the years, biomarkers that use molecular parameters have also been developed; these include clinical chemistry parameters (such as cholesterol, immunoglobulins or fasting glucose), telomere length or 'omics'-based measurements [193, 194]. In the latter category,

almost every layer of biological information can be used to derive a biomarker, including epigenomics (see next section), transcriptomics [195], proteomics [196], metabolomics [197], microbiome [198] or even brain neuroimaging data [199]. Furthermore, composite biomarkers (that combine the biological parameters from molecular layers with measurements of physiological function) [200] and algorithmic innovations (such as deep neural networks) [201] will likely improve the predictions. The **biomarkers of the human ageing process** will serve as personalised risk indicators and will allow monitoring the response to interventions, therefore creating endpoints in clinical trials that target the ageing process.

### 1.3.2 The landscape of epigenetic clocks

**Epigenetic clocks** are mathematical models that predict the biological age of an organism using DNA methylation data. These models exploit the fact that DNA methylation patterns change robustly with age in different tissues and species, as summarised in section 1.2.3. Epigenetic clocks have emerged in the last few years as the most accurate biomarkers of the ageing process in humans, which they can track across the entire lifespan. As a quick comparison, telomere length (one of the other popular ageing biomarkers) achieves a Pearson's correlation coefficient with chronological age of $\sim$ -0.5 in blood leukocytes in the best case scenarios (with many studies reporting much lower values and contradictory results) [202]. On the other hand, the coefficients for Hannum's or Horvath's epigenetic clocks (discussed later) are generally above $\sim$ 0.8 (in virtually all studies assessed) [203].

The idea that DNA methylation patterns behave in a clock-like manner during cellular ageing was already proposed by Holliday and Pugh in 1975 [124]. With the advent of high-throughput DNA methylation technologies, some authors started to test the **ability of DNA methylation patterns to predict chronological age in humans**. In 2010, Bork *et al.* showed that DNA methylation values change at specific CpG sites upon long-term culture and between young and old individuals in mesenchymal stromal cells [204]. Later that same year, studies by Teschendorff [168], Rakyan [167], Gronninger [205] and others identified sets of CpG sites (signatures) that consistently altered their methylation states with age in different tissues and cell types (and interestingly some of them seemed to occur in the same genomic context). In 2011, Bocklandt *et al.* demonstrated that it was possible to predict chronological age in saliva with an average error of 5.2 years using the DNA methylation values of only two CpG sites [206]. Shortly afterwards, Koch *et al.* built what was probably the first multi-tissue predictor of chronological age in humans (which worked using the same 5 CpG sites across different cell types) [207].

The potential role of epigenetic clocks as biomarkers of human ageing was probably realised after the publications, in 2013, of the models by Hannum *et al.* [208] and Horvath (Table 1.1) [209]. Since then, these epigenetic clocks have being validated in a large number of independent cohorts and have become, *de facto*, the default human epigenetic clocks for blood and multi-tissue predictions respectively. Importantly, this inspired other groups to build epigenetic clocks in the mouse [164, 210–213], dogs and wolves [185] or even humpback whales [214]; which will be instrumental to broaden our understanding of the biology of ageing in mammals [210]. A comparison of some of these epigenetic clocks can be found in Table 1.1. The accuracy that they can achieve with a relatively small number of CpG sites as covariates is remarkable.

The predictions from epigenetic clocks are normally referred as epigenetic age (which is equivalent to the concept of biological age previously explained). Interestingly, deviations of epigenetic age from chronological age (a.k.a **epigenetic age acceleration** or EAA) have been associated with many conditions in humans, including time-to-death [203, 215], HIV infection [216], Down syndrome [217], obesity [218], menopause [219] and breast-cancer risk in women [220], Werner syndrome [221] or Huntington's disease [222], among others (reviewed in [223]). Interestingly, females and people of Hispanic ethnicity have lower EAA (after correcting for blood cell composition effects) when compared with males and those of Caucasian origin respectively, highlighting a role for biological sex and genetic background in the rate of the epigenetic ageing clock [224]. In mice, the epigenetic clock is slowed down by dwarfism and calorie restriction [164, 165, 211–213] and is accelerated by ovariectomy and high fat diet [164, 210–212].

Recently, **other epigenetic clocks** have been created for slightly different purposes. For example, Yang and colleagues developed an epigenetic clock that can track the rate of (stem) cell divisions in normal and cancerous tissue (see section 2.3.2) [225]. Furthermore, an epigenetic clock that performs well in skin cells (such as fibroblasts, buccal cells and endothelial cells; known as the skin-blood clock) was developed in order to improve *ex vivo* studies or forensic applications [226]. Moreover, this epigenetic clock enables the detection of EAA in Hutchinson-Gilford progeria, which is not possible with Horvath's clock [226]. Additionally, other epigenetic clocks have been trained to predict more complex dependent variables than chronological age. Levine *et al.* built a model that predicts a combination of chronological age with clinically-relevant variables (such as erythrocytes distribution width or serum glucose), known as *PhenoAge* [227]; while Lu *et al.* built a model that predicts a composite variable mixing information from smoking pack-years and plasma proteins (adrenomedullin, C-reactive protein, plasminogen activation inhibitor 1 and

| **Species** | Human | Human | Mouse | Dog and wolf |
|---|---|---|---|---|
| **Main reference** | Hannum *et al.* [208] | Horvath [209] | Stubbs *et al.* [210] | Thompson *et al.* [185] |
| **DNA methylation technology** | Illumina methylation array (450K) | Illumina methylation array (27K and 450K) | RRBS | RRBS |
| **N samples (in training set)** | $N = 482$ | $N = 3931$ | $N = 129$ | $N = 108$ |
| **Tissues (in training set)** | Blood | Multi-tissue (18) | Multi-tissue (7) | Blood |
| **Age range (in training set)** | 19-101 years | 0-100 years | 1-41 weeks | 0.5-8 years |
| **Number of CpGs in the final model** | 71 | 353 | 329 | 115 |
| **Median absolute error (MAE)** | 4.9 years | 3.6 years | 3.33 weeks | 0.8 years |
| $\frac{MAE}{max.\ age\ in\ model} \cdot 100$ | 4.85% | 3.6% | 8.12% | 10.0% |

**Table 1.1** Comparison of some of the epigenetic clocks available for different species. RRBS: reduced representation bisulfite sequencing (see Chapter 4).

growth differentiation factor 15), known as *GrimAge* [228]. These models perform better than previous epigenetic clocks in predicting the onset of several age-related diseases and therefore they will likely be useful in a clinical context.

From an statistical point of view, most of the epigenetic clocks have been built using linear regression (see section 2.4). A model needs to be trained to predict the dependent variable (normally chronological age) using the methylation values of different cytosines (generally in CpG context) as covariates. Given that the number of covariates is normally several orders of magnitude bigger than the number of samples available for training, regularisation (i.e. 'shrinking' of the linear regression coefficients, many of which become zero) needs to be performed. More specifically, elastic net (a combination of lasso and ridge regularisation) has been successfully applied [229]. **Many epigenetic clocks with similar performance can be built from different sets of CpG sites** (i.e. the construction of epigenetic clocks is highly statistically degenerate) [212]. Therefore, it is important to understand that the CpG sites that constitute an epigenetic clock are not necessarily the most important biologically, but rather they are probably a lower-dimensional representation of the main processes that shape the epigenome with age.

### 1.3.3 Molecular mechanisms of the epigenetic ageing clock

At this point it is probably useful to clarify a few concepts that I will refer to throughout this work. I define the **epigenetic ageing clock** as the biological mechanisms that give rise to the genome-wide epigenetic changes that occur during ageing (in a given species); a definition in line with the one reported in [223]. These changes have been widely studied in the context of DNA methylation and can be utilised to train predictors of chronological age (or other more complex variables). These predictors constitute different types of *epigenetic clocks*, and I will try to refer to them by the specific model being mentioned (e.g. Horvath's epigenetic clock, Hannum's epigenetic clock, etc.). As such, specific epigenetic clocks capture the changes associated with the underlying epigenetic ageing clock.

The molecular mechanisms that control the rate of the epigenetic ageing clock are still mysterious [223, 230]. Steve Horvath proposed that his multi-tissue epigenetic clock captures the workings of an **epigenetic maintenance system**, although the molecular nature of this hypothetical system is unknown to this date [209]. Furthermore, we still do not know whether these changes are functional at all or whether they are just downstream consequences of other molecular processes that drive ageing.

As mentioned in section 1.2.3, many studies have characterised changes in DNA methylation patterns during mammalian ageing, some of which seemed to be evolutionarily conserved [186, 209]. Interestingly, changes that involve a gain in methylation during ageing seem to be more conserved across tissues, whilst changes involving hypomethylation are generally more tissue-specific [209, 225]. Furthermore, many of these changes occur in **regions normally occupied by Polycomb Repressing Complex 2**, which are marked by the repressive histone mark H3K27me3. Therefore, it is likely that disruptions of H3K27me3 domains (which are generally inherited after cell division) play a role in epigenetic ageing. A specific instance would be bivalent promoters (which are marked by both H3K27me3 and H3K4me3); these tend to gain methylation with age (see section 1.2.3). These signals are captured by most epigenetic clocks trained to predict chronological age [223].

The mere existence of multi-tissue epigenetic clocks supports the idea that some of the mechanisms behind the epigenetic ageing clock are shared across tissues. Furthermore, Hannum's epigenetic clock (trained exclusively in blood) explains 72% of variation in chronological age across other tissues (such as breast, kidney, lung and skin), although there is generally a tissue-specific offset [208]. Interestingly, Horvath's epigenetic clock (which is multi-tissue) presents positive epigenetic age acceleration in breast tissue [231], whilst the cerebellum looks younger than expected [232] and some tissues are poorly calibrated (uterine endometrium, dermal fibroblasts, skeletal muscle and heart) [209]. Moreover, Horvath's epigenetic clock dramatically underestimates epigenetic age in sperm [209], which highlights differences between somatic cells and the germline. Altogether, this raises the possibility that **some of the mechanisms behind the epigenetic ageing clock may be shared across tissues** but that they may operate at different rates (e.g. because of different exposure to hormones, differences in proliferation rate, etc.).

Horvath's epigenetic clock works in primary tissues and cell types, and also *in vitro* (both in cell culture and organoids) [209, 233]. Furthermore, recipients of allogeneic hematopoietic stem cell transplantations show an epigenetic age in their blood that corresponds to the age of the donor, even 17 years after the transplantation took place [234]. This suggests that **the epigenetic ageing clock is a stable cell-intrinsic property**, as opposed to the idea that it is highly influenced by the systemic environment (such as the effects observed in heterochronic parabiotic experiments) [235]. The stability is further demonstrated by experiments showing that human fibroblasts that have been reprogrammed into neurons maintain their original epigenetic age [236].

Epigenetic age acceleration (EAA) has been proposed as a way to capture the ageing phenotype in GWAS analysis. Genetic variants associated with EAA have been found in

*TERT*, the catalytic subunit of telomerase [237]. Epigenetic age increases *in vitro* with cell passage, but it requires the expression of *TERT* to keep linearly increasing after a certain number of passages [237]. This suggests that bypassing replicative senescence is required for the epigenetic ageing clock to keep ticking, at least *in vitro*. Interestingly, inducing senescence in TERT-immortalised cells via an oncogene makes the cells age faster in culture, but induction of senescence via DNA damage does not increase epigenetic age [238]. Overall, this could imply that **the epigenome of senescent cells does not contribute substantially to the changes captured by the epigenetic ageing clock**. Furthermore, it has been proposed that epigenetic ageing could serve a complementary role to that of senescence, by suppressing potential cancer development (e.g. by protecting against dedifferentiation signals) [223]. The molecular connections between cell division (positive EAA is generally found in cancerous tissue) [208, 209], alternative non-telomeric functions of *TERT* and the epigenetic ageing clock need to be further studied. Moreover, these experiments do not discard an indirect effect of senescent cells on the epigenetic ageing clock (i.e. via the SASP by inducing changes in the epigenomes of other cells in the tissue) that could occur *in vivo*.

The rate of the epigenetic ageing clock is substantially faster during post-natal organismal growth (something that Horvath's model accounts for) [209], which could be related to the high levels of *TERT* expression during this period [237]. Interestingly, epigenetic ageing according to Horvath's epigenetic clock (but not according to other epigenetic clocks, such as Hannum's clock, the skin-blood clock, *PhenoAge* or *GrimAge*) seems to start a few weeks post-conception in fetal tissues [233]. This could imply that the molecular processes responsible for mammalian epigenetic ageing are operative even during pre-natal development, potentially with different consequences.

This **molecular continuum between development and ageing** is further reinforced by the fact that embryonic stem cells have an epigenetic age around zero [209]. Notably, *in vitro* reprogramming of somatic cells into induced pluripotent stem cells (iPSCs) also reduces epigenetic age to values close to zero (or even negative) both in humans [209] and mice [211, 213]. Moreover, the induction of *in vivo* partial reprogramming (short and cyclic exposure to reprogramming factors) in progeric mice ameliorates several ageing phenotypes and extends lifespan [239]. We are currently testing whether a similar protocol applied to physiologically aged mice can reduce epigenetic age. This is of extreme importance since it shows that the epigenetic changes associated with the epigenetic ageing clock are reversible, which opens the door to further mechanistic studies and to the development of rejuvenation therapies [240–243].

The goal of this thesis is to improve our understanding of the epigenetic ageing clock in humans. For this purpose, I will first review statistical methods to quantify epigenetic ageing in human blood (Chapter 2). Then, I will study how different proteins of the epigenetic machinery affect the rate of the epigenetic ageing clock (Chapter 3). Next, I will discuss a technological improvement with the potential to make future epigenetic clocks more cost-effective (Chapter 4). Finally, I will provide interesting future avenues that should be explored in order to unravel the molecular mechanisms of the epigenetic ageing clock (Chapter 5).

# Chapter 2

# Statistical aspects

> 'I often say that when you can measure what you are speaking about, and express it in numbers, you know something about it; but when you cannot measure it, when you cannot express it in numbers, your knowledge is of a meagre and unsatisfactory kind; it may be the beginning of knowledge, but you have scarcely, in your thoughts, advanced to the stage of *science*, whatever the matter may be'
>
> Lord Kelvin, 1889 [244]

## 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 [245, 246]. Many of these datasets can be obtained from public repositories, such as the NCBI-hosted Gene Expression Omnibus (GEO) [247]. Given its clinical accessibility and ease of collection, blood is one of the most commonly profiled tissues in human DNA methylation studies [246], including published studies on developmental disorders [248] (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 [246]. Additionally, given that the different array versions have different chemistries, biases and number of probes [249–251], I decided to focus on 450K data for my analyses. Using the *GEOquery* R package [252], 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 a minimum requirement.
- In order to study physiological ageing, the blood samples were collected from individuals without prior disease diagnoses. 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 also sample people in their middle ages, before the diseases are normally diagnosed.

### 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 [253–255]. More specifically, a recent study by Je Liu and Kimberly D.

| **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 (control). All the batches were downloaded from GEO [247], with the exception of 'Europe' and 'Feb_2016', which were generated in-house by my 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).

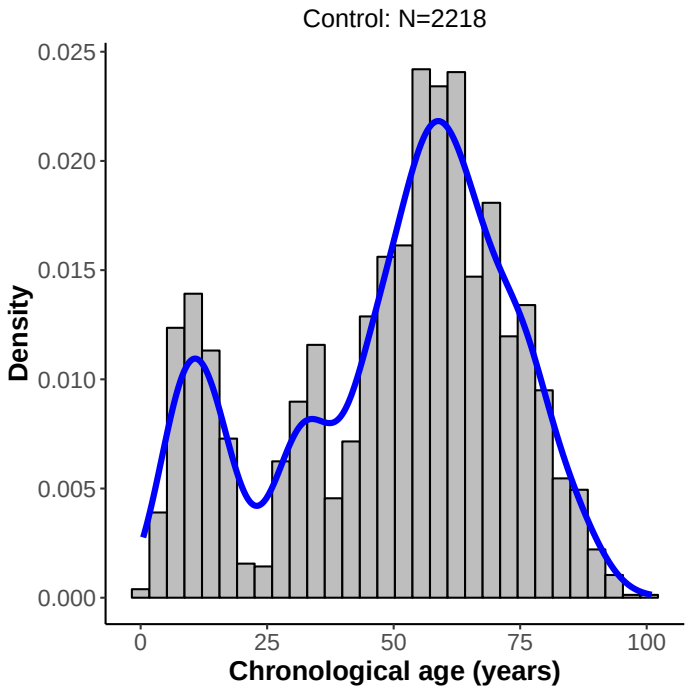



**Fig. 2.1** Histogram showing the chronological age distribution for all the healthy individuals included in the DNA methylation dataset. The blue line represents the 1D kernel density estimate, as calculate by the *stat_density* function in R with default parameters.

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 [255]. Inspired by their results, I implemented a pre-processing pipeline for the 450K data using the *minfi* R package [256] embedded in the following steps (Fig. 2.2):

1. **Background correction**. I used the *noob* method [257], as implemented in the *preprocessNoob* function from the *minfi* R package [256]. *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 ($\beta$-values) obtained (Fig. 2.2b, Fig. S1.1) [257]. 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 [257].

2. **Quality control** (QC). Following guidelines from the *minfi* R package [256], I kept only those samples that satisfied the following criteria:

   (a) The sex predicted from the DNA methylation data ($\text{Sex}_p$) was the same as the reported sex in the metadata. The sex was predicted using the *getSex* function from the *minfi* R package [256], 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 [256].

- Cross-reactive probes, as defined by Chen *et al.* [258]. These are probes that can co-hybridise to alternative genomic sequences that are highly homologous to the target sequences [258].
- 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 [253, 254]. However, I did not include these filters since it was not pointed out in the *minfi* guidelines [256, 259] and it could complicate further downstream analyses (e.g. different sets of probes missing across different batches).

4. **$\beta$-value calculation**. The methylation status of a given CpG site in one of the array probes is normally quantified using the $\beta$-value statistic, which can be calculated as [253, 260]:

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

where $M_i$ and $U_i$ represent the methylated and unmethylated intensity measurements for the $i$th-probe and $\alpha$ is a constant offset (in this work $\alpha = 100$, as recommended by Illumina) [260].

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), $\beta$-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 $\beta$-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:

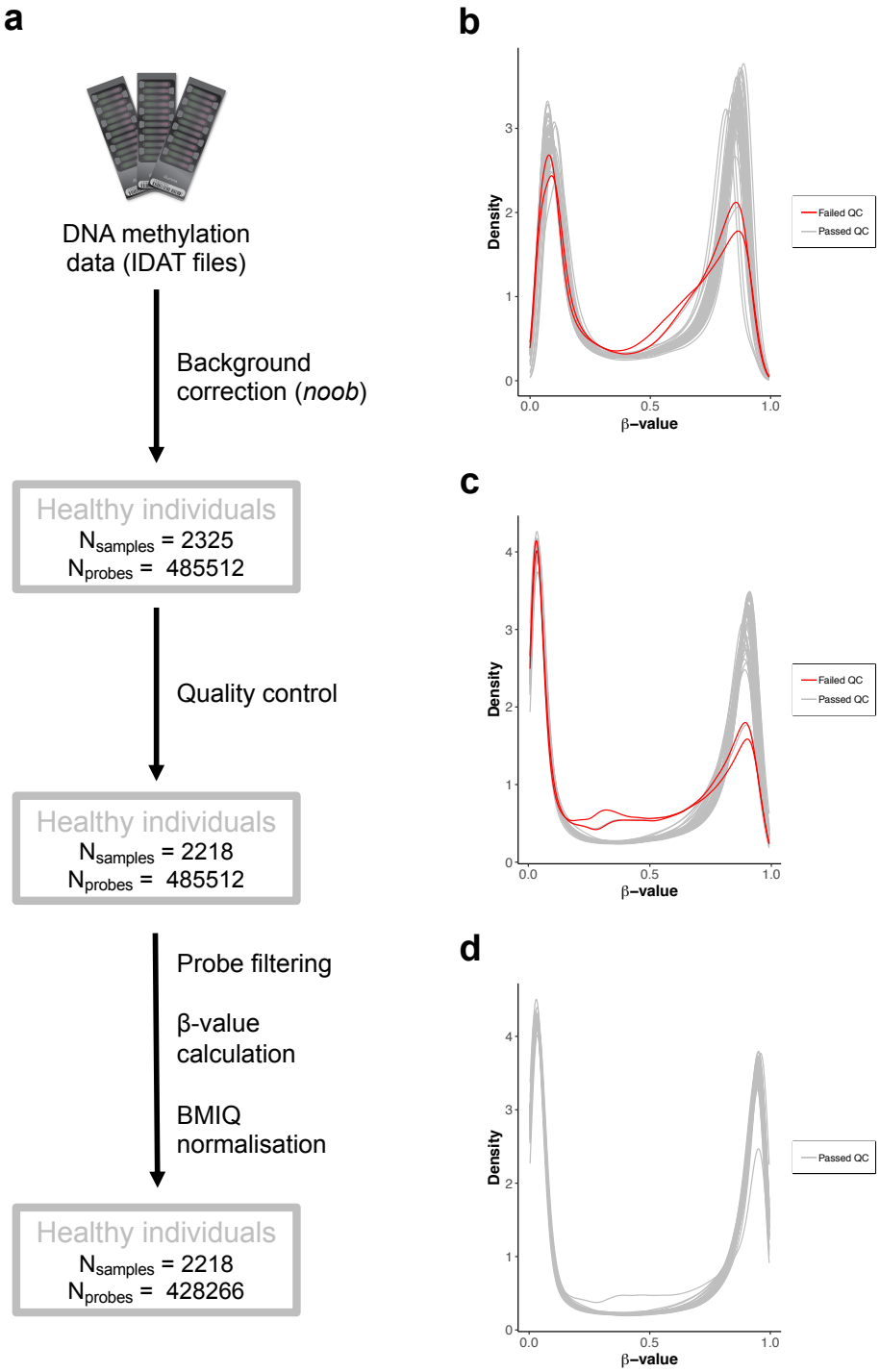



**Fig. 2.2** Main DNA methylation data pre-processing pipeline. **a.** Flowchart showing the main steps implemented to pre-process the DNA methylation data from the 450K methylation arrays. The number of samples ($N_{samples}$) and the number of array probes ($N_{probes}$) left after each step are also specified for the samples from the healthy individuals. **b.** $\beta$-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 $\beta$-values after background correction. **d.** As in b., but calculating the $\beta$-values after background correction, QC, probe filtering and BMIQ normalisation (i.e. the final $\beta$-values that I used for downstream analyses). Note that the samples that failed QC have been removed.

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

with a default offset value of $\alpha = 1$. Du *et al.* reported that $\beta$-values suffer from severe heteroscedasticity (i.e. differences in the variance) for highly methylated or unmethylated CpG sites and therefore the M-values have more desirable statistical properties [260]. However, Zhuang *et al.* later showed that this only becomes a problem in studies with small sample sizes [261] (which is not the case for my analyses). Furthermore, $\beta$-values are easier to interpret biologically and can be readily used in the context of BMIQ normalisation (see below). For these reasons, I choose $\beta$-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 $\beta$-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 [262–265]. 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 ($\beta$-values close to 0) and methylated ($\beta$-values close to 1) probes, this is done by transforming probabilities into quantiles. In the case of 'hemimethylated' probes (intermediate $\beta$-values), a dilation transformation is applied to preserve the monotonicity and continuity of the data [262]. I applied BMIQ to my samples and discarded those that failed the normalisation step.

### 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 [266]. 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 [267].

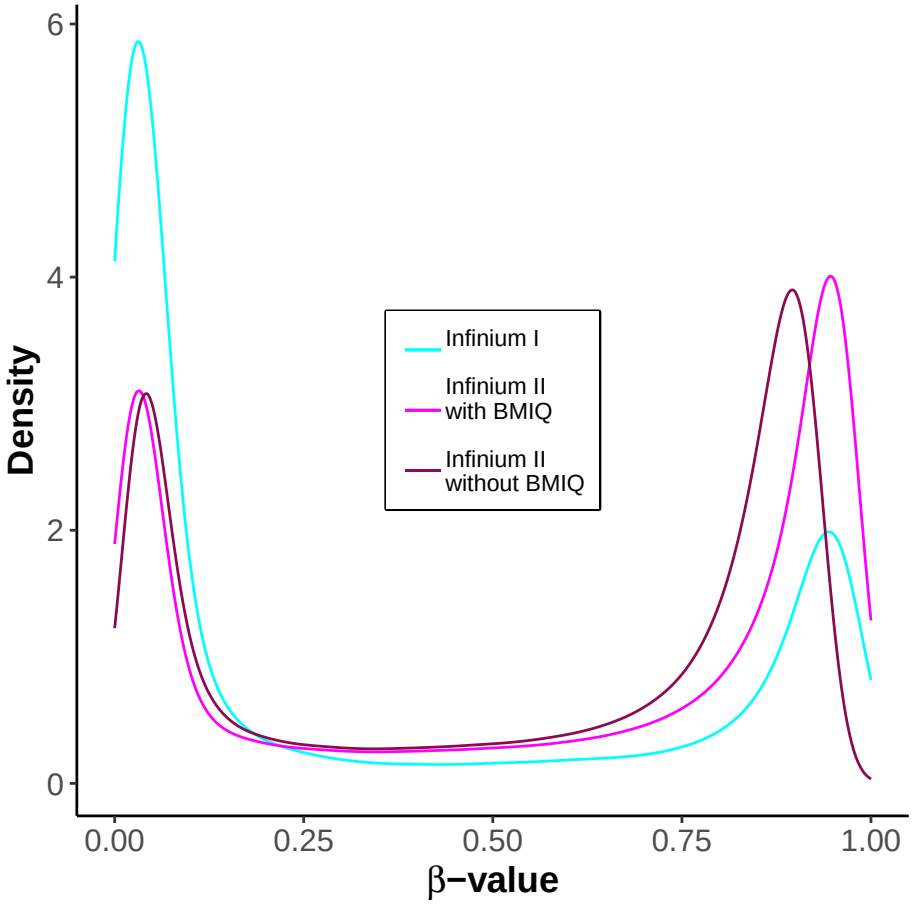



**Fig. 2.3** Effect of BMIQ normalisation on the $\beta$-value distribution of different subsets of array probes with different chemistries (Infinium I, Infinium II). These results correspond to a DNA methylation sample from the GSE41273 batch. It can be appreciated how BMIQ transforms the distribution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

Accounting for this cellular heterogeneity is really important in epigenome-wide association studies (EWAS) [146, 268, 269]. Furthermore, previous research has highlighted changes in blood cell composition with age, which could be one of the causes behind immunosenescence [270–274]. 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 [203, 224, 268].

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) [266, 275–277]. 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 $\beta$-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 $C$ 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 $\sum_{c=1}^{C} w_c \leq 1$) [266]. In mathematical terms:

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

  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 $c$th cell type and $\varepsilon$ is the error [276]. 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) [278] 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 [266]. These methods become particularly useful when no references are available for the cell types that constitute the tissue [266].

However, reference-free approaches rarely provide estimates for the specific cell types in a given sample [266] (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 [276], 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 [276, 279]. 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 six 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 single'computational cell type' as previously done [224, 203]. 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 [256], 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) [268].
    2. The reference implemented in the *EpiDISH* R package (*centDHSbloodDMC.m*) [280]. 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 [276].
    3. The reference implemented as part of the IDOL strategy (IDentifying Optimal DNA methylation Libraries) [279]. 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 [279].

The three references were built using the dataset from Reinius *et al.* (GSE35069) [267], which I obtained directly from the *FlowSorted.Blood.450k* R package [281]. This dataset contains DNA methylation data generated in the 450K array for the six cell types considered, all of which were isolated using flow cytometry [267]. The $\beta$-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 [276, 279] (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.* [278]), RPC (robust partial correlations) [276] and CIBERSORT (which was originally developed for cell-type deconvolution using RNA expression data) [276, 282]. One of the key differences between the algorithms is how the normalisation constrain ($\sum_{c=1}^{C} w_c \leq 1$) is implemented [276]. All the algorithms were run using the implementations in the *epidish* function from the *EpiDISH* R package [280], with the exception of the run in the *minfi* reference, for which I used the *estimateCellCounts* function with default parameters for the 450K array [256].

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 [279]. I assessed the accuracy of the predictions using three different metrics:

- Root mean squared error (*RMSE*), which is 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 is 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 is 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 [279]. However, the gold-standard samples were used as an independent validation of the IDOL reference and should not influence 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

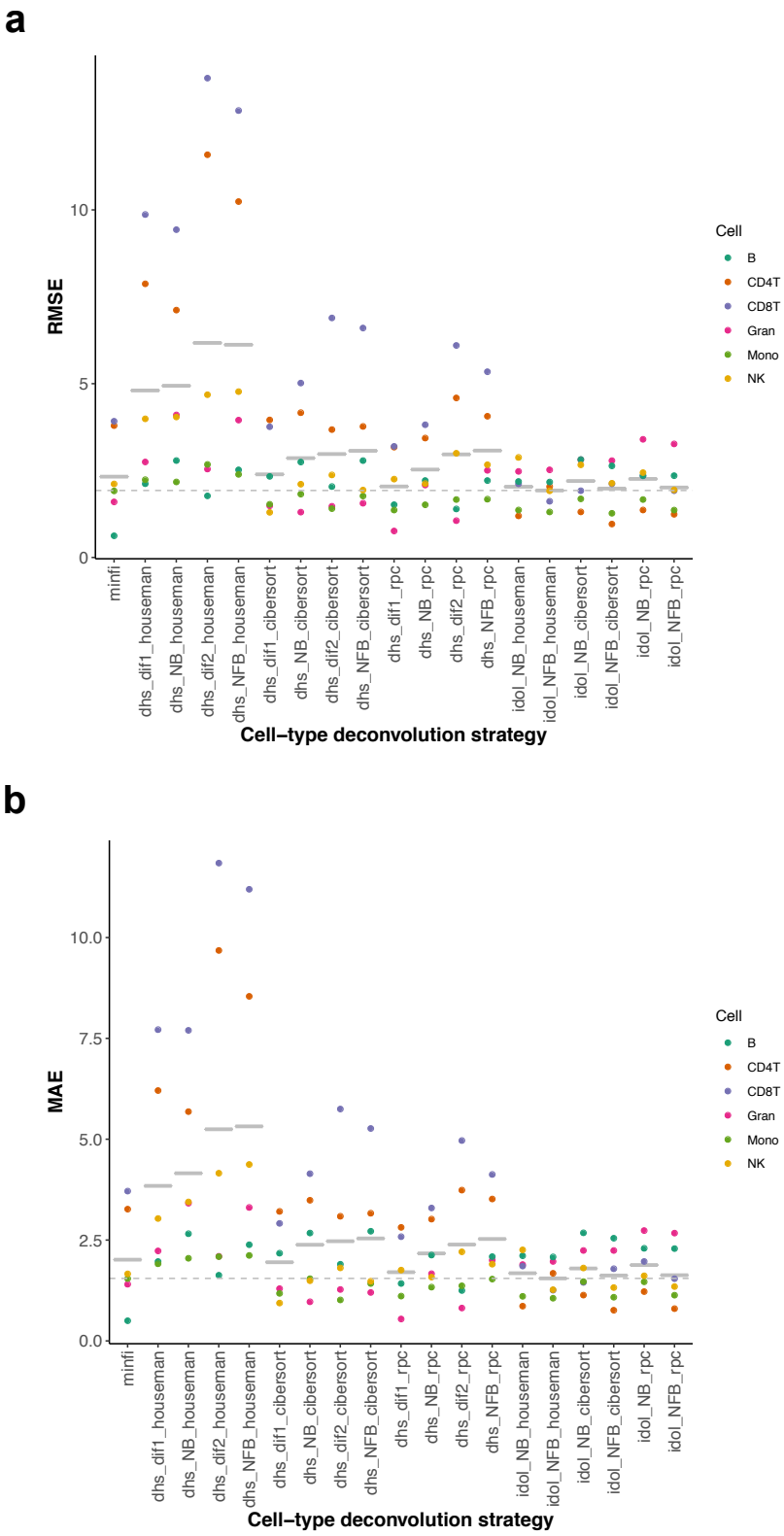



**Fig. 2.4** Benchmarking of the cell-type deconvolution strategies in blood. The x-axis shows the different strategies that were tested (for a detailed description see Fig. S1.4). The y-axis shows the results for **a.** the root mean squared error (*RMSE*) and **b.** the mean absolute error (*MAE*) when comparing the predictions with the real proportions of cells in a gold-standard dataset (GSE77797) [279]. 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.

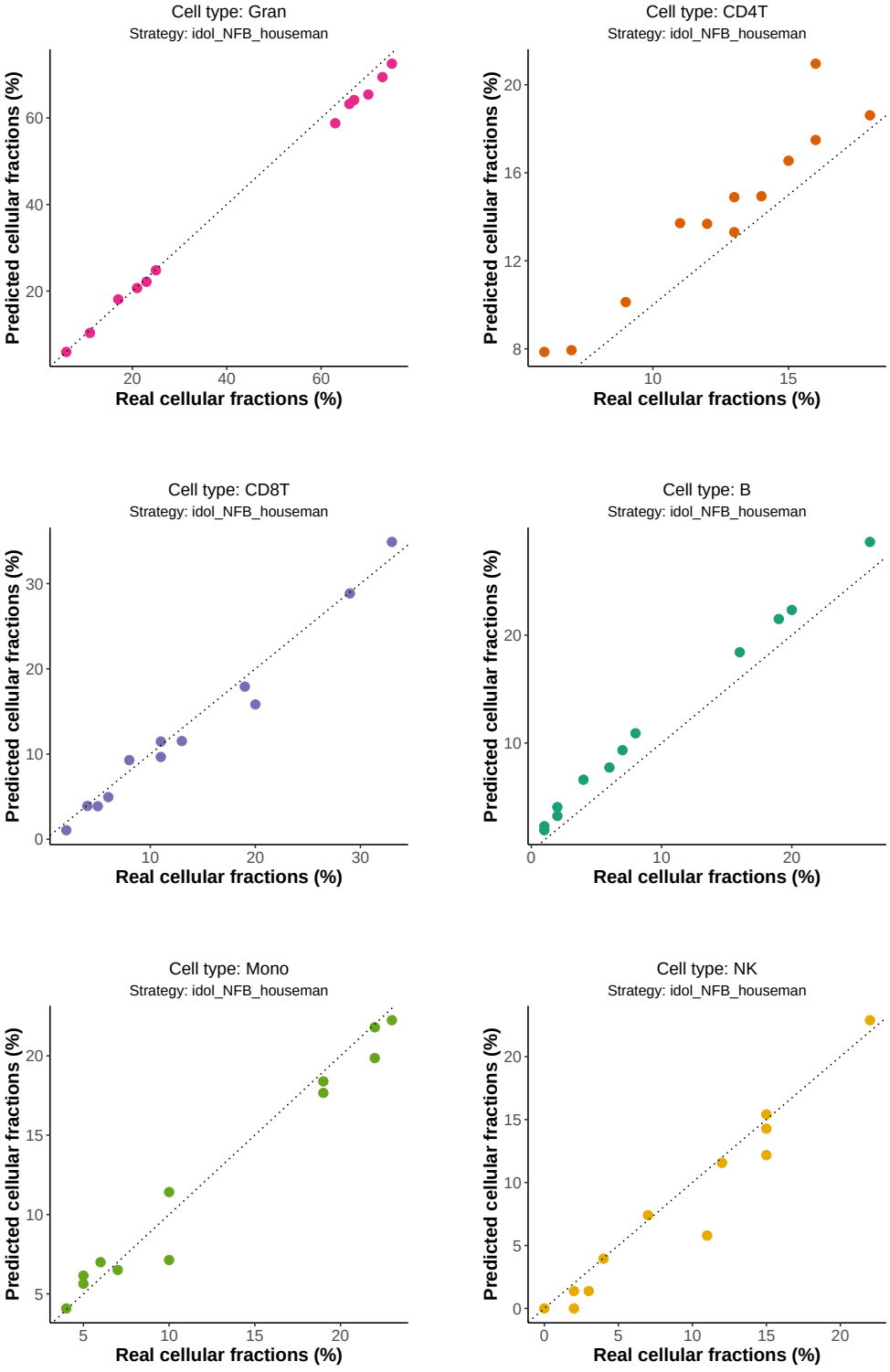



**Fig. 2.5** Comparison of the predictions for the different cell types using the optimal deconvolution strategy ('idol_NFB_houseman') with the real cell type fractions in the gold-standard dataset (GSE77797) [279]. Each point corresponds to a different sample in the gold-standard. The black dashed line represents the diagonal to aid visual interpretation.

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 [283]. 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) [284].

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 reported [270].
- 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 [268, 270–274], 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) [285].

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, etc.), which means that these conclusions must be taken with care [283]. Furthermore, the sex of the individual can influence the proportions of blood leukocytes [272] and it should be taken into account in future analyses.

### 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 [253, 254, 277]. 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

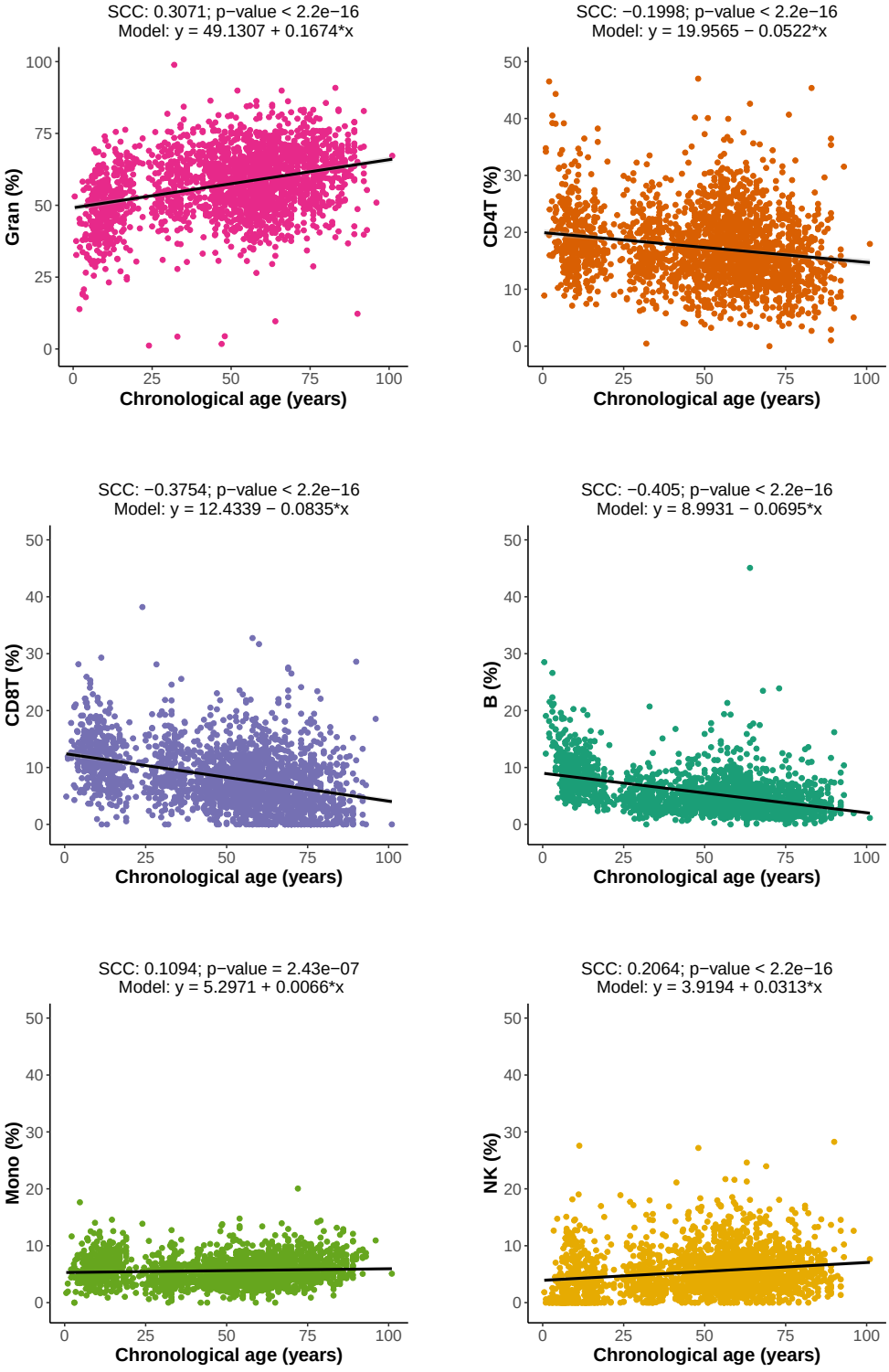



**Fig. 2.6** Changes in blood cell composition during human ageing. Scatterplots showing the changes in the proportions of the six cell types considered (inferred using the cell-type deconvolution strategy) as a function of chronological age. Each point represents a different DNA methylation human sample 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.

disease, when compared with a healthy sample). It is worth mentioning that DMPs are also called differentially methylated cytosines (DMCs) in the literature [277].

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 [286], can also been adapted to find aDMPs [261, 277]. In the case of a continuous variable (such as chronological age) the association is performed using a linear regression modelling framework [261] (see section 2.4 for a short description of linear regression and the nomenclature used throughout 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 $\beta$-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 $\beta$-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 [277], 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 [256] to identify the aDMPs, which internally uses the *limma* framework [286]. 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 [286].

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. However, it is worth mentioning that Bonferroni correction provides a very conservative picture of the methylomic changes (when compared with other methods to control for type-I error, like FDR) and it is likely that an even greater proportion of the methylome is indeed altered with age [179]. 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 $\beta$-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 [3]. It is worth mentioning that around 50% of the CpG sites that constitute the Horvath epigenetic clock are blood aDMPs according to my analysis (Fig. 2.7c,d). Overall, these results are consistent with previous studies [177–179, 287].

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 [178, 288–291, 208]. 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 fatty acids [291]. 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

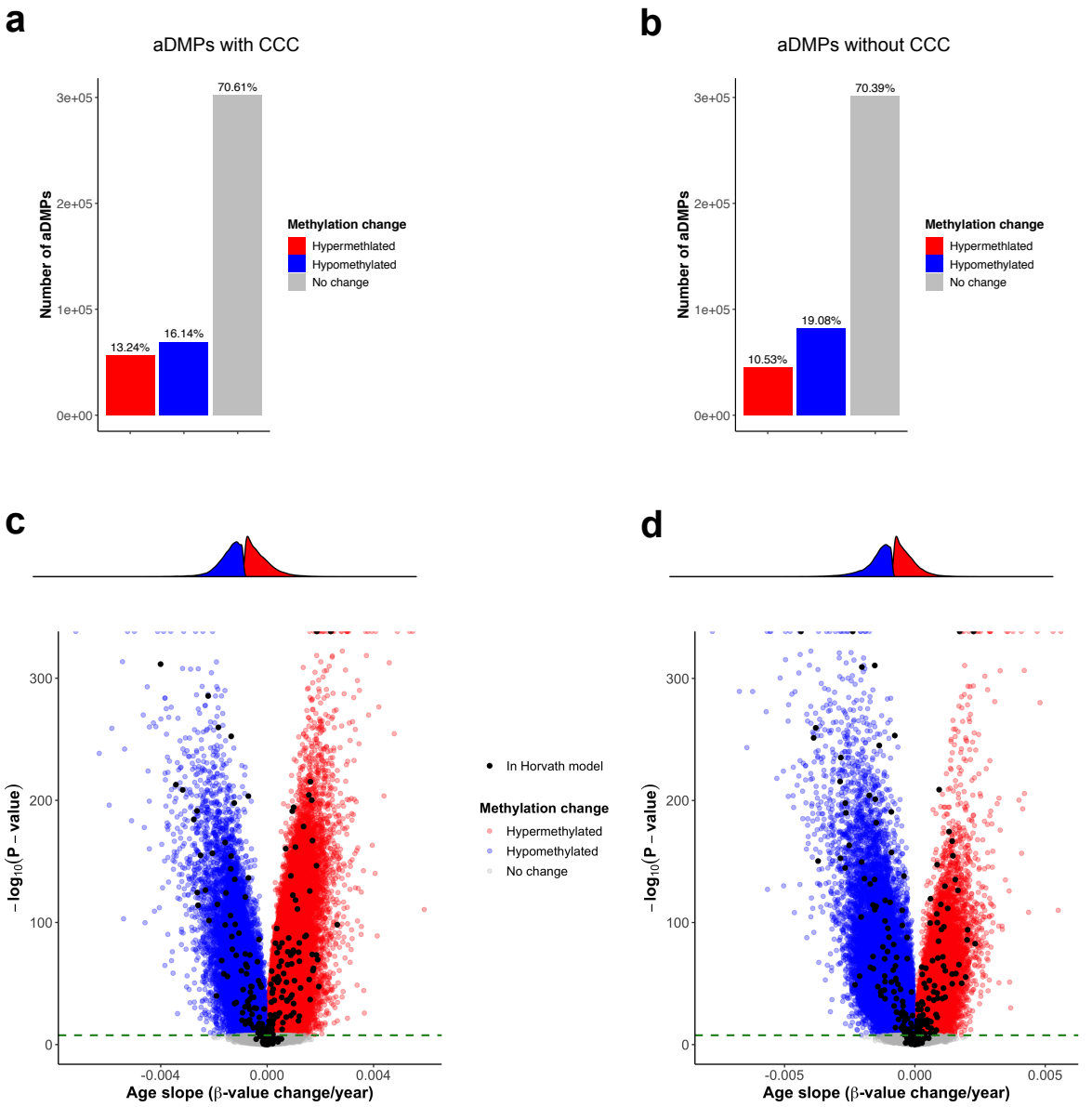



**Fig. 2.7** The blood methylome changes during physiological human ageing. **a.** Barplot showing the total number of differentially methylated positions during ageing (aDMPs) that were identified (in grey: probes that did not reach statistical significance). In this case, the model with cell composition 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 [209]. **d.** As in c., but using the model without CCC.

aDMP, Fig. 2.8b) [288]. These results validate the statistical methods used so far to process the DNA methylation data and to identify aDMPs.

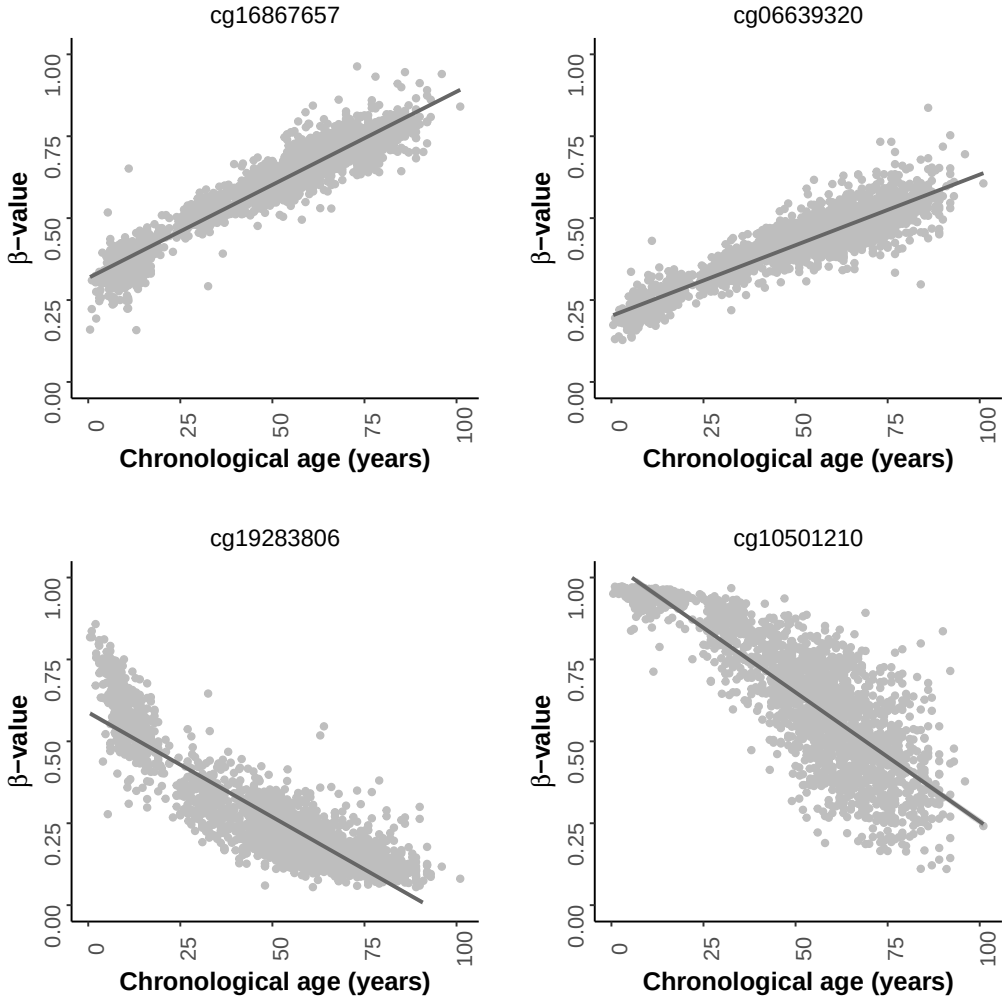



**Fig. 2.8** Changes in the $\beta$-values of four different differentially methylated positions during ageing (aDMPs) in the blood of the healthy individuals. cg16867657 and cg06639320 are the top aDMPs that gain methylation with age (i.e. become hypermethylated) according to the model that accounts 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 $\beta$-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 [209] and that is why he transformed the age 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 [208, 287]. 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 statistic, by estimating the empirical null distribution of the observed set of test statistics [292]. 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) [177], differentially methylated regions (DMRs, which consider several correlated CpGs at the same time) [277] or differentially methylated cytosines in individual cell types (DMCTs, which consider interactions between the phenotypic variable and the proportions of cell types) [293].

### 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 [164, 177, 208, 294, 295]. I calculated it using the same approach as in Hannum *et al.* [208]:

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

where $\beta_i$ represents the methylation $\beta$-value for the $i$th array probe (or CpG site) and $N = 428266$ if all the array probes that passed 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 [177, 208, 164, 295], 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 batch effect 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.

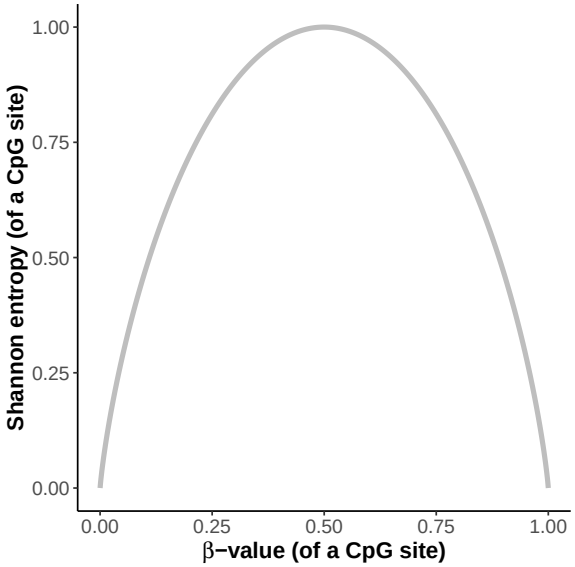



**Fig. 2.9** Plot showing the relationship between the $\beta$-value and the methylation Shannon entropy at a given CpG site (in my case, at a given array probe).

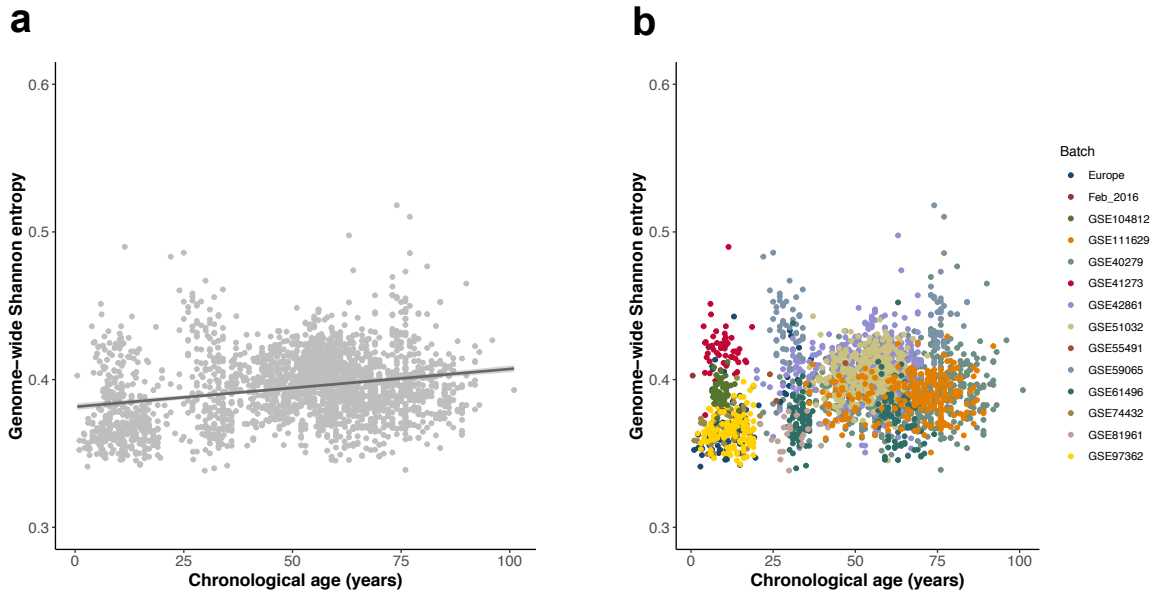



**Fig. 2.10 a.** Scatterplot showing the changes in genome-wide methylation Shannon entropy during ageing in the healthy individuals. Each sample is represented by one point. The black line displays the linear model Entropy $\sim$ Age. **b.** Same as in a., but colouring the samples according to the 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 [209], 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 [296]. 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 [297]. 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 the effect of applying *noob* background correction, before calculating the $\beta$-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 [209]:

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

Therefore, these were the probes selected for downstream analysis.

4. **$\beta$-value calculation**. $\beta$-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 $\beta$-values:

   - Slow imputation (applied when the number of missing $\beta$-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 [298].
   - Fast imputation (applied when the number of missing $\beta$-values is $\geq 3000$). 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 [262]. 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 $\beta$-values for the 21368 probes across all the whole blood samples from [169].

6. **Calculating epigenetic age** (*DNAmAge*). As previously observed for some of the aDMPs, the rate of $\beta$-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).

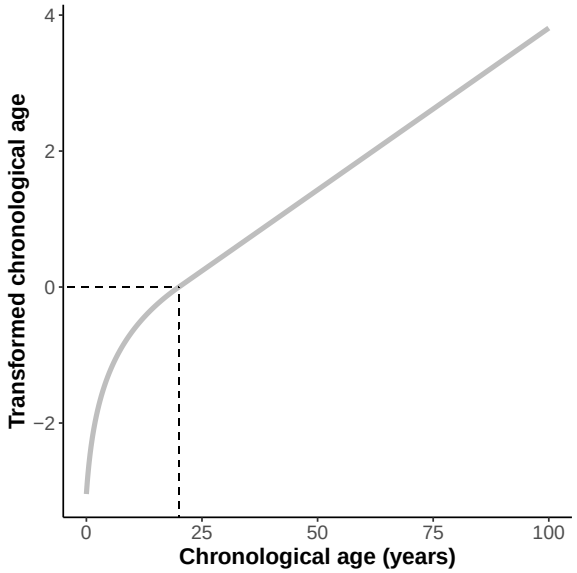



**Fig. 2.11** Plot showing the relationship between the chronological age in years ($c$) and the transformed chronological age ($c_t$) in Horvath's model. This transformation allows accounting for different rates of $\beta$-value change before and after adult age (20 years in humans, as pointed out by 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 $i$th probe (only 353 probes are finally used), $x_i$ is the $\beta$-value for the $i$th probe after gold-standard normalisation and $g(\cdot)$ is 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 $\approx 0$). 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 the 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 [224, 203] (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) [224, 203].

- 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 is calculated as:

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

where $EAA_i$ is the epigenetic age acceleration for the $i$th 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 approach 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) [209]. 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 [209], 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 than 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 [299, 300] and mice [210]. However, to this date, it is unclear

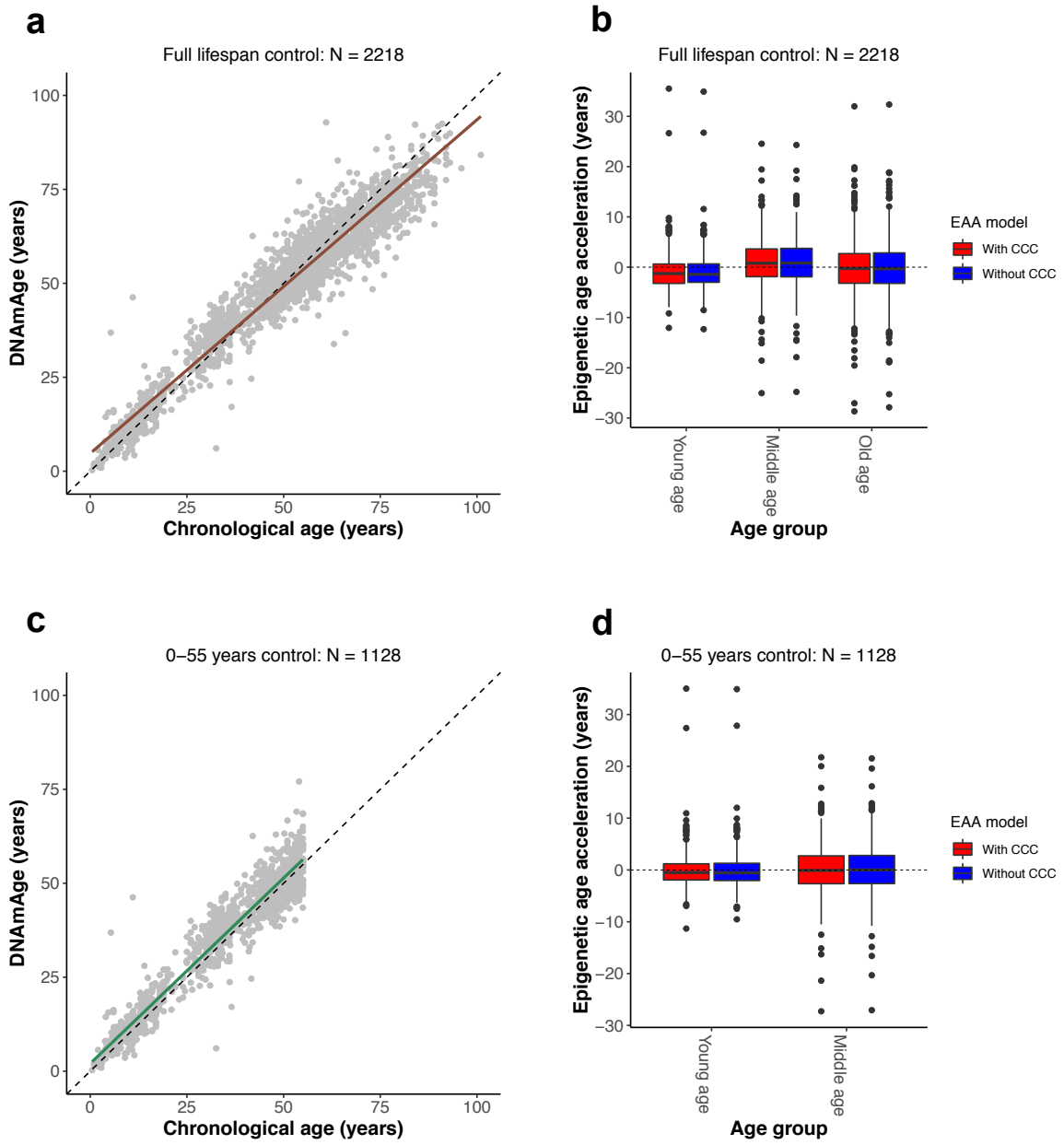



**Fig. 2.12** Horvath's epigenetic clock measures physiological ageing. **a.** Scatterplot showing the relationship between epigenetic age (*DNAmAge*) according to Horvath's model [209] and chronological age of the samples for the healthy individuals. Each sample is represented by one point. The black 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: $\leq 20$ 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).

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

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, when 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 [209]. 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 [301]. 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, etc. [301]. Correcting for batch effects is crucial, especially when integrating data from different studies and sources [302], 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 [303] or M-values [302, 304]. In the context of the epigenetic clock, previous attempts to account for technical variation have used the first five PCs estimated directly from the DNA methylation data (presumably the $\beta$-values) [222]. However, this approach potentially removes meaningful biological variation, especially in studies with global changes in DNA methylation, such as cancer [303] 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 $\beta$-values would impact the predictions [305].

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) [302, 303, 306]. 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 [303, 307]. 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 [224, 203]. The optimal number of PCs was found by making use of the *findElbow* function from [308].

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).

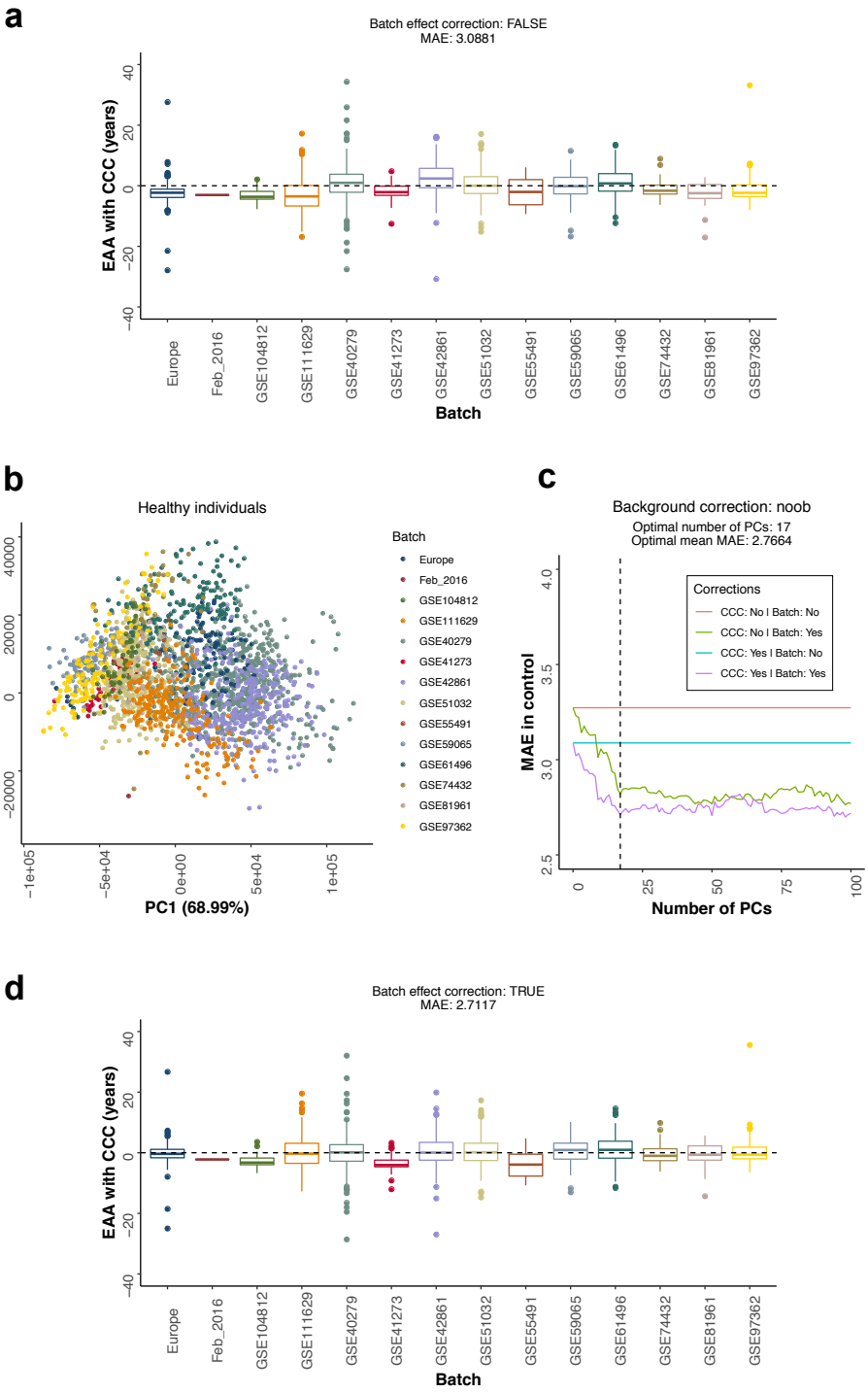



**Fig. 2.13** Correcting for batch effects in the context of the epigenetic clock. **a.** Distribution of the epigenetic age acceleration (EAA) for the different batches of healthy individual samples, using the control model with cell composition correction (CCC) and before applying batch effect correction. 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).

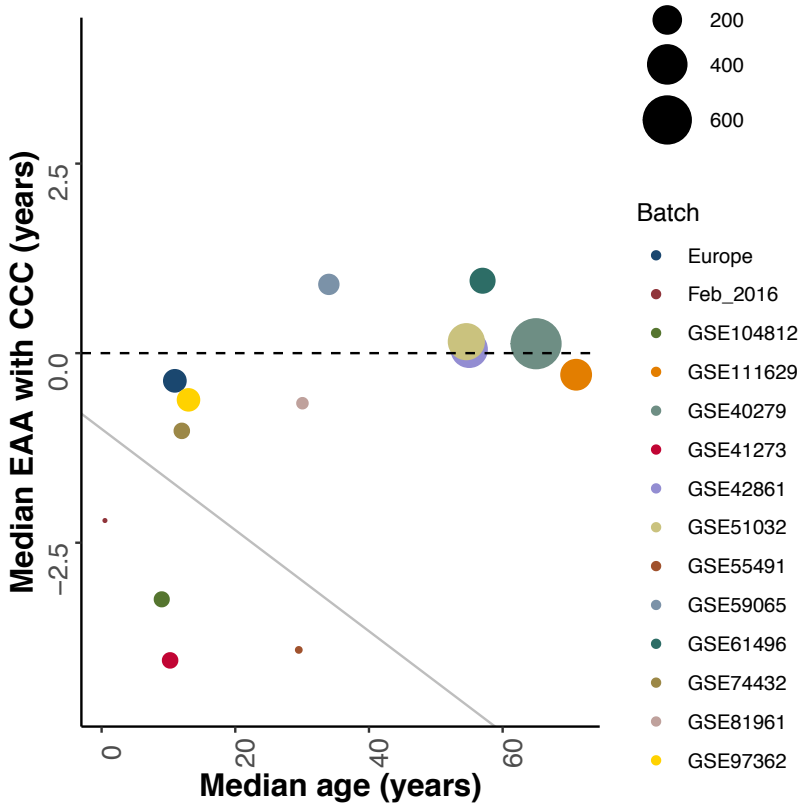



**Fig. 2.14** After applying batch effect correction in the samples from the healthy individuals, deviations from a median epigenetic age acceleration (EAA) of zero (dotted black line) in some of the batches can be explained by other causes. The grey line separates in the lower left corner those weird 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 [224, 203, 300, 215, 309, 310]. Hannum's model was originally trained in whole blood and it makes use of a linear combination of $\beta$-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 performed 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 [215]. 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 [208]. Horvath's and Hannum's epigenetic clocks are correlated (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 three 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 [310]. 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 [215, 310].

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 and is associated with cancer risk [225]. 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 [183]. 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 [225]. 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 [237]. 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) [209, 231] 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

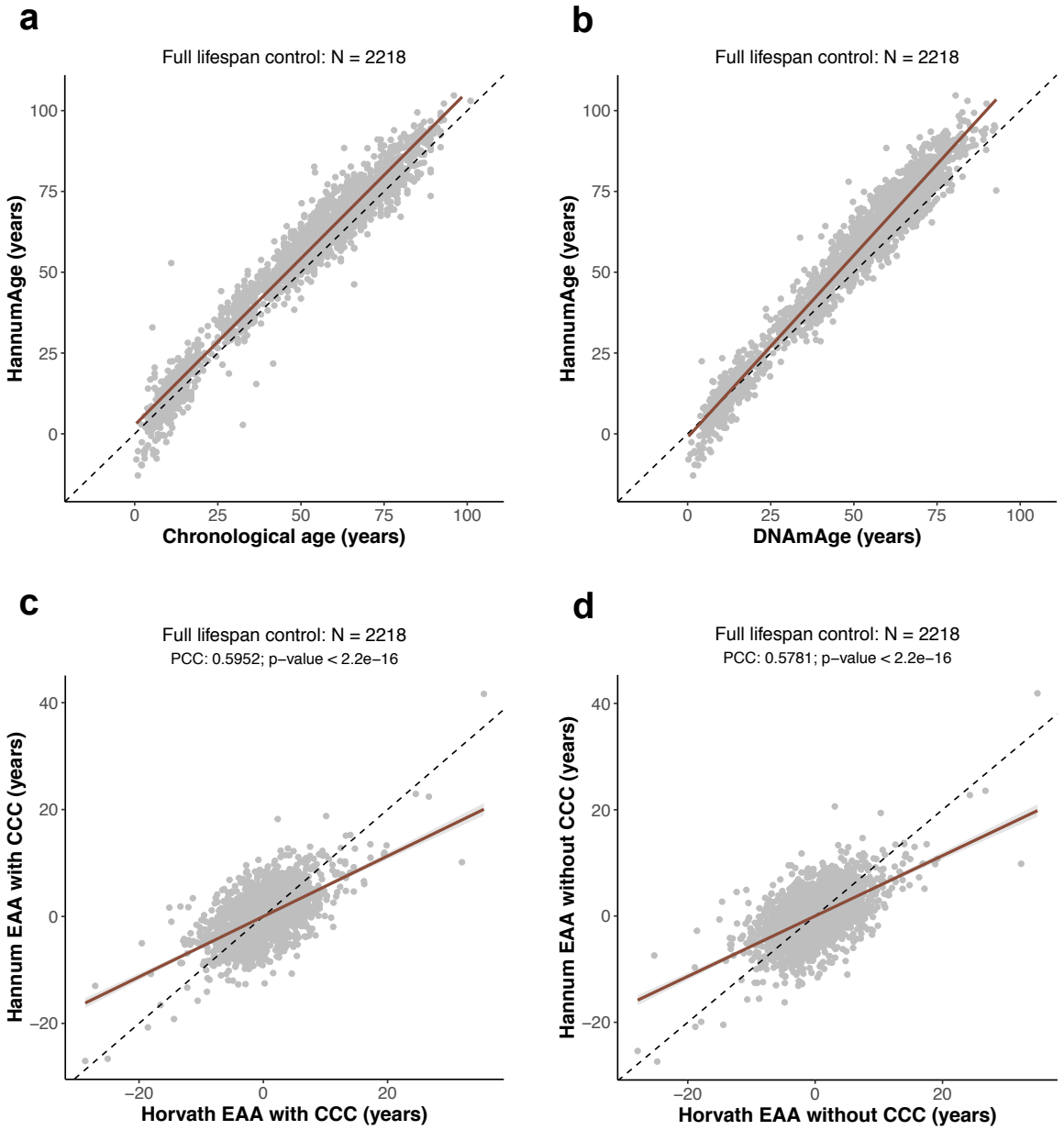



**Fig. 2.15** Behaviour of Hannum's epigenetic clock in the healthy individuals. **a.** Scatterplot showing the relationship between the epigenetic age predicted with Hannum's model (*HannumAge*) [208] and chronological age of the samples for the healthy individuals. Each sample is represented by one 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.

its 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 [286] or genome-wide association studies (GWAS) [311]. An instance of such models is linear regression [312], 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 analytical flexibility [286].

We can describe the dependent variable $\mathbb{Y}$ as a function of the independent variables $\mathbb{X}_j$:

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

where $\beta_j$ are unknown parameters that need to be estimated from the data and $\varepsilon_i$ is the random error. In matrix form:

$$\mathbf{y} = \mathbf{X}\beta + \varepsilon \tag{2.20}$$

where $\mathbf{y} \in R^n$ is the vector $\{y_1,...,y_n\}$, $\mathbf{X} \in R^{n\times k}$ 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 $\mathrm{Cov}(\varepsilon) = \sigma^2\mathbf{I_n}$ (where $\mathbf{I_n}$ is the $n \times n$ identity matrix) and applying the Gauss-Markov theorem [312], it can be demonstrated that:

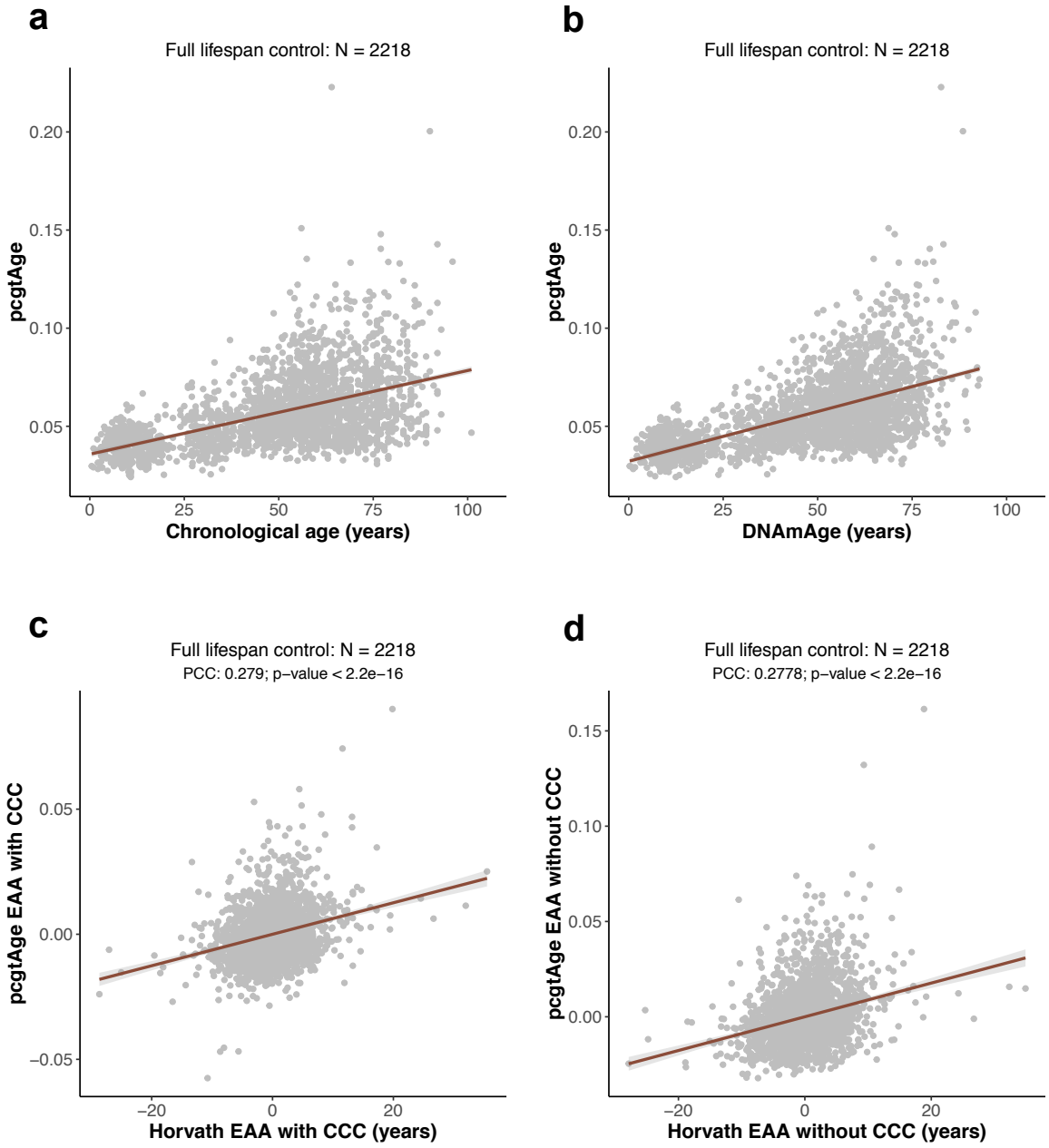



**Fig. 2.16** Behaviour of the epigenetic mitotic clock (*epiTOC*) in the healthy individuals. **a.** Scatterplot showing the relationship between mitotic age (*pcgtAge*) [225] and chronological age of the samples for the healthy individuals. Each sample is represented by one point. The solid brown line 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.

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

where $\mathbf{X}'$ is the transpose of $\mathbf{X}$ and $\hat{\beta}$ is the least-squares estimator of $\beta$, 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 [313]:

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

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. $\beta$-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 $\beta$-value at a certain CpG probe for the $i$th sample, $x_{i1}$ is the age for the $i$th sample, $x_{i2}$ is the sex (e.g. 0 for male and 1 for female) for the $i$th sample, $x_{i3}$ is the

disease status (e.g. 0 for a healthy individual and 1 for an individual with a disease) for the $i$th sample, $\beta_0$ is the intercept coefficient, $\beta_j$ are the covariate coefficients ($j = 1$ for age, $j = 2$ for sex, $j = 3$ for disease status) and $\varepsilon_i$ is the error for the $i$th sample.

Throughout this thesis, I use the following nomenclature to describe the model above ('R-style' nomenclature):

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

# Chapter 3

# Biological aspects

> 'At a fundamental level evolutionary survival is the preservation of a dynamic balance between information, or order, and entropy, or disorder.'
>
> T. B. L. Kirkwood, 1977 [10]

### Declaration

This chapter in mainly the product of my own work. Additionally, I would like to recognise the contributions of Janet M. Thornton, Wolf Reik and Thomas M. Stubbs (who helped designing the study and interpreting the data), Erfan Aref-Eshghi (who run some of the analyses using my code and provided part of the samples in the dataset), Marc Jan Bonder and Oliver Stegle (who provided statistical input) and Bekim Sadikovic (who provided part of the samples in the dataset). All of them also helped in the revision of the final text. This work is now under peer-review in the journal *Genome Biology* [314].

## 3.1 Background

Epigenetic clocks can be understood as a proxy to quantify the changes of the epigenome with age. However, little is known about the molecular mechanisms that determine the rate of these clocks. Steve Horvath proposed that the multi-tissue epigenetic clock captures the workings of an **epigenetic maintenance system** [209]. Recent GWAS studies have found several genetic variants associated with epigenetic age acceleration in genes such as *TERT* (the catalytic subunit of telomerase) [237], *DHX57* (an ATP-dependent RNA helicase) [315] or *MLST8* (a subunit of both mTORC1 and mTORC2 complexes) [315]. Nevertheless, to my

knowledge no genetic variants in epigenetic modifiers have been found and the molecular nature of this hypothetical system is unknown to this date.

I decided to take a reverse genetics approach and look at the **behaviour of the epigenetic clock in patients with developmental disorders**, many of which harbour mutations in proteins of the epigenetic machinery [248, 316]. I performed an unbiased screen for epigenetic age acceleration and found that Sotos syndrome accelerates epigenetic ageing, potentially revealing a role of H3K36 methylation maintenance in the regulation of the rate of the epigenetic clock.

## 3.2 Screening for genes that accelerate the epigenetic clock

The main goal of this analysis is to identify genes, mainly components of the epigenetic machinery, that can **affect the rate of epigenetic ageing in humans** (as measured by Horvath's epigenetic clock) [209]. For this purpose, I assembled a dataset with all the DNA methylation data from patients with different developmental disorders that I could find, in order to perform an unbiased screen. This dataset combines samples publicly available in GEO [247] with in-house data generated by my collaborators at the London Health Sciences Centre, Canada (Table S2.1, Fig. S2.1). All these data were generated from blood using the Illumina 450K methylation array, as in the case of the healthy individuals described in Chapter 2.

Many of these developmental syndromes have overlapping clinical features [248, 316]. Furthermore, in some cases with a clinical diagnosis, the genetic cause remains unknown, probably due to locus heterogeneity or difficulty to assess the clinical significance of some genetic variants [317]. Therefore, several studies have explored the ability of DNA methylation signatures to aid differential diagnoses of these syndromes [248, 317–327]. Given that most of the diagnoses for developmental disorders are carried out early in life, this dataset has a bias towards younger ages (Fig. 3.1). In order to maximise the ability to detect ageing-associated effects, I kept only those developmental disorders with at least 5 samples, of which at least 2 had a chronological age $\geq 20$ years (which, according to Horvath's model, is the adult age for humans) [209]. This filtering resulted in a dataset for the main screen with $N = 367$ samples from cases, which had ages between 0 and 55 years (Fig. 3.2, Table 3.1).

The purpose of the screen is to **test whether the epigenetic ages of the samples from a given developmental disorder (cases) deviate from their chronological age** i.e. identify those developmental disorders that present epigenetic age acceleration (EAA). For a given

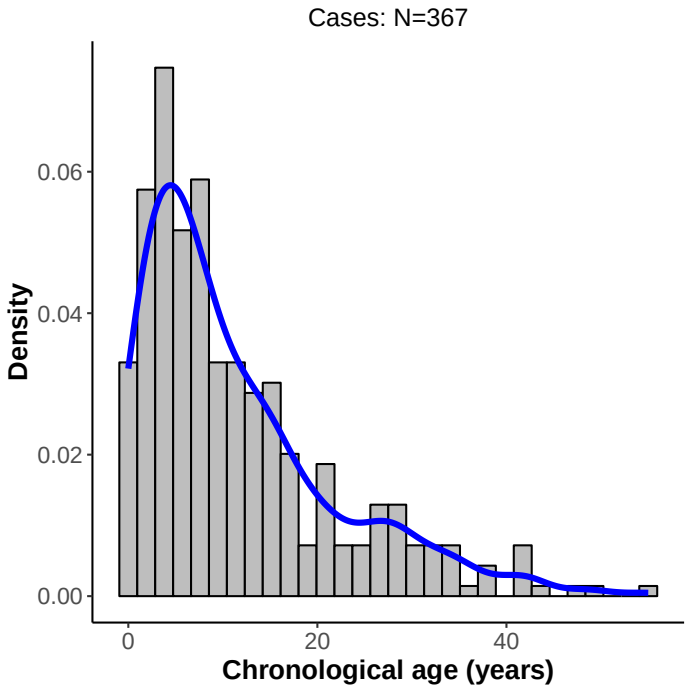



**Fig. 3.1** Histogram showing the chronological age distribution for all the individuals with developmental disorders (cases) included in the final dataset (i.e. after QC and filtering). The blue line represents the 1D kernel density estimate, as calculate by the *stat_density* function in R with default parameters.

| Developmental disorder | Gene(s) involved | Gene(s) function | Molecular cause | N | Age range (years) |
|---|---|---|---|---|---|
| Angelman | *UBE3A* | Ubiquitin protein ligase E3A | Imprinting, mutation | 14 | 1 to 55 |
| Autism spectrum disorder (ASD) | - | - | - | 119 | 1.83 to 35.16 |
| Alpha thalassemia/mental retardation X-linked syndrome (ATR-X) | *ATRX* | Chromatin remodelling | Mutation | 15 | 0.7 to 27 |
| Claes-Jensen | *KDM5C* | H3K4 demethylase | Mutation | 10 | 2 to 42 |
| Coffin-Lowry | *RPS6KA3* | Serine / threonine kinase | Mutation | 10 | 1.3 to 22.8 |
| Floating-Harbour | *SRCAP* | Chromatin remodelling | Mutation | 17 | 4 to 42 |
| Fragile X syndrome (FXS) | *FMR1* | Translational control | Mutation (CGG expansion) | 32 | 0.08 to 48 |
| Kabuki | *KMT2D* | H3K4 methyltransferase | Mutation | 46 | 0 to 24.1 |
| Noonan | *PTPN11*, *RAF1*, *SOS1* | RAS/ MAPK signalling | Mutation | 15, 11, 14 | 0.2 to 49 |
| Rett | *MECP2* | Transcriptional repression | Mutation | 15 | 1 to 34 |
| Saethre-Chotzen | *TWIST1* | Transcription factor | Mutation | 22 | 0 to 38 |
| Sotos | *NSD1* | H3K36 methyltransferase | Mutation | 20 | 1.6 to 41 |
| Weaver | *EZH2* | H3K27 methyltransferase | Mutation | 7 | 2.58 to 43 |
| **Total** | - | - | - | 367 | 0 to 55 |

**Table 3.1** Overview of the developmental disorders that were included in the screening after quality control and filtering (total $N = 367$).

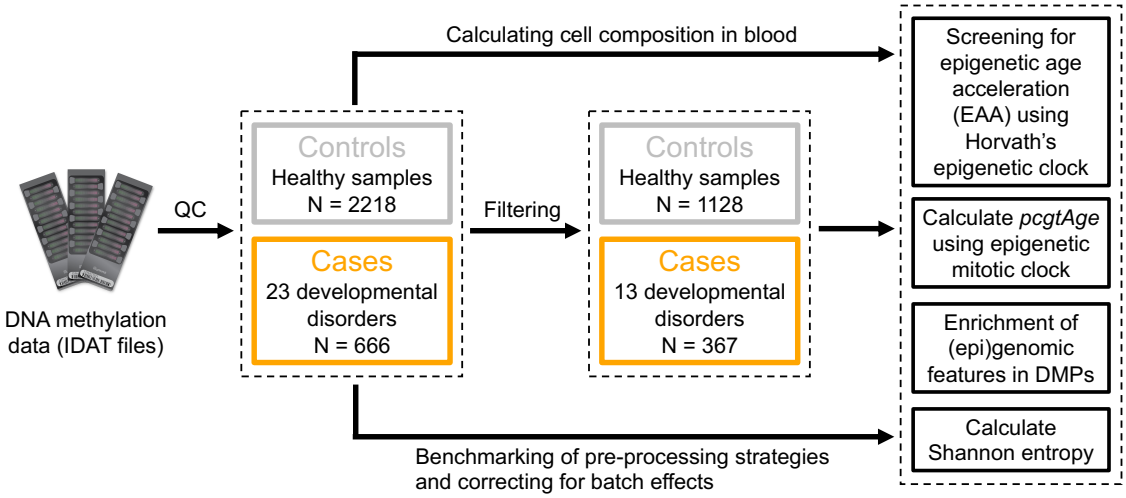



**Fig. 3.2** Flow diagram that portrays an overview of the different analyses that are carried out in the raw DNA methylation data (IDAT files) from human blood for cases (developmental disorders samples) and controls (healthy samples). The control samples are filtered to match the age range of the cases (0-55 years). The cases are filtered based on the number of 'adult' samples available (for each disorder, at least 5 samples, with 2 of them with an age $\geq$ 20 years). QC: quality control. DMPs: differentially methylated positions.

sample, a positive EAA indicates that the epigenetic (biological) age of the sample is higher than the one expected for someone with that chronological age. In other words, it means that the epigenome of that person resembles the epigenome of an older individual. The opposite is true when a negative EAA is found (i.e. the epigenome looks younger than expected). I calculated the epigenetic ages (*DNAmAge*) of all the samples according to Horvath's epigenetic clock (see section 2.2.1) and I fitted the control models to the samples from the healthy individuals, including models with and without blood cell composition correction (CCC) and always accounting for potential batch effects (see equations 2.16 and 2.17). As previously discussed (see section 2.2.2), due to the fact that Horvath's model underestimates the epigenetic age of old samples, the age distribution of the control samples can have an impact on the results of the screen. Therefore, I filtered the ages of the healthy individual samples to make them match the age range of the developmental disorders (0-55 years, $N = 1128$, see Fig. 3.2).

The EAA for the control samples corresponds to the residuals from the control models (see section 2.2.2). On the other hand, the EAA for a case sample is calculated by taking the difference between the epigenetic age (*DNAmAge*) and the predicted value from the

corresponding control model (with or without cell composition). Finally, I compared the distributions of the EAA for the different developmental disorders against the EAA distributions for the healthy controls using the non-parametric two-sided Wilcoxon's test. P-values were adjusted for multiple testing using Bonferroni correction and a significance level of $\alpha = 0.01$ was applied. It is worth mentioning that some of the developmental disorders included in the screen (such as autism spectrum disorder or Coffin-Lowry syndrome) are not necessarily caused by alterations in the epigenetic machinery, but were still included to maintain the unbiased nature of the screen.

## 3.3 Sotos syndrome accelerates epigenetic ageing

The results from the screen are portrayed in Fig. 3.3. Most syndromes do not show evidence of accelerated epigenetic ageing, but **Sotos syndrome presents a clear positive EAA** (median $\text{EAA}_{\text{with CCC}}$ = + 7.64 years, median $\text{EAA}_{\text{without CCC}}$ = + 7.16 years), with p-values considerably below the significance level of 0.01 after Bonferroni correction ($\text{p-value}_{\text{corrected, with CCC}} = 3.40 \cdot 10^{-9}$, $\text{p-value}_{\text{corrected, without CCC}} = 2.61 \cdot 10^{-7}$). Additionally, Rett syndrome (median $\text{EAA}_{\text{with CCC}}$ = + 2.68 years, median $\text{EAA}_{\text{without CCC}}$ = + 2.46 years, $\text{p-value}_{\text{corrected, with CCC}}$ = 0.0069, $\text{p-value}_{\text{corrected, without CCC}}$ = 0.0251) and Kabuki syndrome (median $\text{EAA}_{\text{with CCC}}$ = - 1.78 years, median $\text{EAA}_{\text{without CCC}}$ = - 2.25 years, $\text{p-value}_{\text{corrected, with CCC}}$ = 0.0011, $\text{p-value}_{\text{corrected, without CCC}}$ = 0.0035) reach significance, with a positive and negative EAA respectively. Finally, fragile X syndrome (FXS) shows a positive EAA trend (median $\text{EAA}_{\text{with CCC}}$ = + 2.44 years, median $\text{EAA}_{\text{without CCC}}$ = + 2.88 years) that does not reach significance in the screen ($\text{p-value}_{\text{corrected, with CCC}}$ = 0.0680, $\text{p-value}_{\text{corrected, without CCC}}$ = 0.0693).

Next, I tested the effect of changing the median age used to build the healthy control model (i.e. the median age of the controls) on the screening results (Fig. S2.2). Sotos syndrome is robust to these changes, whilst Rett, Kabuki and FXS are much more sensitive to the control model used. This again highlights the importance of choosing an appropriate age-matched control when testing for epigenetic age acceleration, given that Horvath's epigenetic clock underestimates epigenetic age for advanced chronological ages [299, 300].

Moreover, all but one of the Sotos syndrome patients (19/20 = 95%) show a consistent deviation in EAA (with CCC) in the same direction (Fig. 3.4a,b), which is not the case for the rest of the disorders, with the exception of Rett syndrome (Fig. S2.3). Even though these data suggest that there are already some methylomic changes at birth, the EAA seems to increase with age in the case of Sotos patients (Fig. 3.4b). This implies that at least some of the

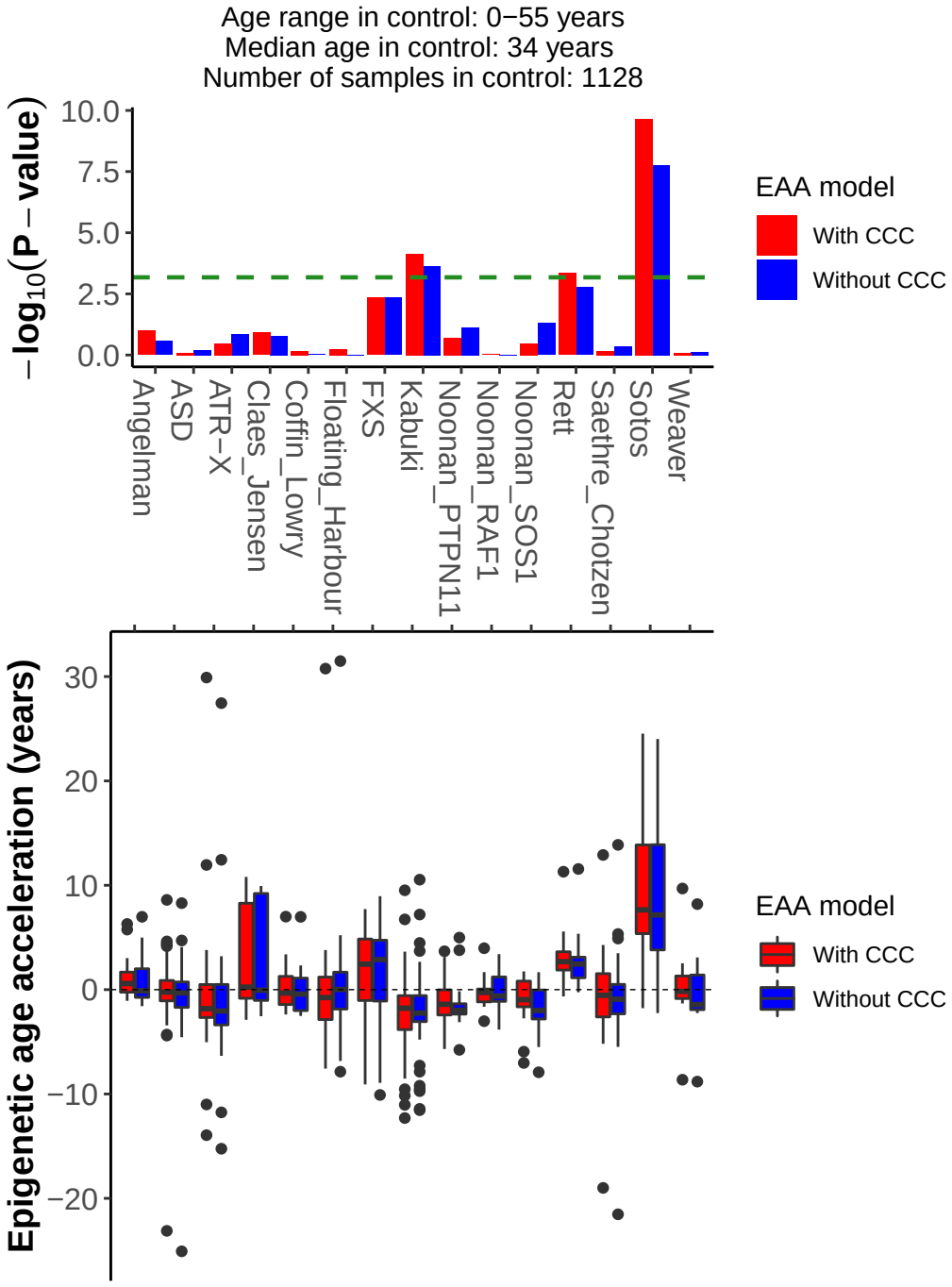



**Fig. 3.3** Screening for epigenetic age acceleration (EAA) in developmental disorders. The upper panel shows the p-values derived from comparing the EAA distributions for the samples in a given developmental disorder and the control (two-sided Wilcoxon's test). The dashed green line displays the significance level of $\alpha = 0.01$ after Bonferroni correction. The bars above the green line reach statistical significance. The lower panel displays the actual EAA distributions, which allows assessing the direction of the EAA (positive or negative). In red: EAA model with cell composition correction (CCC). In blue: EAA model without CCC. ASD: autism spectrum disorder. ATR-X: alpha thalassemia/mental retardation X-linked syndrome. FXS: fragile X syndrome.

changes that normally affect the epigenome with age are happening at a faster rate in Sotos syndrome patients during their lifespan (as opposed to the idea that the Sotos epigenetic changes are only acquired during prenatal development and remain constant afterwards).

Finally, I investigated whether Sotos syndrome leads to a higher rate of (stem) cell division in blood when compared with the healthy population. I employed the epigenetic mitotic clock, that makes use of the fact that some CpGs in promoters that are bound by Polycomb group proteins become hypermethylated with age (captured by a metric called *pcgtAge*; see section 2.3.2). This hypermethylation correlates with the number of cell divisions in the tissue and is also associated with an increase in cancer risk [225]. I calculated *pcgtAge* for the Sotos samples and compared them against the healthy controls (using a model similar to the one in equation 2.16, although in this case the dependent variable was *pcgtAge*; see section 2.3.2). I found a trend suggesting that **the epigenetic mitotic clock might be accelerated in Sotos patients** (p-value = 0.0112, Fig. 3.4c,d), which could explain the higher cancer predisposition reported in these patients and might relate to their overgrowth [328].

Consequently, I report that individuals with Sotos syndrome present an accelerated epigenetic age, which makes their epigenome look, on average, more than 7 years older than expected. These changes seem to be the consequence of a higher ticking rate of the epigenetic clock (or at least part of its machinery), with epigenetic age acceleration increasing during lifespan: the youngest Sotos patient (1.6 years) has an $EAA_{\text{with CCC}}$ = 5.43 years and the oldest (41 years) has an $EAA_{\text{with CCC}}$ = 24.53 years. Additionally, Rett syndrome, Kabuki syndrome and fragile X syndrome could also have their epigenetic ages affected, but more evidence is required to be certain about this conclusion.

## 3.4 Comparing Sotos syndrome and physiological ageing

Sotos syndrome is caused by loss-of-function heterozygous mutations in the NSD1 gene, a histone H3K36 methyltransferase [320, 329]. These mutations lead to a specific DNA methylation signature in Sotos patients, potentially due to the crosstalk between the histone and DNA methylation machinery [320]. In order to gain a more detailed picture of the reported epigenetic age acceleration, I decided to compare the genome-wide (or at least array-wide) changes observed in the methylome during ageing with those observed in Sotos syndrome. For this purpose, I **identified differentially methylated positions (DMPs) for both conditions**, using the models that account for cell composition correction (see equations 2.10 and 3.1). Ageing DMPs (aDMPs) were calculated in this case using the healthy samples in the age range 0-55 years. aDMPs were composed almost equally of CpG

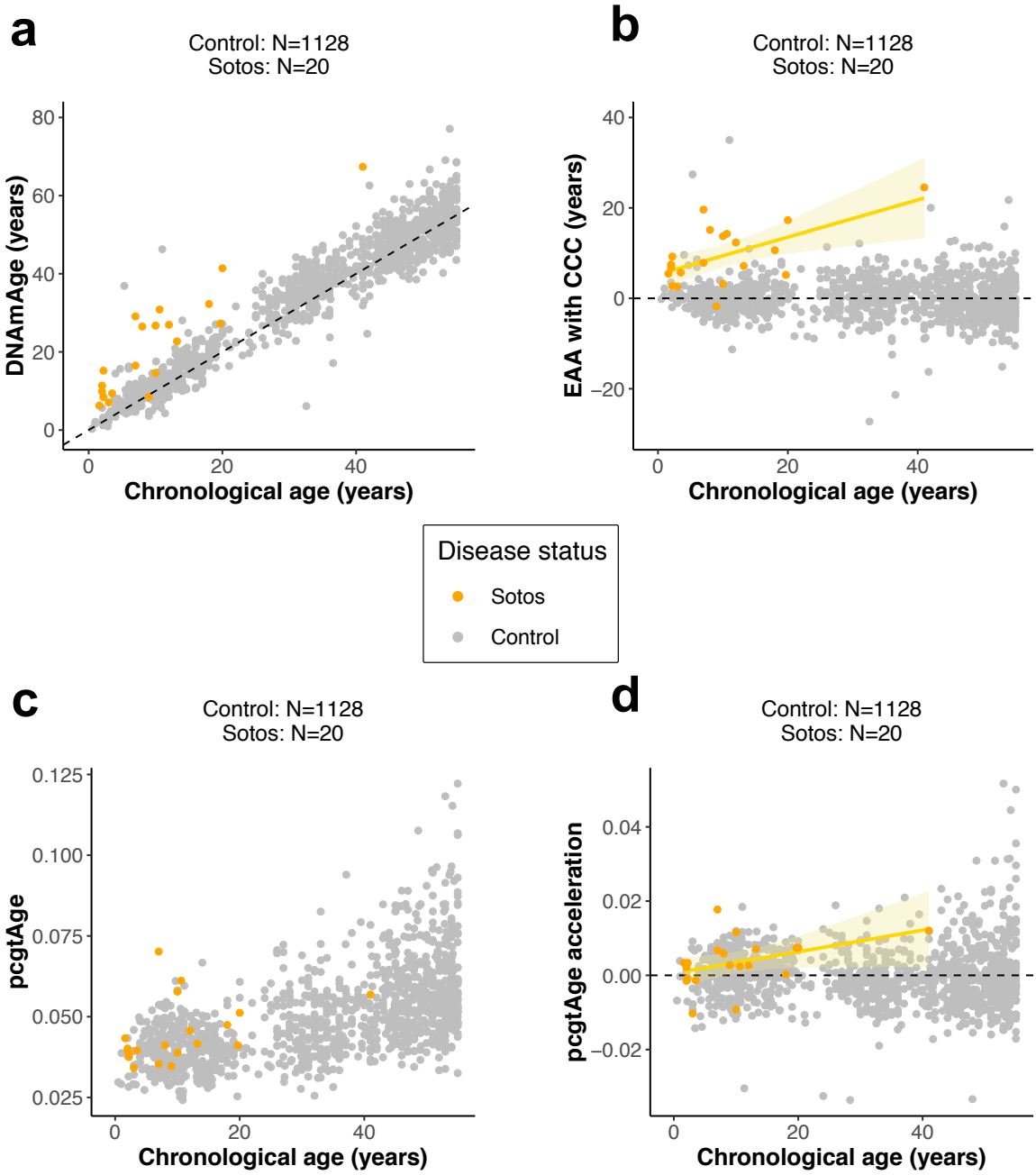



**Fig. 3.4** Sotos syndrome accelerates epigenetic ageing. **a.** Scatterplot showing the relationship between epigenetic age (*DNAmAge*) according to Horvath's model [209] and chronological age of the samples for Sotos (orange) and control (grey). Each sample is represented by one point. The black dashed line represents the diagonal to aid visualisation. **b.** Scatterplot showing the relationship between the epigenetic age acceleration (EAA) and chronological age of the samples for Sotos (orange) and control (grey). Each sample is represented by one point. The yellow line represents the linear model EAA $\sim$ Age, with the standard error shown in the light yellow shade. **c.** Scatterplot showing the relationship between the score for the epigenetic mitotic clock (*pcgtAge*) [225] and chronological age of the samples for Sotos (orange) and control (grey). Each sample is represented by one point. A higher value of *pcgtAge* is associated with a higher number of cell divisions in the tissue. **d.** Scatterplot showing the relationship between the epigenetic mitotic clock (*pcgtAge*) acceleration (with CCC) and chronological age of the samples for Sotos (orange) and control (grey). Each sample is represented by one point. The yellow line represents the linear model pcgtAge_EAA$_{\text{with CCC}}$ $\sim$ Age, with the standard error shown in the light yellow shade.

sites that gain methylation with age (i.e. become hypermethylated, 51.69%) and CpG sites that lose methylation with age (i.e. become hypomethylated, 48.31%, barplot in Fig. 3.5a), a picture that resembles previous studies [179]. It is worth mentioning that in this case fewer aDMPs were identified when compared with the full lifespan analysis presented in section 2.1.4, where the hypomethylated aDMPs were also slightly more frequent when compared with the hypermethylated ones. This highlights the importance of the age range and/or the sample size when calculating aDMPs. On the contrary, DMPs in Sotos were clearly dominated by CpGs that have lower methylation levels in individuals with the syndrome (i.e. hypomethylated, 99.27%, barplot in Fig. 3.5a), consistent with previous reports [320].

Then, I compared the intersections between the hypermethylated and hypomethylated DMPs in ageing and Sotos. Most of the DMPs were specific for ageing or Sotos (i.e. they did not overlap), but a subset of them were shared (table in Fig. 3.5a). Interestingly, there were 1728 DMPs that became hypomethylated both during ageing and in Sotos (**'Hypo-Hypo DMPs'**). This subset of DMPs is of special interest because it could be used to understand in more depth some of the mechanisms that drive hypomethylation during physiological ageing. Thus, I tested whether the different subsets of DMPs are found in specific genomic contexts (Fig. S2.4, Fig. S2.5). DMPs that are hypomethylated during ageing and in Sotos were both enriched (odds ratio >1) in enhancer categories (such as 'active enhancer 1' or 'weak enhancer 1', see the chromatin state model used, from the K562 cell line, in section 3.7) and depleted (odds ratio <1) for active transcription categories (such as 'active TSS' or 'strong transcription'), which was also observed in the 'Hypo-Hypo DMPs' subset (Fig. 3.5b). Interestingly, age-related hypomethylation in enhancers seems to be a characteristic of both humans [177, 178] and mice [165]. Furthermore, both *de novo* DNA methyltransferases (DNMT3A and DNMT3B) have been shown to bind in an H3K36me3-dependent manner to active enhancers [330], consistent with these results.

When looking at the levels of total RNA expression (depleted for rRNA) in blood, I confirmed a significant reduction in the RNA levels around these hypomethylated DMPs when compared with the controls sets (Fig. 3.5c, see section 3.7 for more details on how the control sets were defined). Interestingly, hypomethylated DMPs in both ageing and Sotos were depleted from gene bodies (Fig. 3.5b) and were located in areas with lower levels of H3K36me3 when compared with the control sets (Fig. 3.5d, Fig. S2.5). Moreover, hypomethylated aDMPs and hypomethylated Sotos DMPs where both generally enriched or depleted for the same histone marks in blood (Fig. S2.5), which adds weight to the hypothesis that they share the same genomic context and could become hypomethylated through similar molecular mechanisms.

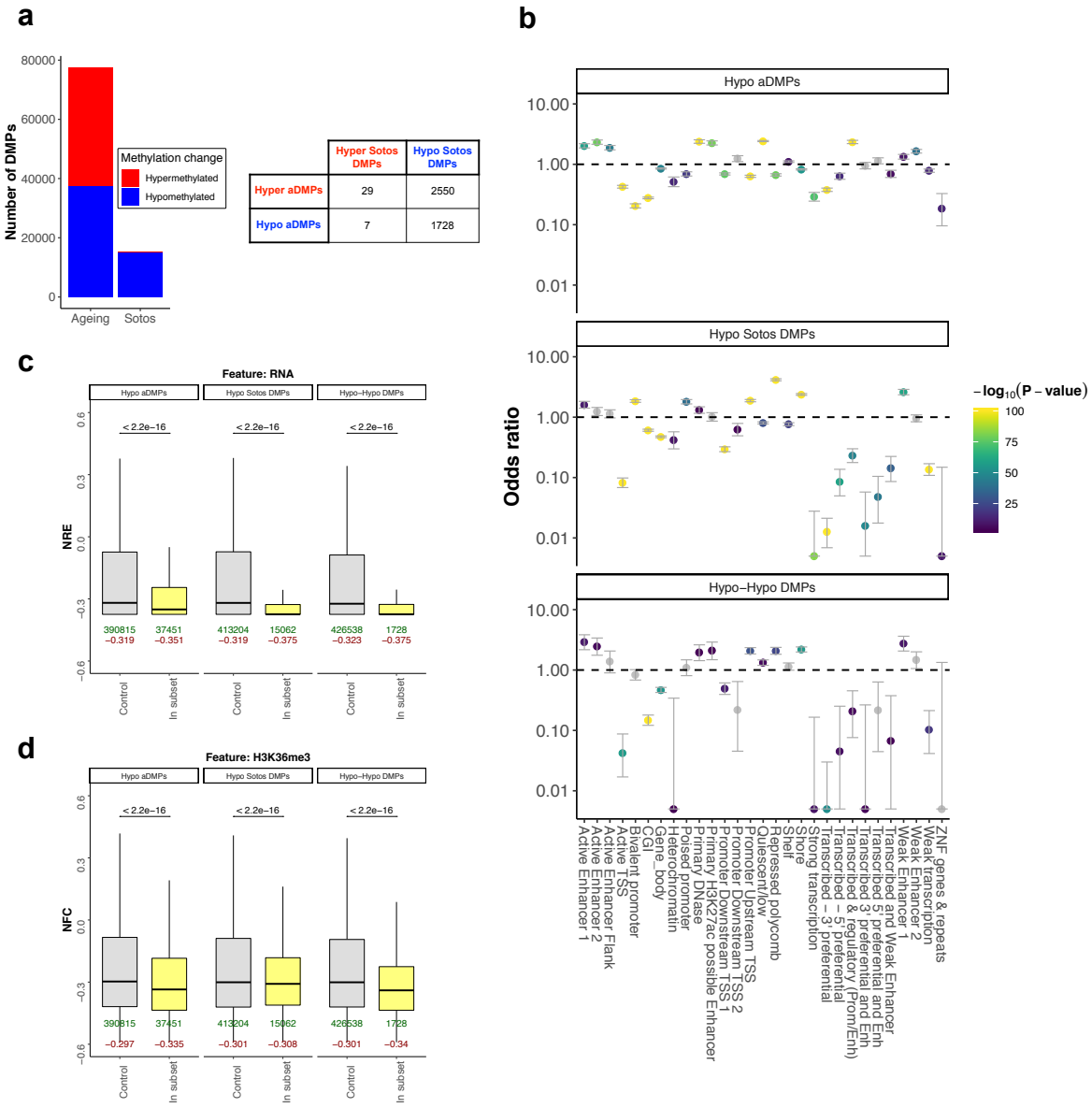



**Fig. 3.5** Comparison between the DNA methylation changes during physiological ageing and in Sotos. **a.** On the left: barplot showing the total number of differentially methylated positions (DMPs) found during physiological ageing and in Sotos syndrome. CpG sites that increase their methylation levels with age in the healthy population or those that are elevated in Sotos patients (when compared with a control) are displayed in red. Conversely, those CpG sites that decrease their methylation levels are displayed in blue. On the right: table that represents the intersection between the ageing (aDMPs) and the Sotos DMPs. The subset resulting from the intersection between the hypomethylated DMPs in ageing and Sotos is called the 'Hypo-Hypo DMPs' subset (N=1728). **b.** Enrichment for the categorical (epi)genomic features considered when comparing the different genome-wide subsets of differentially methylated positions (DMPs) in ageing and Sotos against a control (see section 3.7). The y-axis represents the odds ratio (OR), the error bars show the 95% confidence interval for the OR estimate and the colour of the points codes for $-\log_{10}$(p-value) obtained after testing for enrichment using Fisher's exact test. An OR > 1 shows that the given feature is enriched in the subset of DMPs considered, whilst an OR < 1 shows that it is found less than expected. In grey: features that did not reach significance using a significance level of $\alpha = 0.01$ after Bonferroni correction. **c.** Boxplots showing the distributions of the 'normalised RNA expression' (NRE) when comparing the different genome-wide subsets of differentially methylated positions (DMPs) in ageing and Sotos against a control (see section 3.7). NRE represents normalised mean transcript abundance in a window of $\pm$ 200 bp from the CpG site coordinate (DMP) being considered. The p-values (two-sided Wilcoxon's test, before multiple testing correction) are shown above the boxplots. The number of DMPs belonging to each subset (in green) and the median value of the feature score (in dark red) are shown below the boxplots. **d.** As in c., but showing the 'normalised fold change' (NFC) for the H3K36me3 histone modification (representing normalised mean ChIP-seq fold change for H3K36me3 in a window of $\pm$ 200 bp from the DMP being considered).

Intriguingly, I also identified a subset of DMPs (2550) that were hypermethylated during ageing and hypomethylated in Sotos (Fig. 3.5a). These **'Hyper-Hypo DMPs'** seem to be enriched for categories such as 'bivalent promoter' and 'repressed polycomb' (Fig. S2.4), which are normally associated with developmental genes [181, 184]. These categories are also a defining characteristic of the hypermethylated aDMPs, highlighting that even though the direction of the DNA methylation changes is different in some ageing and Sotos DMPs, the genomic context in which they happen is shared.

Finally, I looked at the DNA methylation patterns in the 353 **Horvath's epigenetic clock CpG sites for the Sotos samples**. For each clock CpG site, I modelled the changes of DNA methylation with age in the healthy control individuals (0-55 years) and then calculated the deviations from these patterns for the Sotos samples (Fig. 3.6, see equation 3.3). As expected, the landscape of clock CpG sites is dominated by hypomethylation in the Sotos samples, although only a small fraction of the clock CpG sites seems to be significantly affected (Fig. 3.6c). Overall, I confirmed the trends reported for the genome-wide analysis (Fig. S2.6, Fig. S2.7, Fig. S2.8). However, given the much smaller number of CpG sites to consider in this analysis, very few comparisons reached significance.

I have demonstrated that the ageing process and Sotos syndrome share a subset of hypomethylated CpG sites that is characterised by an enrichment in enhancer features and a depletion of active transcription activity. This highlights the usefulness of **developmental disorders as a model to study the mechanisms that may drive the changes in the methylome with age**, since they permit stratification of the ageing DMPs into different functional categories that are associated with alterations in the function of specific genes and hence specific molecular components of the epigenetic ageing clock.

## 3.5 Methylation Shannon entropy and the epigenetic clock

In section 2.1.5 I have discussed how Shannon entropy can be applied in the context of DNA methylation data in order to measure the genome-wide epigenetic information loss that happens during ageing. It is possible to apply a methodology similar to the one described in section 2.2.2 to compare the methylation Shannon entropy in healthy controls (0-55 years) and Sotos patients (i.e. using a linear model similar to equation 2.16, although in this case the dependent variable is the entropy value). This allows testing whether Sotos syndrome patients present genome-wide Shannon entropy acceleration i.e. deviations from the expected genome-wide Shannon entropy for their age. Despite detailed analysis, I did not find evidence that this was the case when looking genome-wide (p-value = 0.71, Fig. 3.7a,b, Fig. S2.9a).

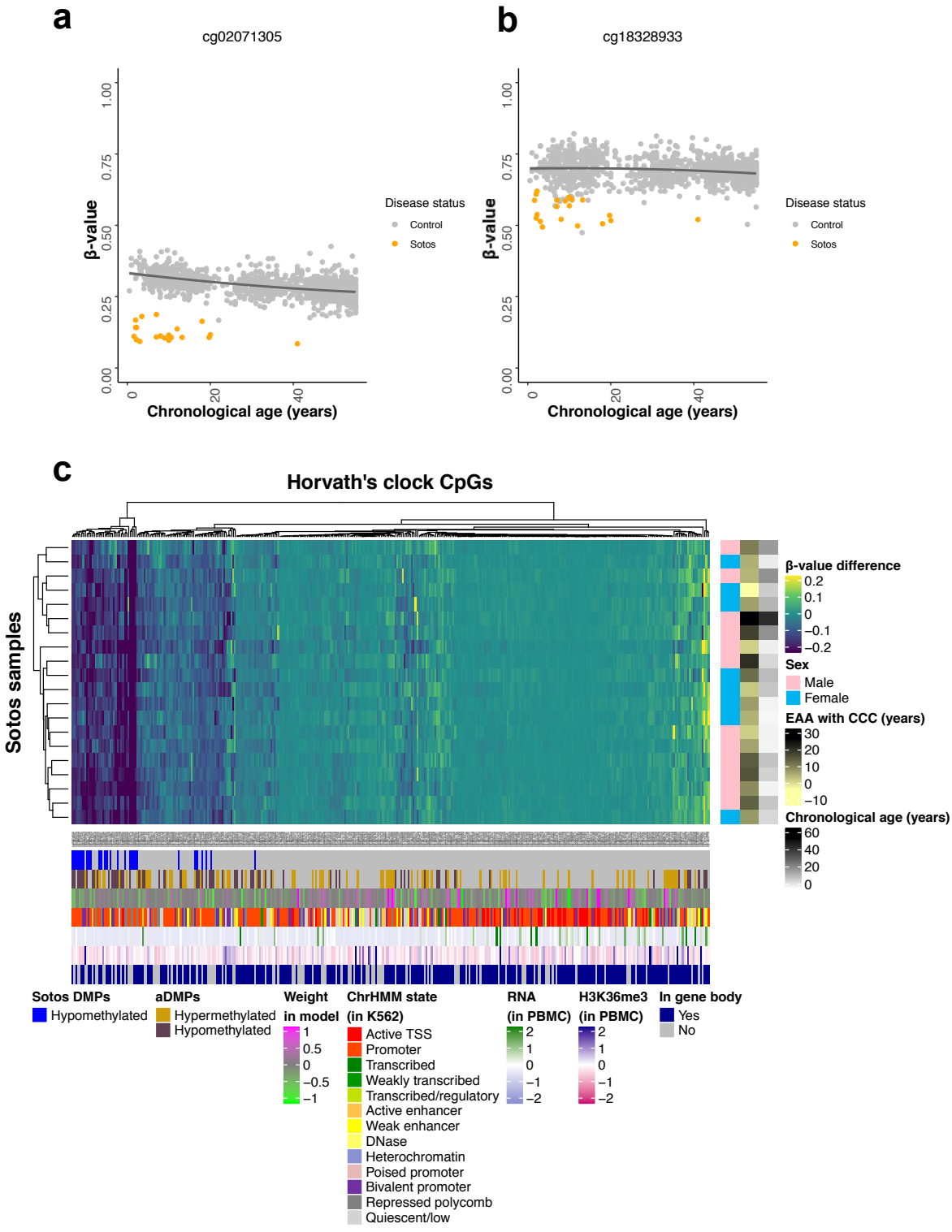



**Fig. 3.6** The landscape of Horvath's epigenetic clock CpG sites in Sotos syndrome. **a.** and **b.** DNA methylation ($\beta$-value) profiles for two of the clock CpG sites (cg02071305 and cg18328933). A linear model (displayed in dark grey, see equation 3.3) can be fixed to each CpG site to model the changes in $\beta$-value with chronological age in the controls (grey). Afterwards, the difference of the Sotos samples $\beta$-values (orange) with the controls can be estimated. **c.** Heatmap displaying the differential methylation patterns for Sotos samples (rows) when compared with controls in each one of the 353 epigenetic clock CpGs (columns). Hierarchical clustering was performed in both rows and columns. RNA refers to the 'normalised RNA expression' (NRE). H3K36me3 refers to the H3K36me3 histone modification 'normalised fold change' (NFC). aDMPs: differentially methylated positions during ageing. EAA: epigenetic age acceleration. CCC: cell composition correction. PBMC: peripheral blood mononuclear cells.

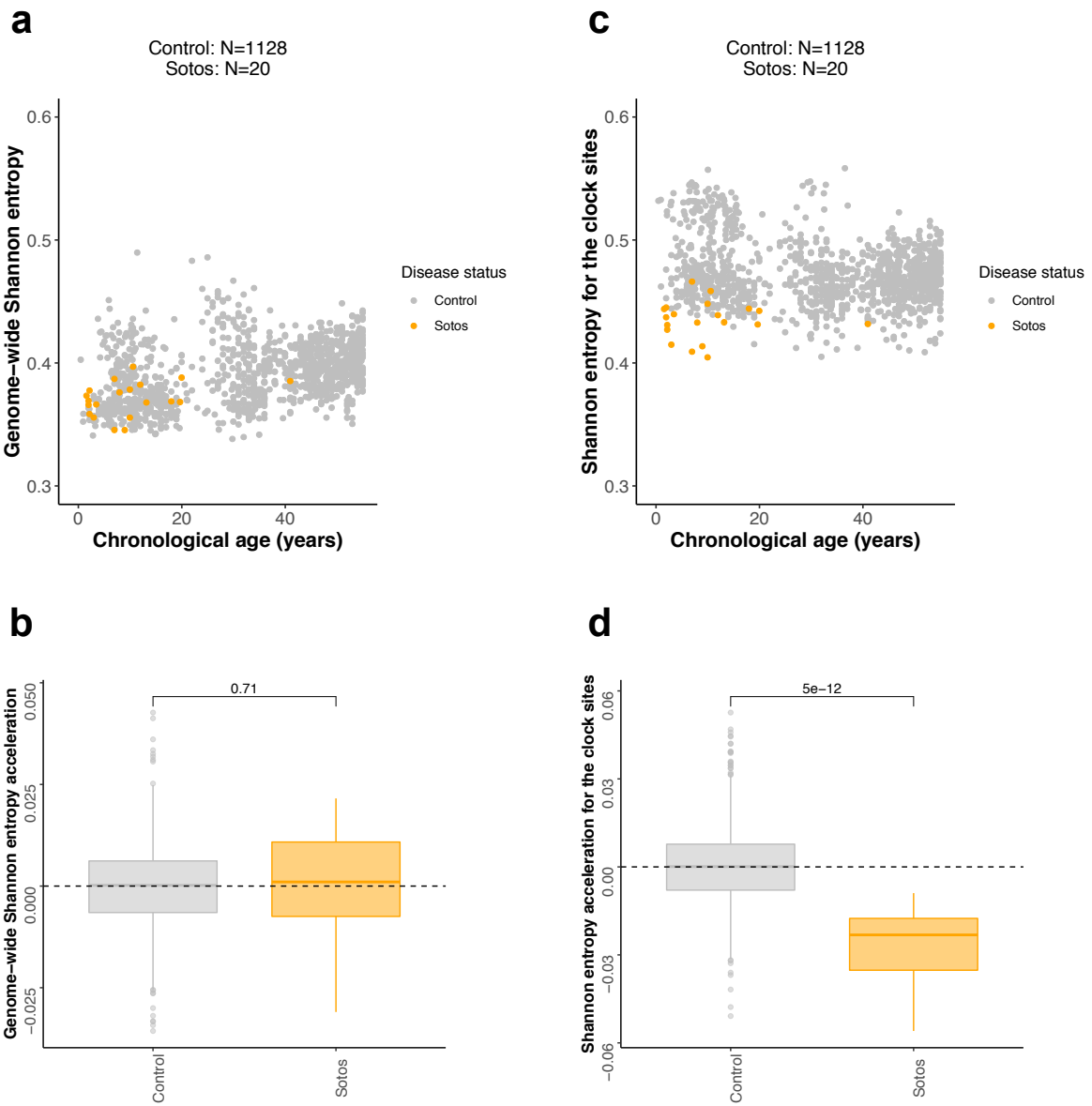



**Fig. 3.7** Analysis of methylation Shannon entropy during physiological ageing and in Sotos syndrome. **a.** Scatterplot showing the relation between genome-wide Shannon entropy (i.e. calculated using the methylation levels of all the CpG sites in the array) and chronological age of the samples for Sotos (orange) and healthy controls (grey). Each sample is represented by one point. **b.** Boxplots showing the distributions of genome-wide Shannon entropy acceleration (i.e. deviations from the expected genome-wide Shannon entropy for their age) for the control and Sotos samples. The p-value displayed on top of the boxplots was derived from a two-sided Wilcoxon's test. **c.** As in a., but using the Shannon entropy calculated only for the 353 CpG sites in the Horvath's epigenetic clock. **d.** As in b., but using the Shannon entropy calculated only for the 353 CpG sites in the Horvath's epigenetic clock.

When I considered only the 353 Horvath's epigenetic clock CpG sites for the entropy calculations, the picture was different. Shannon entropy for the 353 clock sites slightly decreased with age in the controls when I included all the batches, showing the opposite direction when compared with the genome-wide entropy (SCC $= -0.1223$, p-value = $3.8166 \cdot 10^{-5}$, Fig. 3.7c). However, when I removed the 'Europe' batch (which was an outlier even after pre-processing, Fig. S2.10), this trend was reversed and I observed a weak increase of clock Shannon entropy with age (SCC $= 0.1048$, p-value = $8.6245 \cdot 10^{-5}$). This shows that Shannon entropy calculations are very sensitive to batch effects, especially when considering a small number of CpG sites, and the results must be interpreted carefully, as already discussed in section 2.1.5.

Interestingly, the mean Shannon entropy across all the control samples was higher in the epigenetic clock sites (mean = 0.4726, Fig. 3.7c) with respect to the genome-wide entropy (mean = 0.3913, Fig. 3.7a). Sotos syndrome patients displayed a lower clock Shannon entropy when compared with the control (p-value = $5.0449 \cdot 10^{-12}$, Fig. 3.7d, Fig. S2.9b), which is probably driven by the hypomethylation of the clock CpG sites. Furthermore, this highlights that the **Horvath's epigenetic clock sites could have slightly different characteristics in terms of the methylation entropy associated with them** when compared with the genome as a whole, something that to my knowledge has not been reported before.

## 3.6 Discussion

The epigenetic ageing clock has emerged as the most accurate biomarker of the ageing process and it seems to be a conserved property in mammalian genomes [223, 230]. However, it is still unknown whether the age-related DNA methylation changes measured are functional at all or whether they are related to some fundamental process of the biology of ageing. Developmental disorders in humans represent an interesting framework to look at the biological effects of mutations in genes that are fundamental for the integrity of the epigenetic landscape and other core processes, such as growth or neurodevelopment [248, 316]. Therefore, using a reverse genetics approach, I aimed to identify genes that disrupt aspects of the behaviour of the epigenetic ageing clock in humans.

Most of the studies have looked at the epigenetic ageing clock using Horvath's epigenetic clock [209], and I decided to employ it as a tool to measure the epigenetic age of my samples. The results from the screen strongly suggest that Sotos syndrome accelerates epigenetic ageing. Sotos syndrome is caused by loss-of-function mutations in the NSD1 gene [320, 329], which encodes a histone H3 lysine 36 (H3K36) methyltransferase. This leads to a phenotype

which can include pre-natal and post-natal overgrowth, facial gestalt, advanced bone age, developmental delay, higher cancer predisposition and, in some cases, heart defects [328]. Remarkably, many of these characteristics could be interpreted as ageing-like, identifying **Sotos syndrome as a potential human model of accelerated physiological ageing**.

NSD1 catalyses the addition of either monomethyl (H3K36me) or dimethyl groups (H3K36me2) and indirectly regulates the levels of trimethylation (H3K36me3) by altering the availability of the monomethyl and dimethyl substrates for the trimethylation enzymes (SETD2 in humans, whose mutations cause a 'Sotos-like' overgrowth syndrome ) [331, 332]. H3K36 methylation has a complex role in the regulation of transcription [331] and has been shown to regulate nutrient stress response in yeast [333]. Moreover, experiments in model organisms (yeast and worm) have demonstrated that **mutations in H3K36 methyltranferases decrease lifespan and, remarkably, mutations in H3K36 demethylases increase it** [334–336].

In humans, DNA methylation patterns are established and maintained by three conserved enzymes: the maintenance DNA methyltransferase DNMT1 and the *de novo* DNA methyltransferases DNMT3A and DNMT3B [337]. Both DNMT3A and DNMT3B contain PWWP domains that can read the H3K36me3 histone mark [338, 339]. Therefore, the H3K36 methylation landscape can influence DNA methylation levels in specific genomic regions through the recruitment of the *de novo* DNA methyltransferases. Mutations in the PWWP domain of DNMT3A impair its binding to H3K36me2 and H3K36me3 and cause an undergrowth disorder in humans (microcephalic dwarfism) [340]. This redirects DNMT3A, which is normally targeted to H3K36me2 and H3K36me3 throughout the genome, to DNA methylation valleys (DMVs, a.k.a DNA methylation canyons), which become hypermethylated [340]; a phenomenon that also seems to happen during physiological ageing in humans [177, 167, 168] and mice [165]. DMVs are hypomethylated domains conserved across cell types and species, often associated with Polycomb-regulated developmental genes and marked by bivalent chromatin (with H3K27me3 and H3K4me3) [341–344]. Therefore, I suggest a model (Fig. 3.8) where the **reduction in the levels of H3K36me2 and/or H3K36me3, caused by a proposed decrease in H3K36 methylation maintenance during ageing or NSD1 function in Sotos syndrome, could lead to hypomethylation in many genomic regions (because DNMT3A is recruited less efficiently) and hypermethylation in DMVs (because of the higher availability of DNMT3A)**. Indeed, I observe enrichment for categories such as 'bivalent promoter' or 'repressed polycomb' in the hypermethylated DMPs in Sotos and ageing (Fig. S2.4), which is also supported by higher levels of Polycomb Repressing Complex 2 (PRC2, represented by EZH2) and H3K27me3, the mark deposited by

PRC2 (Fig. S2.5).This is also consistent with the results obtained for the epigenetic mitotic clock [225], where I observe a trend towards increased hypermethylation of Polycomb-bound regions in Sotos patients. Furthermore, it is worth mentioning that a mechanistic link between PRC2 recruitment and H3K36me3 has also been unravelled via the Tudor domains of some polycomb-like proteins [345, 346].

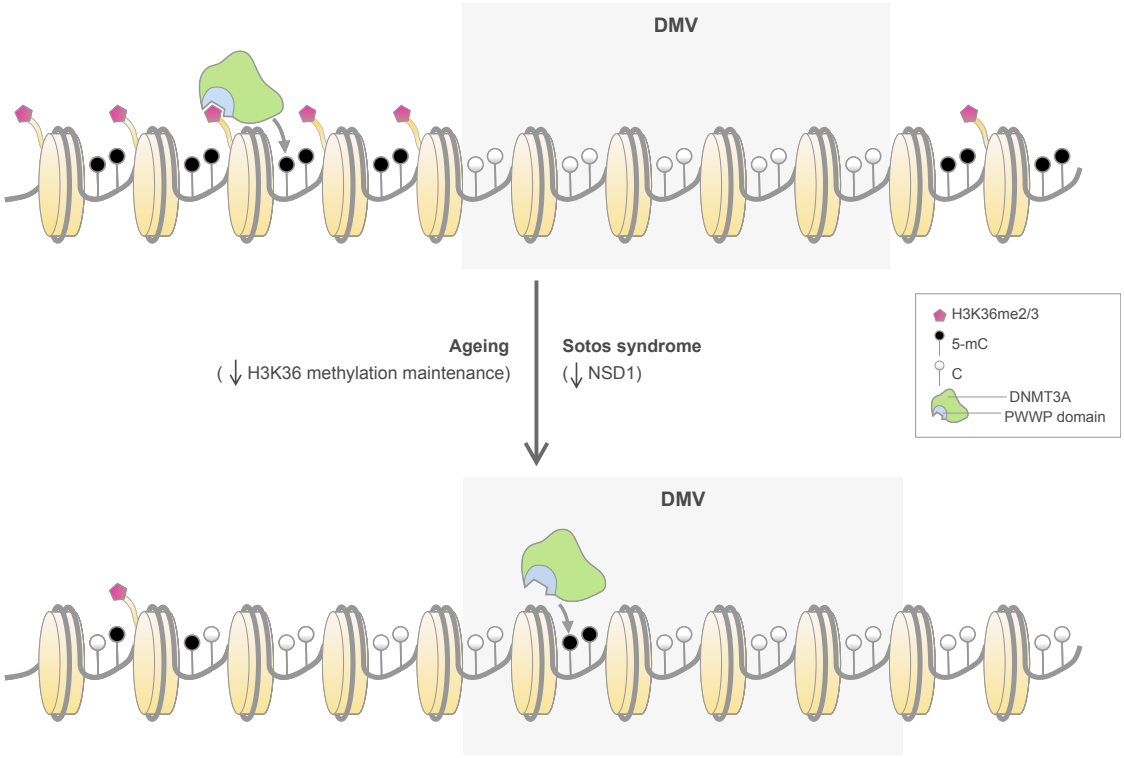



**Fig. 3.8** Proposed model that highlights the role of H3K36 methylation maintenance on epigenetic ageing. The H3K36me2/3 mark allows recruiting *de novo* DNA methyltransferases DNMT3A (in green) and DNMT3B (not shown) through their PWWP domain (in blue) to different genomic regions (such as gene bodies or pericentric heterochromatin) [339, 347, 348], which leads to the methylation of the cytosines in the DNA of these regions (5mC, black lollipops). On the contrary, DNA methylation valleys (DMVs) are conserved genomic regions that are normally found hypomethylated and associated with Polycomb-regulated developmental genes [341–344]. During ageing, the H3K36 methylation machinery could become less efficient at maintaining the H3K36me2/3 landscape. This would lead to a relocation of *de novo* DNA methyltransferases from their original *genomic reservoirs* (which would become hypomethylated) to other non-specific regions such as DMVs (which would become hypermethylated and potentially lose their normal boundaries), with functional consequences for the tissues. This is also partially observed in patients with Sotos syndrome, where mutations in NSD1 potentially affect H3K36me2/3 patterns and accelerate the epigenetic ageing clock as measured with the Horvath's model [209]. Given that DNMT3B is enriched in the gene bodies of highly transcribed genes [339] and that I found these regions depleted in the differential methylation analysis, I hypothesise that the hypermethylation of DMVs could be mainly driven by DNMT3A instead. However, it is important to mention that my analysis does not discard a role of DNMT3B during epigenetic ageing.

A recent preprint has shown that loss-of-function mutations in DNMT3A, which cause Tatton-Brown-Rahman overgrowth syndrome, also lead to a higher ticking rate of the epigenetic ageing clock [349]. They also report positive epigenetic age acceleration in Sotos syndrome and negative acceleration in Kabuki syndrome, consistent with my results. Furthermore, they observe a DNA methylation signature in the DNMT3A mutants characterised by widespread hypomethylation, with a modest enrichment of DMPs in regions upstream of the transcription start site, shores and enhancers [349], which I also detect in the 'Hypo-Hypo DMPs' (those that become hypomethylated both during physiological ageing and in Sotos). Therefore, **the hypomethylation observed in the 'Hypo-Hypo DMPs' is consistent with a reduced methylation activity of DNMT3A**, which in my analysis could be a consequence of the decreased recruitment of DNMT3A to genomic regions that have lost H3K36 methylation (Fig. 3.8).

Interestingly, H3K36me3 is required for the selective binding of the *de novo* DNA methyltransferase DNMT3B to the bodies of highly transcribed genes [339]. Furthermore, DNMT3B loss reduces gene-body methylation, which leads to intragenic spurious transcription (a.k.a cryptic transcription) [118]. An increase in this so-called cryptic transcription seems to be a conserved feature of the ageing process [335]. Therefore, the changes observed in the 'Hypo-Hypo DMPs' could theoretically be a consequence of the loss of H3K36me3 and the concomitant inability of DNMT3B to be recruited to gene bodies. However, the 'Hypo-Hypo DMPs' were depleted for H3K36me3, active transcription and gene bodies when compared with the rest of the probes in the array (Fig. 3.5b-d), prompting me to suggest that the DNA methylation changes observed are likely mediated by DNMT3A instead (Fig. 3.8). Nevertheless, it is worth mentioning that the different biological replicates for the blood H3K36me3 ChIP-seq datasets were quite heterogeneous and that the absolute difference in the case of the hypomethylated Sotos DMPs, although significant due to the big sample sizes, is quite small. Thus, I cannot exclude the existence of this mechanism during human ageing and an exhaustive study on the prevalence of cryptic transcription in humans and its relation to the ageing methylome should be carried out.

H3K36me3 has also been shown to guide deposition of the $N^6$-methyladenosine mRNA modification ($m^6A$), an important post-transcriptional mechanism of gene regulation [350]. Interestingly, a decrease in overall $m^6A$ during human ageing has been previously reported in PBMCs [351], suggesting another biological route through which an alteration of the H3K36 methylation landscape could have functional consequences for the organism.

Because of the way that the Horvath epigenetic clock was trained [209], it is likely that its constituent 353 CpG sites are a **low-dimensional representation of the different**

**genome-wide processes that are eroding the epigenome with age**. My analysis has shown that these 353 CpG sites are characterised by a higher Shannon entropy when compared with the rest of the genome, which is dramatically decreased in the case of Sotos patients. This could be related to the fact that Horvath's clock CpGs are enriched in regions of bivalent chromatin (marked by H3K27me3 and H3K4me3), conferring a more dynamic or plastic regulatory state with levels of DNA methylation deviated from the collapsed states of 0 or 1. Interestingly, EZH2 (part of Polycomb Repressing Complex 2, responsible for H3K27 methylation) is an interacting partner of DNMT3A and NSD1, with mutations in NSD1 affecting the genome-wide levels of H3K27me3 [352]. Furthermore, Kabuki syndrome was weakly identified in my screen as having an epigenome younger than expected, which could be related to the fact that they show post-natal dwarfism [317, 319]. Kabuki syndrome is caused by loss-of-function mutations in KMT2D [317, 319], a major mammalian H3K4 mono-methyltransferase [353]. Additionally, H3K27me3 and H3K4me3 levels can affect lifespan in model organisms [151]. It will be interesting to test whether bivalent chromatin is a general feature of multi-tissue epigenetic ageing clocks.

Thus, **DNMT3A, NSD1 and the machinery in control of bivalent chromatin (such as EZH2 and KMT2D) contribute to an emerging picture on how the mammalian epigenome is regulated during ageing**, which could open new avenues for anti-ageing drug development. Mutations in these proteins lead to different developmental disorders with impaired growth defects [316], with DNMT3A, NSD1 and potentially KMT2D also affecting epigenetic ageing. Interestingly, EZH2 mutations (which cause Weaver syndrome, Table 3.1) do not seem to affect the epigenetic clock in my screen. However, this syndrome has the smallest number of samples ($N = 7$) and this could limit the power to detect any changes.

My screen has also revealed that **Rett syndrome and fragile X syndrome (FXS) could potentially have an accelerated epigenetic age**. It is worth noting that FXS is caused by an expansion of the CGG trinucleotide repeat located in the 5' UTR of the FMR1 gene [321]. Interestingly, Huntington's disease, caused by a trinucleotide repeat expansion of CAG, has also been shown to accelerate epigenetic ageing of human brain [222], pointing towards trinucleotide repeat instability as an interesting molecular mechanism to look at from an ageing perspective. It is important to notice that the conclusions for Rett syndrome, FXS and Kabuki syndrome were very dependent on the age range used in the healthy control (Fig. S2.2) and these results must therefore be treated with caution.

This study has several **limitations that I tried to address in the best possible way**. First of all, given that DNA methylation data for patients with developmental disorders is relatively rare, some of the sample sizes were quite small. It is thus possible that some of the

other developmental disorders assessed are epigenetically accelerated but I lack the power to detect this. Furthermore, people with the disorders tend to get sampled when they are young i.e. before reproductive age. Horvath's clock adjusts for the different rates of change in the DNA methylation levels of the clock CpGs before and after adult/reproductive age (20 years in humans) [209], but this could still have an effect on the predictions, especially if the control is not properly age-matched. My solution was to discard those developmental disorders with less than 5 samples and I required them to have at least 2 samples with an age $\geq$ 20 years, which reduced the list of final disorders included to the ones listed in Table 3.1.

Future studies should increase the sample size and follow the patients during their entire lifespan in order to confirm these findings. Furthermore, it would be interesting to identify mutations that affect, besides the mean, the variance of epigenetic age acceleration, since changes in methylation variability at single CpG sites with age have been associated with fundamental ageing mechanisms [177]. Finally, testing the influence of H3K36 methylation on the epigenetic clock and lifespan in mice will provide deeper mechanistic insights.

## 3.7 Additional methods

### Sample generation and annotation

I collected DNA methylation data generated with the Illumina Infinium HumanMethylation450 BeadChip (450K array) from human blood. In the case of the developmental disorder samples, I combined public data with data generated in-house by my collaborators in Canada (Table S2.1, Fig. S2.1). The wet-lab protocols used in the public datasets can be found in their respective GEO repositories. DNA methylation data from my Canadian collaborators was generated according to the manufacturer's protocol [354, 355].

Basic metadata (including the chronological age) was also stored. All the mutations in the developmental disorder samples were manually curated using Variant Effect Predictor [356] in the GRCh37 (hg19) human genome assembly. Those samples with a variant of unknown significance that had the characteristic DNA methylation signature of the disease were also included (they are labelled as 'YES_predicted' in Fig. S2.1). In the case of fragile X syndrome (FXS), only male samples with full mutation (>200 repeats) [321] were included in the final screen. As a consequence, only samples with a clear molecular and clinical diagnosis were kept for the final screen.

## Identifying differentially methylated positions in Sotos syndrome

Following a strategy similar to the one outlined in section 2.1.4, I identified those array probes that were differentially methylated in patients with Sotos syndrome. I compared the Sotos samples (N=20) against the internal control samples (N=51) from the same dataset (GSE74432) [320], fitting the following linear model to each one of the array probes:

$$Beta \sim Disease\_status + Age + Sex + Gran + CD4T + CD8T + B + Mono + NK + PC1 + ... + PC17 \quad (3.1)$$

where *Beta* is the $\beta$-value for the array probe being evaluated; *Disease_status* indicates whether a sample comes from a healthy individual (0) or a Sotos syndrome patient (1); *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).

P-values and regression coefficients were extracted for the *Disease_status* covariate. I selected as my final Sotos DMPs those CpG probes that survived the analysis after Bonferroni multiple testing correction with a significance level of $\alpha = 0.01$.

## (Epi)genomic annotation of the CpG sites

Different (epi)genomic features were extracted for the CpG sites of interest. All the data were mapped to the *hg19* assembly of the human genome. The **continuous features** were calculated by extracting the mean value in a window of $\pm$ 200 bp from the CpG site coordinate using the *pyBigWig* package [357]. I chose this window value based on the methylation correlation observed between neighbouring CpG sites in previous studies [358]. The continuous features included (Fig. S2.11):

- ChIP-seq data from ENCODE (histone modifications from peripheral blood mononuclear cells or PBMC; EZH2, as a marker of Polycomb Repressing Complex 2 binding, from B cells; RNF2, as a marker of Polycomb Repressing Complex 1 binding, from the K562 cell line). I obtained Z-scores (using the *scale* function in R) for the values of 'fold change over control' as calculated in ENCODE [93]. When needed, biological replicates of the same feature were aggregated by taking the mean of the Z-scores in order to obtain the 'normalised fold change' (NFC).

- ChIP-seq data for LaminB1 (GSM1289416, quantified as 'normalised read counts' or NRC) and Repli-seq data for replication timing (GSM923447, quantified as 'wavelet-transformed signals' or WTS). I used the same data from the IMR90 cell line as in [359].

- Total RNA-seq data (rRNA depleted, from PBMC) from ENCODE. I calculated Z-scores after aggregating the 'signal of unique reads' (*sur*) for both strands (+ and - ) in the following manner:

$$RNA_i = \log_2(1 + sur_{i+} + sur_{i-}) \tag{3.2}$$

  where $RNA_i$ represents the RNA signal (that then needs to be scaled to obtain the 'normalised RNA expression' or NRE) for the *i*th CpG site.

The **categorical features** were obtained by looking at the overlap (using the *pybedtools* package) [360] of the CpG sites with the following:

- Gene bodies, from protein-coding genes as defined in the basic gene annotation of GENCODE release 29 [361].

- CpG islands (CGIs) were obtained from the UCSC Genome Browser [362]. Shores were defined as regions 0 to 2 kb away from CGIs in both directions and shelves as regions 2 to 4 kb away from CGIs in both directions as previously described [358, 363].

- Chromatin states were obtained from the K562 cell line in the Roadmap Epigenomics Project (based on imputed data, 25 states, 12 marks) [364]. A visualisation for the association between chromatin marks and chromatin states can be found in [365]. When needed for visualisation purposes, the 25 states were manually collapsed to a lower number of them.

I compared the different genomic features for each one of the subsets of CpG sites (hypomethylated aDMPs, hypomethylated Sotos DMPs, etc.) against a control set. This control set was composed of all the probes from the background set from which I removed the subset that I was testing. In the case of the comparisons against the 353 Horvath clock CpG sites, a background set of the 21368 (21K) CpG probes used to train the original Horvath model [209] was used. In the case of the genome-wide comparisons for ageing and Sotos syndrome, a background set containing all 428266 probes that passed my pre-processing pipeline was used (see section 2.1.2).

For each continuous feature, the feature score distributions for a given subset of CpG sites and the control set were compared using the non-parametric two-sided Wilcoxon's test. For each categorical feature, I first created a $2x2$ contingency table, with the two variables indicating whether a given CpG site overlaps with the categorical feature under consideration (Yes/No) and whether the CpG site is in the subset (e.g. hypomethylated aDMPs) being considered (Yes/No). Using Fisher's exact test (as implemented in the *fisher.test* function in R) I calculated the p-value and the odds ratio (OR), which allows determining whether the categorical feature under consideration is enriched in the CpGs subset.

## Differences in the epigenetic clock CpGs $\beta$-values for Sotos syndrome

To compare the $\beta$-values of the Horvath clock CpG sites between the healthy samples and Sotos samples I fitted the following linear model to each array probe from the Horvath's epigenetic clock (353 in total) in the healthy individuals samples (Fig. 3.6a,b):

$$Beta \sim Age + Age^2 + Sex + Gran + CD4T + CD8T + B + Mono + NK + PC1 + ... + PC17 \tag{3.3}$$

where *Beta* is the $\beta$-value for the clock 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). The $Age^2$ covariate allows accounting for non-linear relationships between chronological age and the $\beta$-values.

Finally, I calculated the difference between the $\beta$-values in Sotos samples and the predictions from the models in equation 3.3 and displayed these differences in an annotated heatmap (Fig. 3.6c).

# Chapter 4

# Technological aspects

> 'It is perfectly true, as the philosophers say, that life must be understood backwards. But they forget the other proposition, that it must be lived forwards.'
>
> Søren Kierkegaard, 1843 [366]

### Declaration

The content of this chapter was joint work with Tom Stubbs, with whom I designed and developed cuRRBS. Nevertheless, almost all the text, code and plots here presented were produced by myself. Additionally, I would like to recognise the contributions of Janet M. Thornton and Wolf Reik (who helped designing the study), Antonio J. M. Ribeiro (who implemented the last version of cuRRBS to make it more computationally efficient) and Felix Krueger (who processed the RRBS datasets). All of them also helped in the revision of the final text. This work has been published in the journal *Nucleic Acids Research* [363].

## 4.1 Background

With the advent of next-generation sequencing, scientists are studying the biology of life at unprecedented resolution [367]. 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) [368–370], 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 [371, 372]. 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 [373].

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 [141, 374–377]. 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) [378], a recombinant binding protein (e.g. methyl-CpG-binding domains or MBD) [379], covalent biotin tagging [380], capture probes/baits for the sites of interest [381–383], array-based approaches (e.g. 27K, 450K and EPIC arrays in human) [249–251, 384] or PCR-based approaches [385–390]. These methods have many limitations, including enrichment biases, complex protocols and difficulties in quantification [374, 375].

Current evidence shows that depletion-based methods do not have enrichment biases, tend to be simpler and are more readily quantifiable [374, 376]. 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 [391–395]. 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) [396, 397], although enzymes such as BglII [398], XmaI [399], Taq$^{\alpha}$I [400, 401], MspJI [402] , ApeKI [403], HpyCH4IV or HpaII [404] have also been used. RRBS has proven extremely useful for cost-effective, global studies of DNA methylation [210, 396, 400, 405], 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 [406].

In the context of epigenetic clocks, most studies have used methylation arrays in humans [209, 208, 207] and MspI-based RRBS in mice, dogs and wolves [210–213, 185]. 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 [396]. Nevertheless, it is known that many age-related changes in the methylome occur in other genomic regions (such as enhancers) [177, 178, 165, 314], 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 [226]. 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 [209, 407–412]. 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 [399, 401]. 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).

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 [396] (Fig. 4.2b).

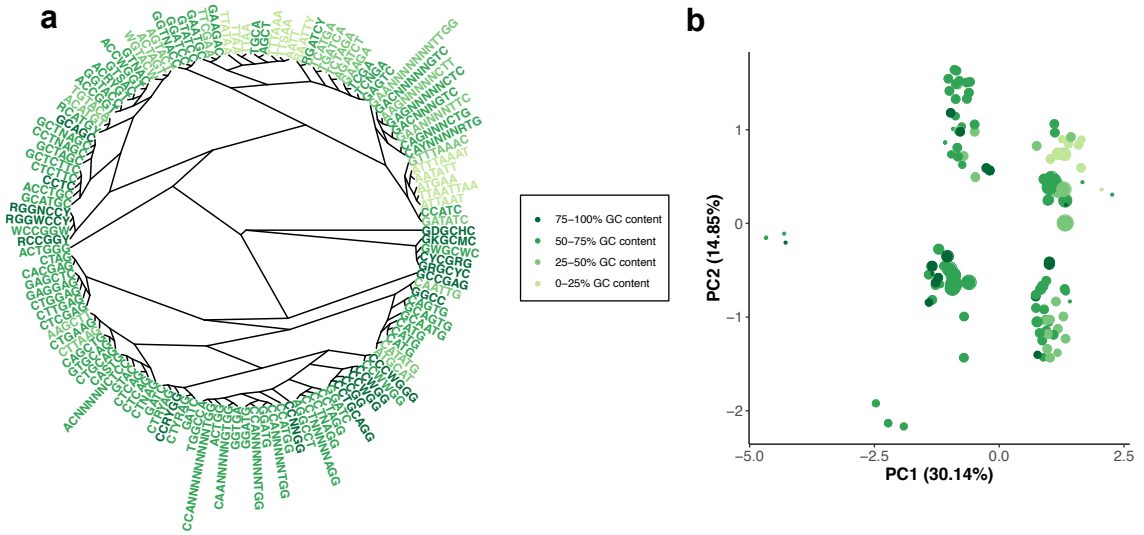



**Fig. 4.1** The landscape of restriction enzyme motifs. **a.** Phylogenetic analysis of the motifs that are recognised by the different commercially-available restriction enzymes which are insensitive to CpG methylation. Each sequence represents a different isoschizomer family considered in this study.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

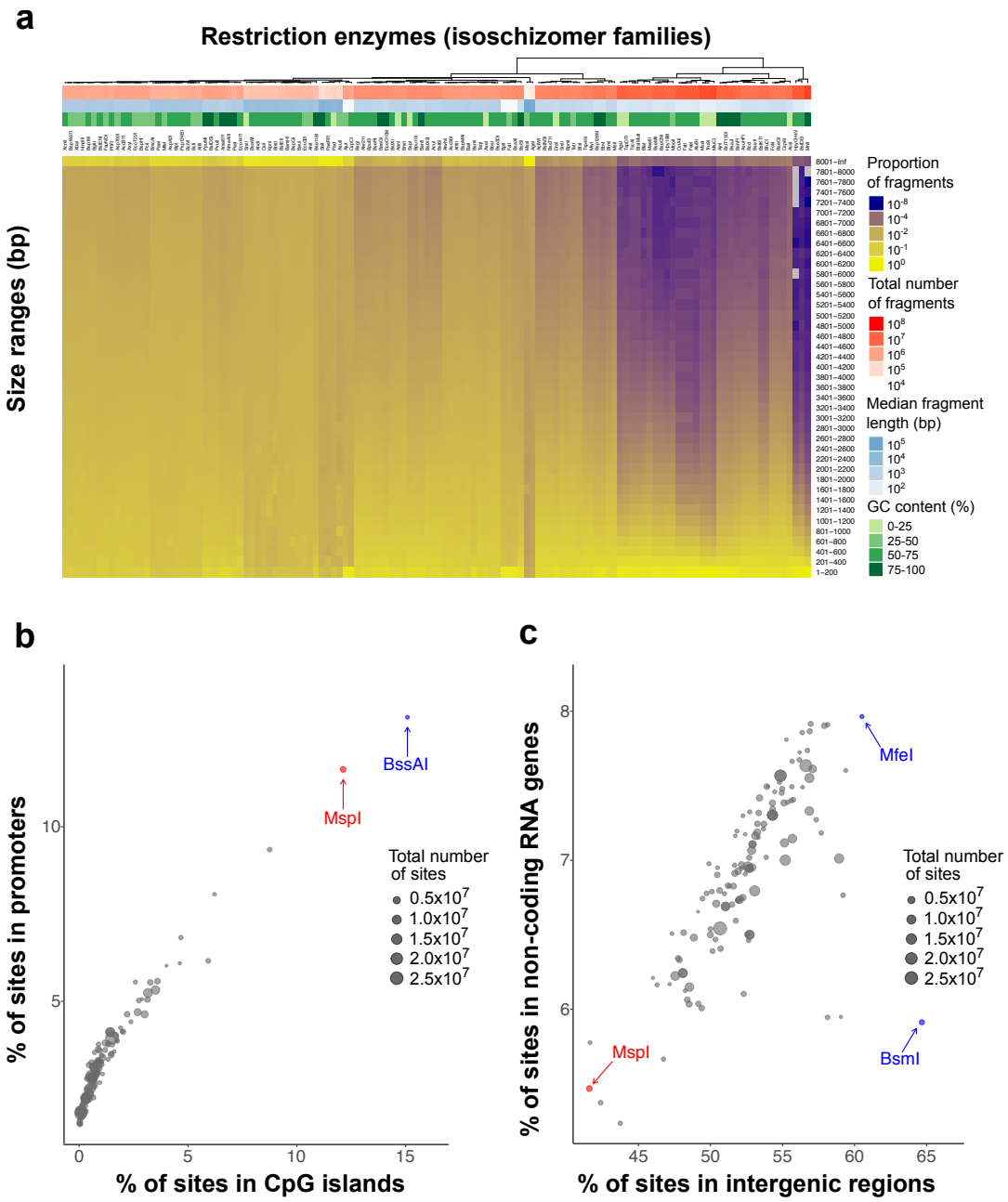



**Fig. 4.2** Restriction enzyme digestion as a tool for genomic enrichment. **a.** Heatmap showing the fragment length distributions generated by different restriction enzymes in the human genome (hg38). Each column represents the distribution for an isoschizomer family of restriction enzymes that contains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

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 in the human genome, 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 [391, 393–395, 399, 400, 403, 404]. However, to our knowledge, there is currently no computational method that systematically explores the capacity of all commercially-available restriction enzymes to generate 'personalised' reduced-representations of the genome whilst minimising the experimental cost (Fig. S3.3).

## 4.3 cuRRBS: customised Reduced Representation Bisulfite Sequencing

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

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

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

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

where $n$ is the total number of sites of interest, $w_i$ is the weight of the $i$th site of interest and $\gamma_i$ is 1 if the $i$th 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 $\delta$ 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 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,

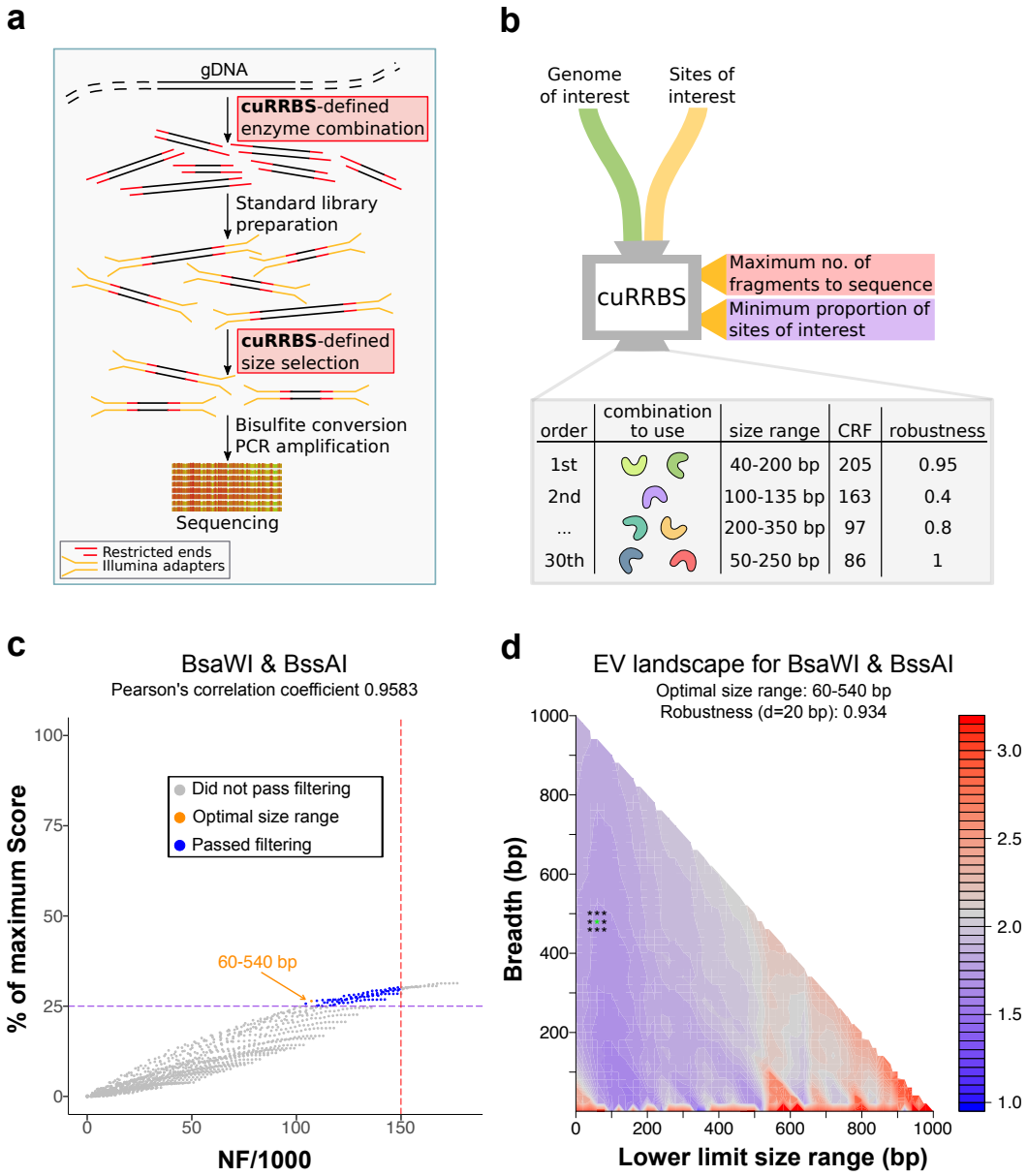



**Fig. 4.3** cuRRBS overview. **a.** Outline of an RRBS protocol. Highlighted are the two steps that would be modified according to the output produced by cuRRBS (i.e. the restriction enzymes used for the genomic digestion and the size selection). Legend is displayed on the bottom left. **b.** Schematic 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 [407]. *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 [407]. *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.

such as plants [413], which currently lack an MspI-based RRBS protocol, owing to the enzyme's CHG methylation sensitivity [414].

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 [408], Horvath's epigenetic clock [209], transcription factor binding sites that are affected by DNA methylation [410, 412], imprinted loci [407], CpGs found in the exon-intron boundaries [411] and CHG sites that are differentially methylated between different arabidopsis accessions [409] (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 [393]. 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) [399, 401]. By evaluating the predictive power of cuRRBS in these two 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

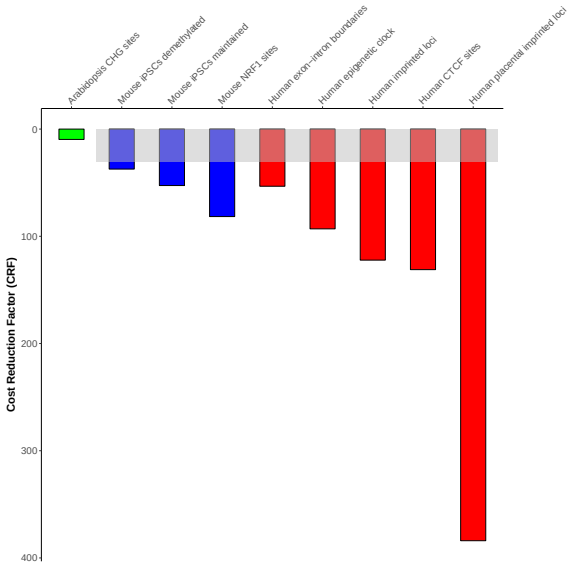



**Fig. 4.4** Running cuRRBS in different biological systems. Barplot showing the values for the *Cost Reduction Factor* (*CRF*) in the different biological systems that were tested (see Fig. S3.5) [209, 407–412]. The colours in the bars represent the different species interrogated (green: *Arabidopsis 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).

enzyme XmaI [399]. 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 [399] (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 i.e. 4.72%), 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) [399]. 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 [401]. 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 [401]. cuRRBS predicted with high accuracy the CpG sites that were observed in this double enzyme experiment (Fig. 4.5c). In addition, the results for sensitivity and specificity were very similar to the ones reported for the XmaI-RRBS dataset (Fig. 4.5d). Therefore, we conclude that cuRRBS produces robust predictions for the sites of interest that will be sequenced in RRBS protocols both for single and double enzyme combinations independent of the genome under study.

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

## 4.6 Conclusions and future directions

cuRRBS provides a new framework that allows the user to optimise RRBS for the biological system of interest by using novel combinations of restriction enzymes. Therefore, cuRRBS makes the study of DNA methylation more affordable across all species for which genomic sequences are available. Furthermore, it can open the door to the design of future studies in a clinical context [400], 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

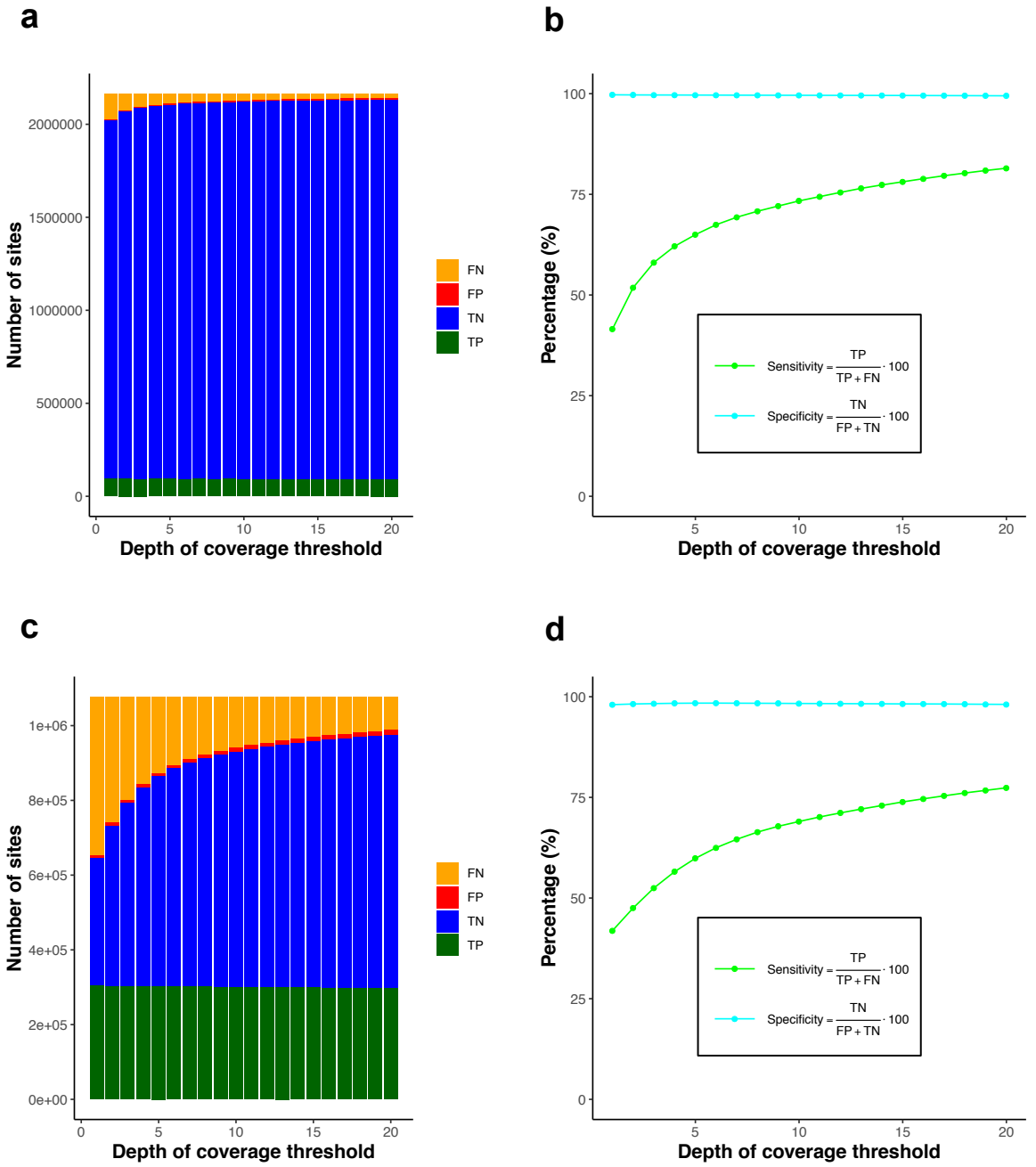



**Fig. 4.5** Experimental validation of cuRRBS. **a.** Barplots showing the number of true positives (TP, in green), true negatives (TN, in blue), false positives (FP, in red) and false negatives (FN, in orange) when comparing cuRRBS theoretical prediction with the actual XmaI-RRBS experimental data [399] (see section 4.7 for more details). 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 [399]. 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 [401]. 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 [401].

contain higher numbers of enzymes, which could theoretically allow targeting the sites of interest even more efficiently [393]. Moreover, there are several methods that are able to impute DNA methylation levels in sites that are not covered experimentally [358, 415]. 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 [416, 417] or to optimise chromosome conformation capture techniques [418, 419]. 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 [420, 421]. 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) [422], with the exception of CpG islands (CGIs), which were extracted from the UCSC Genome Browser [362]. 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 [358]. 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.* [423]. Genomic annotation for the CGIs in the mouse genome (mm10) was also obtained from the UCSC Genome Browser [362]. All annotations were handled using the *pybedtools* library [360, 424].

## 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* [425]. 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 [426]. However, it is also possible for the user to define a personalised set of enzymes by providing a self-generated annotation file. This can be useful, for instance, to reduce the chance of any star activity in the reported cuRRBS protocols.

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

| 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 ($\delta$) | 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.

## 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 [426] 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 [209] and were lifted over to hg38 [427] before running cuRRBS.
- Canonical and placental imprints in human. These loci were obtained from Hanna *et al.* [407]. The sites were lifted over to hg38 [427] 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 [410]. We scanned the peaks for high-scoring motifs according to the CTCF JASPAR model [428]. 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 [410].

- 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 [408], 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 [412], where peaks were called using Peakzilla [429]. 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 [428]. 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 [428].

- CHG sites in *Arabidopsis thaliana*. Non-CpG DMRs arising from the epigenomic diversity between *Arabidopsis thaliana* accessions were obtained from Kawakatsu *et al.* [409]. 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 [209]. All the site annotation files can be found in my GitHub repository [426]

## 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 [399] and MspI&Taq$^{\alpha}$I -RRBS data generated on the Illumina HiSeq platform [401] 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) [430]. 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 [399, 401] (see section 4.5). We ran cuRRBS fixing the theoretical size ranges tested to the ones reported in the publications [399, 401] and we used as our sites of interest the CpGs that overlapped with CpG islands (CGI-CpGs) in the human [399] and the mouse genomes [401] respectively. From the cuRRBS output files we recovered the IDs of the sites that should be theoretically sequenced. Moreover, using the experimental RRBS data [399, 401], 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 [426].

# Chapter 5

# Final remarks

‘Caminante, son tus huellas
el camino, y nada más;
caminante, no hay camino:
se hace camino al andar.

Antonio Machado, 1912 [431]

The purpose of this thesis was to advance our understanding of the epigenetic ageing clock in humans. I now review the main conclusions from this work and propose future directions that could be of interest.

## 5.1 Statistical aspects

In Chapter 2, I have assessed different statistical methods that allowed me to **characterise the epigenetic landscape during human physiological ageing**. To date, DNA methylation data from blood, generated in the Illumina 450K methylation array platform, is the most abundant epigenetic data type available to study human ageing. I built a dataset of this data type for healthy individuals, pre-processed it and benchmarked different methods to correct for blood cell composition changes. I reproduce previous findings showing that a great proportion of the epigenome is affected by the ageing process (in my case around 30%, using a conservative threshold to correct for multiple testing). This highlights that the epigenetic ageing clock is a genome-wide phenomena that extends way beyond the cytosines included in most epigenetic clock models. Furthermore, the small effect sizes suggest that most age-related DNA methylation changes occur only in a small proportion of cells (DNA molecules) in the tissue (around 4% on average for the entire human lifespan).

Finally, I tested the behaviour of different epigenetic clocks (Horvath, Hannum, epiTOC) and developed a strategy to correct for potential batch effects in this context.

Current epigenetic clocks use a linear modelling framework. Nevertheless, many changes in methylation values during ageing are non-linear (for example during organismal growth). Horvath's clock corrects for this by transforming chronological age, but it would be interesting to try to model the changes of individual CpG sites before including them as part of the training. This could also help to identify modules of CpG sites that behave in the same way during ageing and allow deconvoluting the different processes that shape the epigenetic landscape and may be operative at different life stages. Additional **improvements in epigenetic clocks** will likely include integrating longitudinal information (which could help to identify different ageing trajectories) [432] and separating the contributions of mutations and epimutations to the methylation signal. Furthermore, it would be interesting to try to map the shapes of the DNA methylation changes to the changes in mortality rate at a human population level, therefore creating a link between molecular changes and epidemiological observations. This could be further validated in species with extremely different profiles of mortality rate (e.g. naked mole rat).

Current multi-tissue epigenetic clocks have been trained on all tissues available. Nevertheless, it is reasonable to assume that the strategies to maintain stable DNA methylation landscapes over time would significantly differ between highly proliferative tissues (such as blood) from those where cell division is a rare event (such as brain). Thus, building different epigenetic clocks for these two categories of tissues and analysing their genome-wide changes in DNA methylation over time could improve the accuracy of the models and provide **further insights into the role of cell division on the epigenetic ageing clock**.

There are methods that allow imputing DNA methylation patterns based on different genomic features for a 'static' epigenome, both at the bulk [358] and single-cell levels [415, 433]. Given that the regions that change their DNA methylation during ageing seem to share the genomic context, it would be interesting to design an **imputation algorithm for a 'dynamic' epigenome** (e.g. given an epigenome at time $t$, predict what that epigenome would look like at time $t + \Delta t$). This could also give us additional insights into how much the ageing-related changes are hard-coded in the genome and how much the environment and lifestyle contribute to modify it. Moreover, some of these predictions could be tested by introducing exogenous pieces of DNA in ageing mice.

Developmental disorders are useful biological systems to study the effects of altered functions in specific parts of the epigenetic machinery. As such, the analysis presented in

Chapter 3 could be further expanded into a statistical framework that allows quantifying how much certain epigenetic functions contribute to the methylation status of specific regions. In other words, the definition of **epimutational signatures** (e.g. epimutational signature 1 is the consequence of reduced H3K4 methylation in enhancers) that would allow to deconvolute the epigenetic processes behind a specific DNA methylation pattern (e.g. the one caused by ageing or smoking exposure).

## 5.2 Biological aspects

The goal of Chapter 3 was to study **how different parts of the epigenetic machinery affect the rate of the epigenetic ageing clock**, thus providing the first identified components of the hypothetical *epigenetic maintenance system* [209]. For that purpose, I studied the epigenetic age acceleration observed in patients with developmental disorders, many of which harbour mutations in proteins of the aforementioned epigenetic machinery.

This analysis revealed that **mutations in NSD1, an H3K36 methyltransferase, dramatically accelerate epigenetic ageing**. The effect sizes observed (on average > 7 years) are bigger than many of the conditions reported to accelerate the epigenetic ageing clock [223]. Importantly, the genomic context where these changes happen is partially shared with the ageing process. Regions marked by H3K27me3, deposited by Polycomb Repressing Complex 2 (PRC2), were highly enriched for these changes both in ageing and Sotos, consistent with previous reports. Interestingly, global DNA hypomethylation (a characteristic of Sotos patients) causes a redistribution of PRC2 and H3K27me3 from their normal targets (many of them developmental genes marked with bivalent chromatin) to other genomic regions, which leads to the aberrant expression of some of these genes [434]. Importantly, there is a mechanistic link between PRC2 recruitment and H3K36me3 via the Tudor domains of some polycomb-like proteins [345, 346]. As such, it would be expected that perturbations in the H3K36 methylation landscape would affect PRC2 activity. Furthermore, methylation of CpG sites in normally unmethylated CpG islands could also lead to a loss of PRC2 binding [346]. This could be happening in bivalent regions / DNA methylation valleys (DMVs) during ageing and affect the differentiation process of progenitor stem cells in adult tissues. Indeed, this seems to be the case for aged haematopoietic stem cells [182, 183], but whether this applies to other tissues still needs to be elucidated. Importantly, DNA methylation changes affecting progenitor stem cells could be propagated in the tissue, therefore contributing substantially to the signal captured by epigenetic clocks.

Hence, during ageing, there could be a **redistribution of PRC2 from bivalent regions / DMVs to other regions that have become hypomethylated, at the same time that *de novo* DNMT3A/B get relocated in the opposite direction** (as shown in Fig. 3.8), leading to a deregulation in the expression of developmental genes. This model expands and is overall compatible with the one proposed by Zheng, Widschwendter and Teschendorff to explain the increase in cancer risk with age [435]. While this could be induced by the rewiring of the H3K36 methylation landscape, direct evidence needs to be provided to ascertain that this is indeed the case during human physiological ageing. As such, it would be interesting to profile H3K36me3 during ageing in different tissues. Furthermore, differential expression of genes coding for the H3K36 methylation machinery (both methyltransferases and demethylases) during ageing would also be expected (e.g. by hypermethylating the promoter of NSD1, as observed in human neuroblastoma and glioma cells) [436]. Moreover, a study showing if cryptic transcription increases during human ageing (something that seems to happen in model organisms) could contribute to our understanding of the global functional consequences of these epigenetic changes. Finally, genes with lower levels of H3K36me3 should be more prone to cryptic transcription during ageing [336] and potentially display higher transcriptional heterogeneity between cells.

There is conflicting evidence on the literature on whether NSD1 can also catalyse the methylation of H4K20 *in vivo* [436, 437]. H4K20me1 is a histone mark highly enriched in telomeres [438] and depletion of H4K20 methylation leads to genomic instability [439]. This creates another interesting **link between telomere biology and the epigenetic ageing clock** (as discussed in Chapter 1, *TERT* genetic variants are associated with epigenetic age acceleration and its expression is required *in vitro* to ensure epigenetic ageing) [237]. It would be worth testing how the epigenetic ageing clock behaves in cancer-resistant mice that constitutively express *TERT* (which have an extended lifespan) [62].

Ageing-related DNA methylation changes generally **increase the informational entropy of the system** (i.e. the methylation values tend to 0.5, see section 3.5). It is tempting to speculate that, from a biological point of view, this can be interpreted as a dilution of the epigenetic marks that define stable cell types and transcriptional programs and an increase in cell-to-cell epigenetic heterogeneity. Some authors have suggested that epigenetic information is carried by a population of cells as a whole [295, 440]. Furthermore, even populations of a specific cell type (such as primed ESCs) show oscillations in the methylation values of specific regions, which seem to have a particularly high amplitude in enhancers [441] (one of the hotspots of hypomethylation changes during ageing). If a such a population were to be analysed with a bulk DNA methylation method, it would likely display a high

methylation entropy in enhancers. Furthermore, the fact that methylation entropy is higher in the sites of the Horvath clock could indicate that cytosines that display this type of metastable state make good predictors. Thus, it is possible that alterations in the DNA methylation oscillatory behaviour, caused by changes in the activities or the binding of DNMT3s and TETs (which could happen if the H3K36 methylation landscape is altered), are a feature of the epigenetic ageing clock.

Mechanistic advances will require **testing these ideas in the mouse**. First, it would be interesting to confirm whether the effects of heterozygous loss-of-function mutations in NSD1 are evolutionarily conserved, using the mouse multi-tissue epigenetic clock, and test if they affect the lifespan of these mice. Moreover, one of the remaining questions is whether the DNA methylation changes associated with the epigenetic ageing clock are functional at all. Epigenomic editing technologies [442] could help to answer this question. Additionally, testing how conserved these mechanisms are beyond mammals (e.g. in the African turquoise killifish) or whether they behave differently in species with remarkable longevity (such as the naked mole rat) would be of interest.

## 5.3 Technological aspects

In Chapter 4, we have created a computational method (cuRRBS) to **optimise the enrichment of specific sets of genomic sites through the combinatorial use of restriction enzymes**. This could be potentially applied to make future epigenetic clocks more cost-effective (especially if they are composed of several hundreds or thousands of sites). Furthermore, given how statistically degenerate epigenetic clocks are, new models could be trained taking into account the most cost-effective combinations of sites. Reductions in assay cost could lead to the wide adoption of DNA methylation-based biomarkers for high-throughput drug screening.

From a research point-of-view, it is fundamental that we **expand our analysis beyond the biased regions from the Illumina methylation array**. Therefore, whole genome bisulfite sequencing during ageing should become more common, allowing us to characterise the changes in the epigenetic landscape at higher resolution. This will likely become a reality thanks to the fast drop in sequencing costs and to the development of bisulfite-free methods that improve mapping rates [142].

Furthermore, it remains to be seen whether the DNA methylation changes observed during ageing occur in all cell types in the tissue or whether changes in the concentration

of specific cell types (e.g. progenitor stem cells) or clones are responsible for them. In this sense, **single-cell technologies** (specially those that profile transcriptome and epigenome simultaneously) and lineage tracing will become instrumental for future mechanistic advances on the epigenetic ageing clock [443].

# Appendix

# Supplementary figures

## S.1 Supplementary for chapter 2

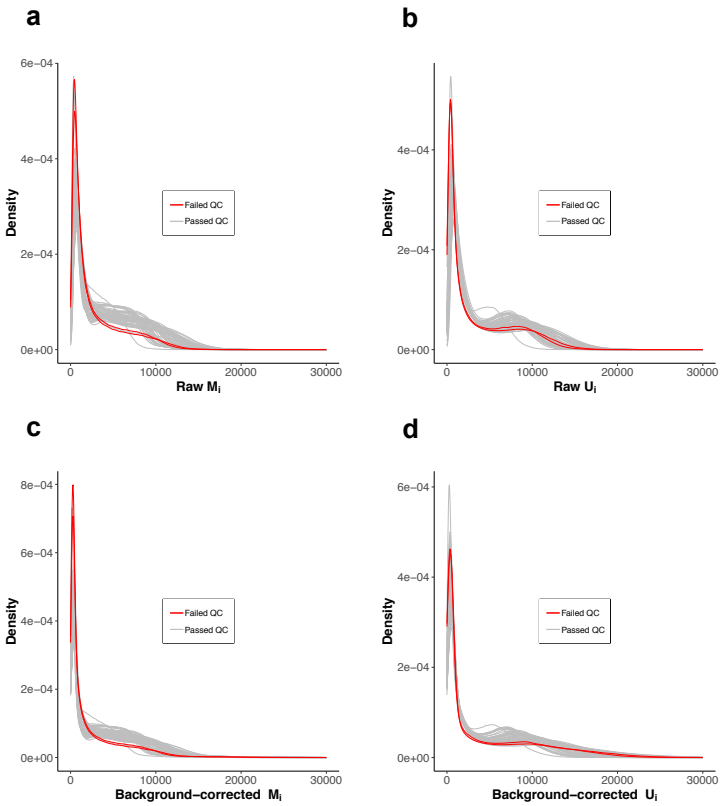



**Fig. S1.1** Effects of *noob* background correction on the array flurescence intensities. Distributions of the array fluorescence intensities for the **a.** methylated signals ($M_i$) before background correction; **b.** unmethylated signals ($U_i$) before background correction; **c.** methylated signals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

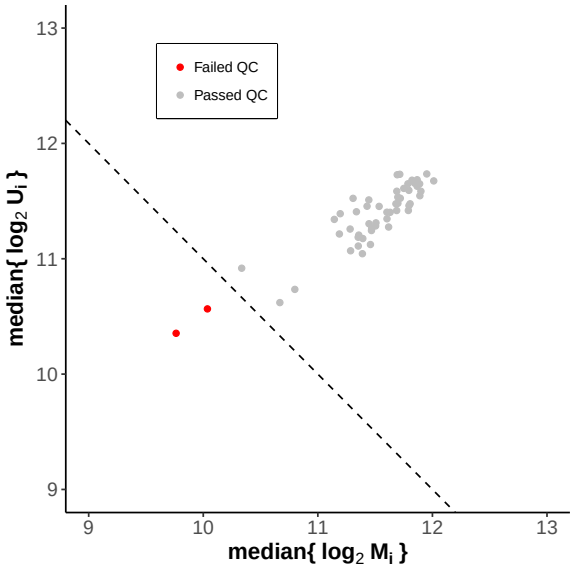



**Fig. S1.2** Quality control (QC) strategy to identify outlier samples, according to their global intensity values, in the GSE41273 batch. Those samples with low median intensity values (see criteria in section 2.1.2) were discarded from downstream analyses (2/53, in red). Each sample is represented 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.

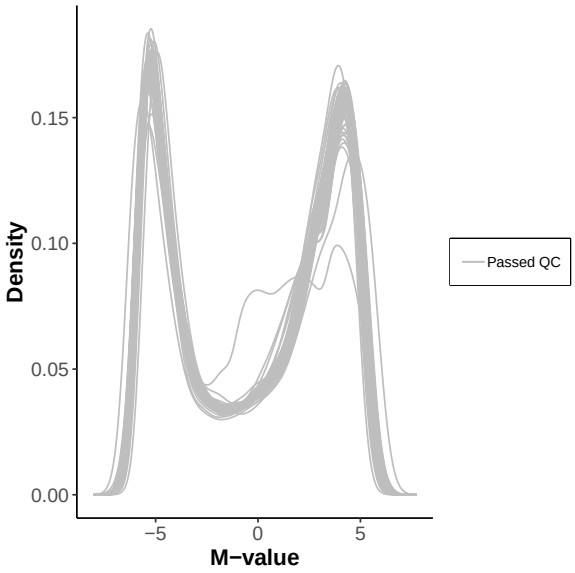



**Fig. S1.3** M-value distributions in the samples of the GSE41273 batch, after all the pre-processing steps have been carried out (background correction, quality control, probe filtering and BMIQ normalisation). M-values were calculated applying the logistic transformation to the $\beta$-values, as described in Du *et al.* [260]. 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*) [276, 280]. See section 2.1.3 in the main text for more details on what the different references refer to.

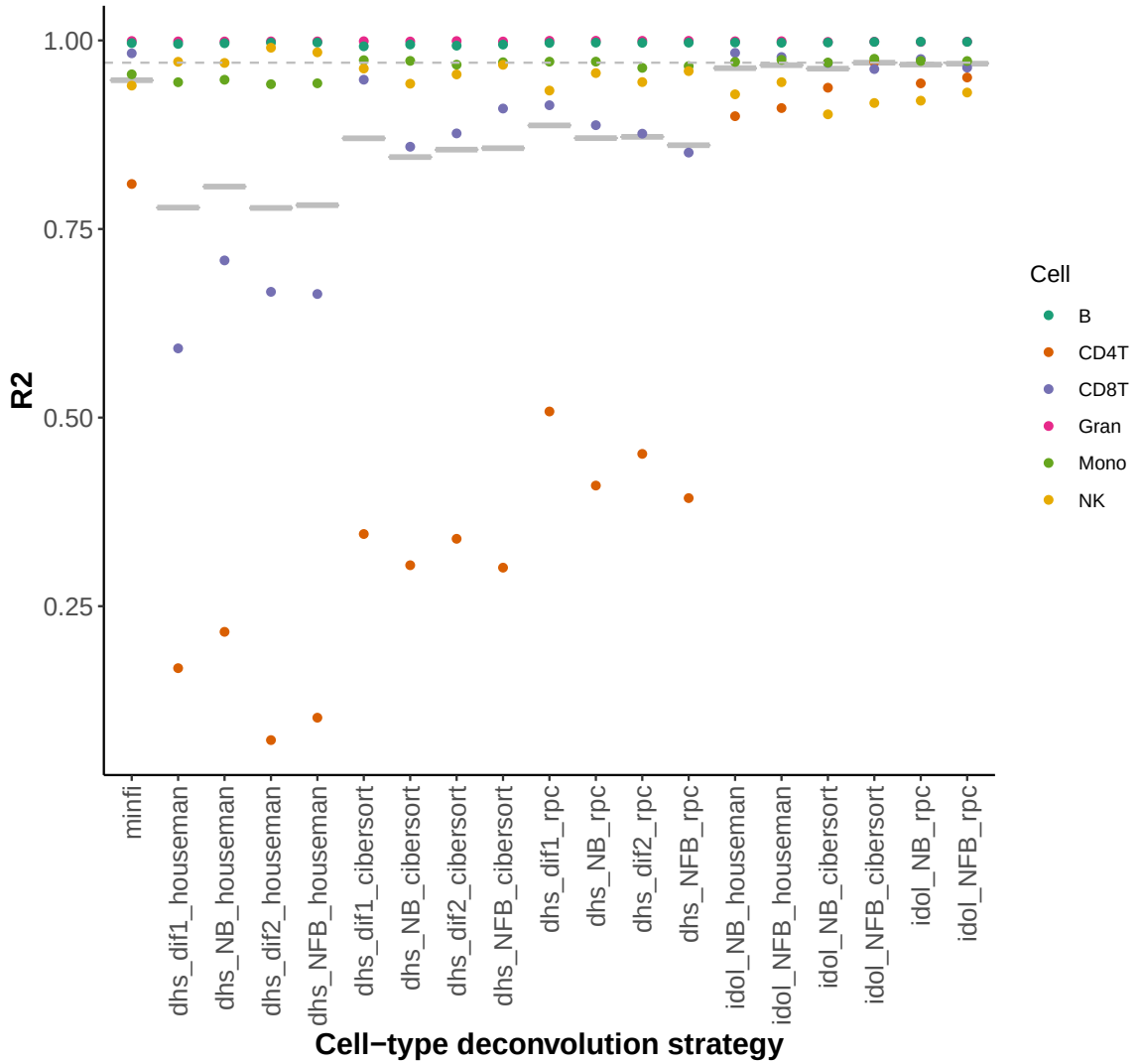



**Fig. S1.5** Benchmarking of the cell-type deconvolution strategies in blood. The x-axis shows the different strategies that were tested (for a detailed description see Fig. S1.4). The y-axis shows the results for the coefficient of determination ($R^2$) when comparing the predictions with the real proportions of cells in a gold-standard dataset (GSE77797) [279]. 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).

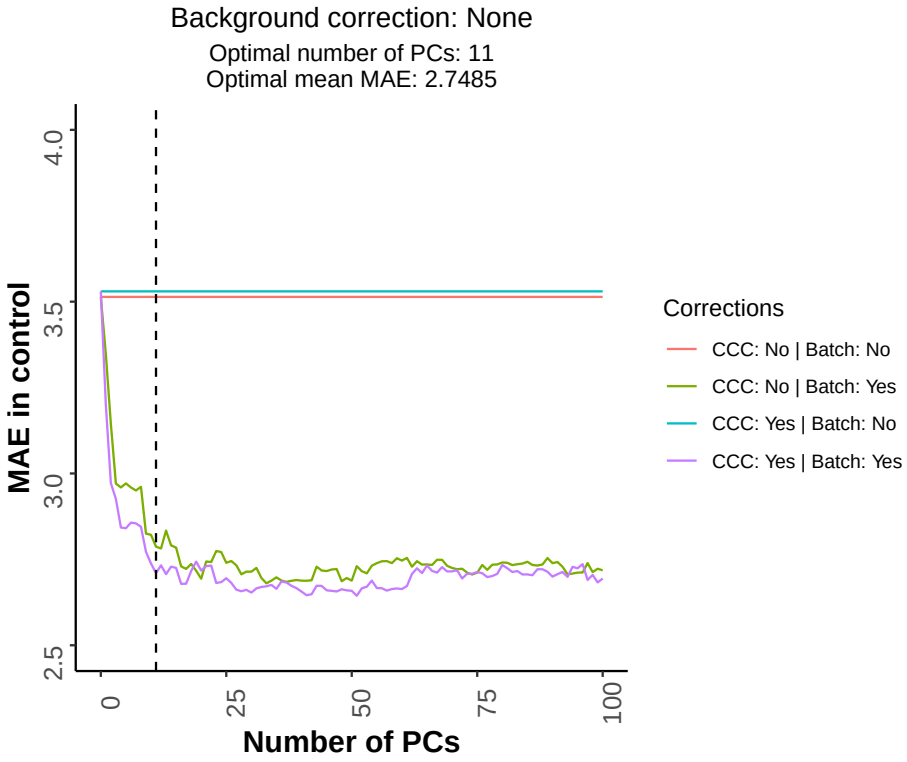



**Fig. S1.7** Plot showing how the median absolute error (MAE) of the prediction in the healthy individual samples, that should tend to zero, is reduced when the PCs capturing the technical variation are included as part of the modelling strategy (see equations 2.16 and 2.17). The dashed line represents 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 [209].

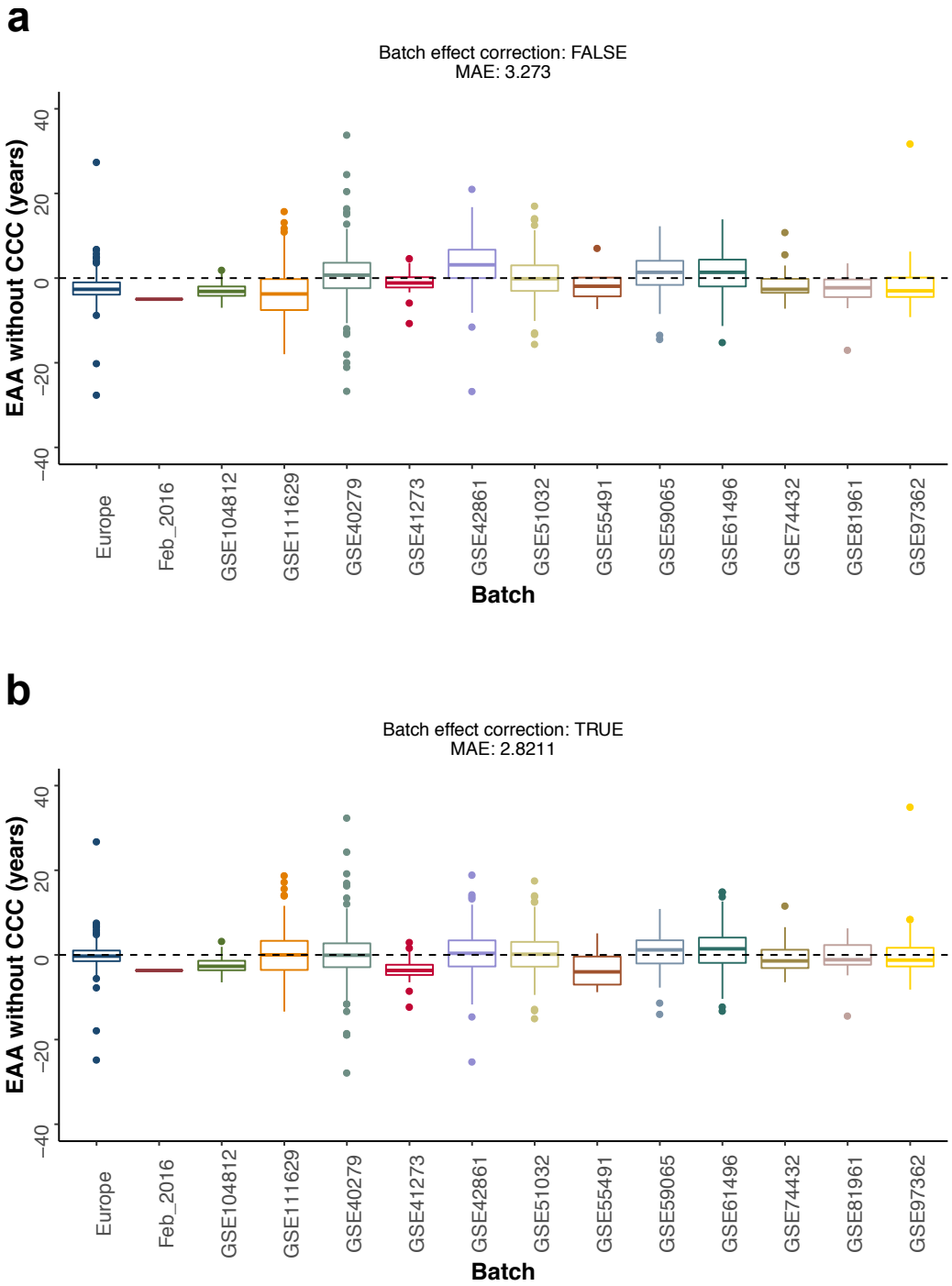



**Fig. S1.8** Correcting for batch effects in the context of the epigenetic clock. **a.** Distribution of the epigenetic age acceleration (EAA) for the different batches of healthy individual samples, using the control model without cell composition correction (CCC) and before applying batch effect correction. 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).

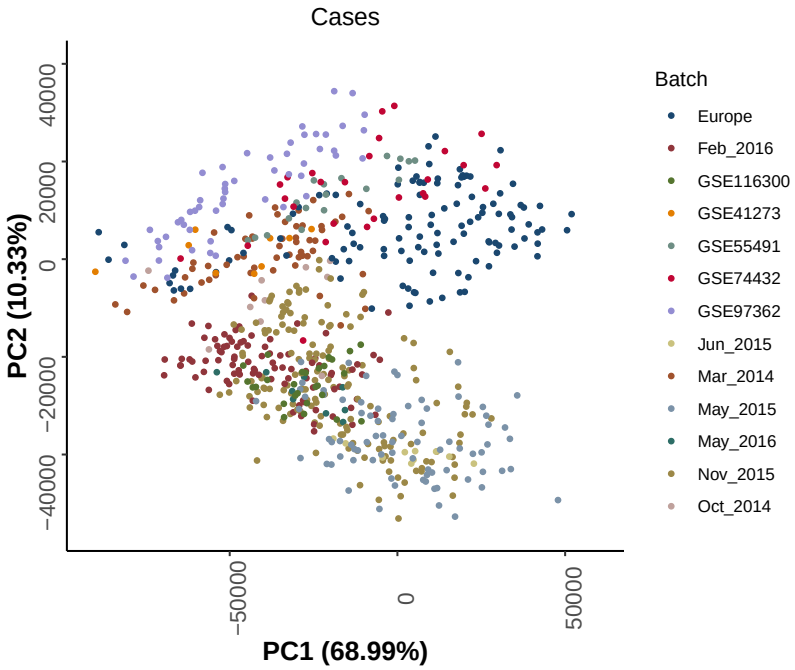



**Fig. S1.9** Scatterplot showing the values of the first two principal components (PCs) for the samples with developmental disorders (cases, see Chapter 3) after performing PCA on the control probes of the 450K arrays. Each point corresponds to a different sample and the colours represent the different 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 ($N = 666$).

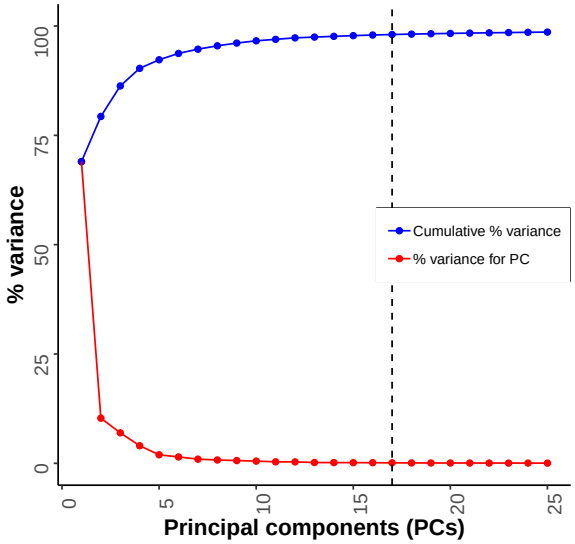



**Fig. S1.10** Plot showing the percentages of technical variance explained by the different PCs from the control probes. The dashed line represents the optimal number of PCs (17) that was finally used.

# S.2 Supplementary for chapter 3

| Batch name | $N_{♀}$ | $N_{♂}$ | N | Median age (years) | Other comments |
|---|---|---|---|---|---|
| Europe | 0 | 119 | 119 | 7.73 | |
| Feb_2016 | 20 | 20 | 40 | 6 | |
| GSE116300 | 4 | 5 | 9 | 3 | |
| GSE41273 | 0 | 9 | 9 | 7.75 | |
| GSE74432 | 11 | 16 | 27 | 10 | |
| GSE97362 | 4 | 9 | 13 | 15 | Samples from the 'validation cohort' were not included in the analysis, since they all seemed outliers on close examination |
| Jun_2015 | 1 | 1 | 2 | 3.5015 | |
| Mar_2014 | 11 | 6 | 17 | 8 | |
| May_2015 | 17 | 49 | 66 | 14 | |
| Nov_2015 | 35 | 30 | 65 | 6.7 | |
| **Total** | 103 | 264 | 367 | 8 | - |

**Table S2.1** Overview of the blood DNA methylation dataset from individuals with developmental disorders. The batches 'Europe', 'Feb_2016', 'Jun_2015', 'Mar_2014', 'May_2015' and 'Nov_2015' were generated in-house by our collaborators in Canada (see Chapter 3). The rest of the batches were downloaded from GEO [247]. $N_{♀}$: 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 and filtering (see section 3.2).

| Batch name | Developmental disorder | Gene | Mutation (DNA) | Mutation (protein) | Mutation effect | Pathogenic | Sex | Age (years) | DNAmAge |
|---|---|---|---|---|---|---|---|---|---|
| Europe | ASD | NA | NA | NA | NA | NA | Male | 23.25 | 29.94120469 |
| Europe | ASD | NA | NA | NA | NA | NA | Male | 25.75 | 23.66579727 |
| Europe | ASD | NA | NA | NA | NA | NA | Male | 23.75 | 22.89490773 |
| Europe | ASD | NA | NA | NA | NA | NA | Male | 26.58 | 31.33521081 |
| Europe | ASD | NA | NA | NA | NA | NA | Male | 11.83 | 13.55540994 |
| Europe | ASD | NA | NA | NA | NA | NA | Male | 12.33 | 12.62567804 |
| Europe | ASD | NA | NA | NA | NA | NA | Male | 11.67 | 11.91444556 |
| Europe | ASD | NA | NA | NA | NA | NA | Male | 12.67 | 15.1433583 |
| Europe | ASD | NA | NA | NA | NA | NA | Male | 15.92 | 20.69231419 |
| Europe | ASD | NA | NA | NA | NA | NA | Male | 16.92 | 18.37736076 |
| Europe | ASD | NA | NA | NA | NA | NA | Male | 15.92 | 14.74270021 |
| Europe | ASD | NA | NA | NA | NA | NA | Male | 19 | 28.69942806 |
| Europe | ASD | NA | NA | NA | NA | NA | Male | 16.75 | 20.84761017 |
| Europe | ASD | NA | NA | NA | NA | NA | Male | 20.16 | 17.69509361 |
| Europe | ASD | NA | NA | NA | NA | NA | Male | 12.92 | 18.28693655 |
| Europe | ASD | NA | NA | NA | NA | NA | Male | 13.25 | 12.24924728 |
| Europe | ASD | NA | NA | NA | NA | NA | Male | 13 | 15.27709141 |
| Europe | ASD | NA | NA | NA | NA | NA | Male | 13.25 | 15.93247357 |
| Europe | ASD | NA | NA | NA | NA | NA | Male | 13.16 | 17.97126245 |
| Europe | ASD | NA | NA | NA | NA | NA | Male | 13.67 | 18.5985271 |
| Europe | ASD | NA | NA | NA | NA | NA | Male | 7.67 | 9.834525429 |
| Europe | ASD | NA | NA | NA | NA | NA | Male | 7.92 | 8.819610809 |
| Europe | ASD | NA | NA | NA | NA | NA | Male | 7.73 | 10.53639331 |
| Europe | ASD | NA | NA | NA | NA | NA | Male | 8 | 8.782413174 |
| Europe | ASD | NA | NA | NA | NA | NA | Male | 7.83 | 8.331080792 |
| Europe | ASD | NA | NA | NA | NA | NA | Male | 8 | 8.412508081 |
| Europe | ASD | NA | NA | NA | NA | NA | Male | 10.83 | 12.94110542 |
| Europe | ASD | NA | NA | NA | NA | NA | Male | 11.5 | 16.52427744 |
| Europe | ASD | NA | NA | NA | NA | NA | Male | 10.83 | 9.546814402 |
| Europe | ASD | NA | NA | NA | NA | NA | Male | 11.5 | 10.75219435 |
| Europe | ASD | NA | NA | NA | NA | NA | Male | 10.83 | 11.7226536 |
| Europe | ASD | NA | NA | NA | NA | NA | Male | 6 | 8.750320884 |
| Europe | ASD | NA | NA | NA | NA | NA | Male | 5.75 | 8.069349936 |
| Europe | ASD | NA | NA | NA | NA | NA | Male | 6 | 8.205893972 |
| Europe | ASD | NA | NA | NA | NA | NA | Male | 5.83 | 8.765912407 |
| Europe | ASD | NA | NA | NA | NA | NA | Male | 6.33 | 6.903468104 |
| Europe | ASD | NA | NA | NA | NA | NA | Male | 5.25 | 5.648518225 |
| Europe | ASD | NA | NA | NA | NA | NA | Male | 5.67 | 5.896253109 |
| Europe | ASD | NA | NA | NA | NA | NA | Male | 5.42 | 6.160793858 |
| Europe | ASD | NA | NA | NA | NA | NA | Male | 5.75 | 8.719005258 |
| Europe | ASD | NA | NA | NA | NA | NA | Male | 5.42 | 6.49657694 |
| Europe | ASD | NA | NA | NA | NA | NA | Male | 3.92 | 4.884904225 |
| Europe | ASD | NA | NA | NA | NA | NA | Male | 4.08 | 4.766905985 |
| Europe | ASD | NA | NA | NA | NA | NA | Male | 4 | 5.462162993 |

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Europe | ASD | NA | NA | NA | NA | NA | Male | 4.08 | 4.557194499 |
| Europe | ASD | NA | NA | NA | NA | NA | Male | 4 | 4.383741212 |
| Europe | ASD | NA | NA | NA | NA | NA | Male | 4.25 | 5.321367013 |
| Europe | ASD | NA | NA | NA | NA | NA | Male | 3.25 | 2.797437125 |
| Europe | ASD | NA | NA | NA | NA | NA | Male | 3.42 | 3.906912403 |
| Europe | ASD | NA | NA | NA | NA | NA | Male | 3.33 | 4.703272329 |
| Europe | ASD | NA | NA | NA | NA | NA | Male | 3.5 | 3.223456196 |
| Europe | ASD | NA | NA | NA | NA | NA | Male | 3.42 | 4.024449964 |
| Europe | ASD | NA | NA | NA | NA | NA | Male | 3.58 | 4.662665584 |
| Europe | ASD | NA | NA | NA | NA | NA | Male | 5.16 | 7.931806871 |
| Europe | ASD | NA | NA | NA | NA | NA | Male | 5.16 | 6.144088681 |
| Europe | ASD | NA | NA | NA | NA | NA | Male | 5.16 | 5.423886319 |
| Europe | ASD | NA | NA | NA | NA | NA | Male | 5.25 | 6.873520458 |
| Europe | ASD | NA | NA | NA | NA | NA | Male | 5.16 | 6.828746343 |
| Europe | ASD | NA | NA | NA | NA | NA | Male | 5.25 | 6.287392617 |
| Europe | ASD | NA | NA | NA | NA | NA | Male | 6.5 | 7.549817595 |
| Europe | ASD | NA | NA | NA | NA | NA | Male | 6.83 | 5.310188113 |
| Europe | ASD | NA | NA | NA | NA | NA | Male | 6.67 | 8.807848811 |
| Europe | ASD | NA | NA | NA | NA | NA | Male | 7.16 | 7.314048584 |
| Europe | ASD | NA | NA | NA | NA | NA | Male | 6.83 | 7.143809294 |
| Europe | ASD | NA | NA | NA | NA | NA | Male | 7.25 | 4.888587648 |
| Europe | ASD | NA | NA | NA | NA | NA | Male | 10.08 | 11.01168613 |
| Europe | ASD | NA | NA | NA | NA | NA | Male | 10.08 | 9.091817984 |
| Europe | ASD | NA | NA | NA | NA | NA | Male | 10.08 | 12.00962928 |
| Europe | ASD | NA | NA | NA | NA | NA | Male | 10.5 | 11.89814401 |
| Europe | ASD | NA | NA | NA | NA | NA | Male | 10.08 | 10.85200361 |
| Europe | ASD | NA | NA | NA | NA | NA | Male | 10.58 | 15.97655481 |
| Europe | ASD | NA | NA | NA | NA | NA | Male | 14.67 | 19.40830372 |
| Europe | ASD | NA | NA | NA | NA | NA | Male | 15.25 | 17.28948864 |
| Europe | ASD | NA | NA | NA | NA | NA | Male | 14.83 | 18.99313794 |
| Europe | ASD | NA | NA | NA | NA | NA | Male | 15.25 | 17.40182035 |
| Europe | ASD | NA | NA | NA | NA | NA | Male | 15.08 | 20.74719227 |
| Europe | ASD | NA | NA | NA | NA | NA | Male | 15.83 | 17.66494621 |
| Europe | ASD | NA | NA | NA | NA | NA | Male | 1.83 | 2.332369997 |
| Europe | ASD | NA | NA | NA | NA | NA | Male | 2.33 | 2.079645877 |
| Europe | ASD | NA | NA | NA | NA | NA | Male | 2.08 | 3.093728905 |
| Europe | ASD | NA | NA | NA | NA | NA | Male | 2.5 | 3.327332717 |
| Europe | ASD | NA | NA | NA | NA | NA | Male | 2.08 | 3.081702301 |
| Europe | ASD | NA | NA | NA | NA | NA | Male | 2.5 | 3.640188937 |
| Europe | ASD | NA | NA | NA | NA | NA | Male | 27.67 | 5.315328746 |
| Europe | ASD | NA | NA | NA | NA | NA | Male | 32.92 | 35.79080593 |
| Europe | ASD | NA | NA | NA | NA | NA | Male | 31.83 | 35.12415194 |
| Europe | ASD | NA | NA | NA | NA | NA | Male | 35.16 | 34.8152863 |
| Europe | ASD | NA | NA | NA | NA | NA | Male | 32.33 | 33.47894995 |
| Europe | ASD | NA | NA | NA | NA | NA | Male | 11.58 | 14.81256772 |
| Europe | ASD | NA | NA | NA | NA | NA | Male | 4.5 | 3.982793413 |

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Europe | ASD | NA | NA | NA | NA | NA | Male | 4.75 | 6.632731853 |
| Europe | ASD | NA | NA | NA | NA | NA | Male | 4.5 | 5.453577973 |
| Europe | ASD | NA | NA | NA | NA | NA | Male | 5 | 6.0536493 |
| Europe | ASD | NA | NA | NA | NA | NA | Male | 4.67 | 4.665684936 |
| Europe | ASD | NA | NA | NA | NA | NA | Male | 5 | 5.538833496 |
| Europe | ASD | NA | NA | NA | NA | NA | Male | 4.33 | 6.826640979 |
| Europe | ASD | NA | NA | NA | NA | NA | Male | 4.42 | 5.074848057 |
| Europe | ASD | NA | NA | NA | NA | NA | Male | 4.33 | 4.069969605 |
| Europe | ASD | NA | NA | NA | NA | NA | Male | 4.5 | 2.914915908 |
| Europe | ASD | NA | NA | NA | NA | NA | Male | 4.33 | 4.177855824 |
| Europe | ASD | NA | NA | NA | NA | NA | Male | 4.5 | 5.359046992 |
| Europe | ASD | NA | NA | NA | NA | NA | Male | 7.33 | 4.981096393 |
| Europe | ASD | NA | NA | NA | NA | NA | Male | 7.5 | 7.521560211 |
| Europe | ASD | NA | NA | NA | NA | NA | Male | 7.33 | 5.632014057 |
| Europe | ASD | NA | NA | NA | NA | NA | Male | 7.58 | 5.381195679 |
| Europe | ASD | NA | NA | NA | NA | NA | Male | 7.42 | 7.07596058 |
| Europe | ASD | NA | NA | NA | NA | NA | Male | 7.58 | 6.118788705 |
| Europe | ASD | NA | NA | NA | NA | NA | Male | 8.83 | 8.225301829 |
| Europe | ASD | NA | NA | NA | NA | NA | Male | 9.08 | 9.139517533 |
| Europe | ASD | NA | NA | NA | NA | NA | Male | 8.83 | 7.154970232 |
| Europe | ASD | NA | NA | NA | NA | NA | Male | 9.67 | 9.966260719 |
| Europe | ASD | NA | NA | NA | NA | NA | Male | 8.92 | 8.69481855 |
| Europe | ASD | NA | NA | NA | NA | NA | Male | 9.67 | 12.84219838 |
| Europe | ASD | NA | NA | NA | NA | NA | Male | 8.08 | 10.35219735 |
| Europe | ASD | NA | NA | NA | NA | NA | Male | 8.25 | 8.849774575 |
| Europe | ASD | NA | NA | NA | NA | NA | Male | 8.16 | 9.464032218 |
| Europe | ASD | NA | NA | NA | NA | NA | Male | 8.33 | 10.51799454 |
| Europe | ASD | NA | NA | NA | NA | NA | Male | 8.16 | 9.41622481 |
| Europe | ASD | NA | NA | NA | NA | NA | Male | 8.75 | 13.39598874 |
| May_2015 | Angelman | *UBE3A* | NA | NA | NA | YES | Female | 7 | 5.473183736 |
| May_2015 | Angelman | *UBE3A* | NA | NA | NA | YES | Male | 13 | 15.48878288 |
| May_2015 | Angelman | *UBE3A* | NA | NA | NA | YES | Male | 55 | 59.49787491 |
| Nov_2015 | Angelman | *UBE3A* | NA | NA | NA | YES | Male | 1 | 2.790549766 |
| Nov_2015 | Angelman | *UBE3A* | NA | NA | NA | YES | Female | 4 | 3.956276247 |
| Nov_2015 | Angelman | *UBE3A* | NA | NA | NA | YES | Female | 15 | 17.87817565 |
| Nov_2015 | Angelman | *UBE3A* | NA | NA | NA | YES | Male | 1 | 2.320603044 |
| Nov_2015 | Angelman | *UBE3A* | NA | NA | NA | YES | Male | 4 | 4.348249902 |
| Nov_2015 | Angelman | *UBE3A* | NA | NA | NA | YES | Male | 1 | 0.959598999 |
| Nov_2015 | Angelman | *UBE3A* | NA | NA | NA | YES | Female | 1 | 1.994091886 |
| Nov_2015 | Angelman | *UBE3A* | NA | NA | NA | YES | Female | 10 | 8.697172131 |
| Nov_2015 | Angelman | *UBE3A* | NA | NA | NA | YES | Female | 14 | 15.7410421 |
| Nov_2015 | Angelman | *UBE3A* | NA | NA | NA | YES | Female | 6 | 5.13374965 |
| Nov_2015 | Angelman | *UBE3A* | NA | NA | NA | YES | Male | 25 | 32.45470863 |
| May_2015 | ATR-X | *ATRX* | c.6254G>A | p.Arg2085His | Missense | YES | Male | 6.3 | 6.19432086 |
| May_2015 | ATR-X | *ATRX* | c.736C>T | p.Arg246Cys | Missense | YES | Male | 18 | 13.11825849 |
| May_2015 | ATR-X | *ATRX* | c.6593A>G | p.His2198Arg | Missense | YES | Male | 1.4 | 2.604328944 |

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| May_2015 | ATR-X | *ATRX* | c.758T>C | p.Leu253Ser | Missense | YES | Male | 18.5 | 6.108170831 |
| May_2015 | ATR-X | *ATRX* | c.4817G>A | p.Ser1606Asn | Missense | YES | Male | 21 | 24.74309568 |
| May_2015 | ATR-X | *ATRX* | c.5786A>G | p.Lys1929Arg | Missense | YES | Male | 0.7 | -0.14552632 |
| May_2015 | ATR-X | *ATRX* | c.730A>C | p.Ile244Leu | Missense | YES | Male | 14 | 11.30064691 |
| May_2015 | ATR-X | *ATRX* | c.7156C>T | p.Arg2386* | Nonsense | YES | Male | 4.6 | 6.236506951 |
| May_2015 | ATR-X | *ATRX* | c.536A>G | p.Asn179Ser | Missense | YES | Male | 4.6 | 33.54375298 |
| May_2015 | ATR-X | *ATRX* | Exon 207 deletion | NA | Exonic deletion | YES | Male | 4.4 | 4.821921423 |
| May_2015 | ATR-X | *ATRX* | c.7366_7367insA | p.Met2456Asnfs*42 | Frameshift | YES | Male | 27 | 39.19917395 |
| May_2015 | ATR-X | *ATRX* | c.109C>T | p.Arg37* | Nonsense | YES | Male | 14.5 | 5.274937882 |
| May_2015 | ATR-X | *ATRX* | c.736C>T | p.Arg246Cys | Missense | YES | Male | 2.5 | 1.113449871 |
| May_2015 | ATR-X | *ATRX* | c.109C>T | p.Arg37* | Nonsense | YES | Male | 17.5 | 22.71435784 |
| May_2015 | ATR-X | *ATRX* | c.109C>T | p.Arg37* | Nonsense | YES | Male | 14 | 11.21597332 |
| Nov_2015 | Claes_Jensen | *KDM5C* | c.1510G>A | p.Val504Met | Missense | YES | Male | 30 | 42.69659356 |
| Nov_2015 | Claes_Jensen | *KDM5C* | c.1439C>T | p.Pro480Leu | Missense | YES_predicted | Male | 6 | 8.103173952 |
| Nov_2015 | Claes_Jensen | *KDM5C* | c.4439_4440delAG | p.Arg1481Glyfs* | Frameshift | YES | Male | 26 | 28.25654272 |
| Nov_2015 | Claes_Jensen | *KDM5C* | Intron 11:+5G>A | NA | Splice site mutation | YES | Male | 42 | 54.3236723 |
| Nov_2015 | Claes_Jensen | *KDM5C* | c.1510G>A | p.Val504Met | Missense | YES | Male | 8 | 10.07007313 |
| Nov_2015 | Claes_Jensen | *KDM5C* | c.1439C>T | p.Pro480Leu | Missense | YES | Male | 2 | 3.619189097 |
| Nov_2015 | Claes_Jensen | *KDM5C* | c.229G>A | p.Ala77Thr | Missense | YES | Male | 37 | 48.42002598 |
| Nov_2015 | Claes_Jensen | *KDM5C* | c.4439_4440delAG | p.Arg1481Glyfs* | Frameshift | YES | Male | 28 | 31.61445991 |
| Nov_2015 | Claes_Jensen | *KDM5C* | c.229G>A | p.Ala77Thr | Missense | YES | Male | 13 | 16.50827759 |
| Nov_2015 | Claes_Jensen | *KDM5C* | c.1510G>A | p.Val504Met | Missense | YES | Male | 26 | 38.69008936 |
| May_2015 | Coffin_Lowry | *RPS6KA3* | c.1520insA | p.Arg507fs | Frameshift | YES | Female | 6 | 4.093225848 |
| May_2015 | Coffin_Lowry | *RPS6KA3* | c.2065C>T | p.Gln689* | Nonsense | YES | Male | 11.5 | 10.63296406 |
| May_2015 | Coffin_Lowry | *RPS6KA3* | c.2186G>A | p.Arg729Gln | Missense | YES_predicted | Male | 4 | 4.62981308 |
| May_2015 | Coffin_Lowry | *RPS6KA3* | c.631_772del142 and c.774+5G>A | NA | Frameshift and intronic mutation | YES_predicted | Male | 7 | 5.068637974 |
| May_2015 | Coffin_Lowry | *RPS6KA3* | c.340C>T | p.Arg114Trp | Missense | YES_predicted | Male | 1.3 | 8.170755226 |
| May_2015 | Coffin_Lowry | *RPS6KA3* | c.727C>T | p.Arg243* | Nonsense | YES | Male | 13 | 14.17141748 |
| May_2015 | Coffin_Lowry | *RPS6KA3* | Intron 14:+1G>A | NA | Splice site mutation | YES | Male | 22.8 | 25.56720654 |
| May_2015 | Coffin_Lowry | *RPS6KA3* | NA | NA | Exonic and intronic deletion | YES | Male | 12 | 10.17620766 |
| May_2015 | Coffin_Lowry | *RPS6KA3* | c.386_387insCTTT | p.Phe130Phefs*141 | Frameshift | YES | Male | 2 | 1.808104516 |
| May_2015 | Coffin_Lowry | *RPS6KA3* | c.1155delT | p.Phe385fs*40 | Frameshift | YES | Male | 8 | 7.406603271 |
| Mar_2014 | Floating_Harbour | *SRCAP* | c.7303C>T | p.Arg2435* | Nonsense | YES | Female | 8 | 11.29885487 |
| Mar_2014 | Floating_Harbour | *SRCAP* | c.7330C>T | p.Arg2444* | Nonsense | YES | Female | 15 | 16.23135534 |
| Mar_2014 | Floating_Harbour | *SRCAP* | c.7282dupC | p.Arg2428Profs*15 | Frameshift | YES | Female | 6 | 5.620915174 |
| Mar_2014 | Floating_Harbour | *SRCAP* | c.7330C>T | p.Arg2444* | Nonsense | YES | Female | 10 | 42.55562244 |
| Mar_2014 | Floating_Harbour | *SRCAP* | c.8117C>G | p.Ser2706* | Nonsense | YES | Male | 4 | 2.815335426 |
| Mar_2014 | Floating_Harbour | *SRCAP* | c.7330C>T | p.Arg2444* | Nonsense | YES | Female | 5 | 4.112348915 |
| Mar_2014 | Floating_Harbour | *SRCAP* | c.7330C>T | p.Arg2444* | Nonsense | YES | Female | 42 | 43.43022309 |
| Mar_2014 | Floating_Harbour | *SRCAP* | c.7330C>T | p.Arg2444* | Nonsense | YES | Male | 12 | 12.37257473 |

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Mar_2014 | Floating_Harbour | *SRCAP* | c.7316dupC | p.Ala2440Serfs*3 | Frameshift | YES | Male | 10 | 4.424381743 |
| Mar_2014 | Floating_Harbour | *SRCAP* | c.7165G>T | p.Glu2389* | Nonsense | YES | Female | 8 | 1.524333568 |
| Mar_2014 | Floating_Harbour | *SRCAP* | c.7218_7219delTC | p.Gln2407Argfs*35 | Frameshift | YES | Male | 12 | 19.26251425 |
| Mar_2014 | Floating_Harbour | *SRCAP* | c.7330C>T | p.Arg2444* | Nonsense | YES | Male | 5 | 4.902256866 |
| Mar_2014 | Floating_Harbour | *SRCAP* | c.7330C>T | p.Arg2444* | Nonsense | YES | Female | 35 | 38.47378886 |
| Mar_2014 | Floating_Harbour | *SRCAP* | c.7330C>T | p.Arg2444* | Nonsense | YES | Female | 15 | 14.81418145 |
| Mar_2014 | Floating_Harbour | *SRCAP* | c.7549delC | p.Gln2517Lysfs*5 | Frameshift | YES | Male | 4 | 3.645524918 |
| Mar_2014 | Floating_Harbour | *SRCAP* | c.7330C>T | p.Arg2444* | Nonsense | YES | Female | 6 | 7.201471688 |
| Mar_2014 | Floating_Harbour | *SRCAP* | c.7219C>T | p.Gln2407* | Nonsense | YES | Female | 6 | 6.552720685 |
| GSE41273 | FXS | *FMR1* | NA | NA | CGG repeat expansion | YES | Male | 5 | -0.26537653 |
| GSE41273 | FXS | *FMR1* | NA | NA | CGG repeat expansion | YES | Male | 10.41667 | 4.620596743 |
| GSE41273 | FXS | *FMR1* | NA | NA | CGG repeat expansion | YES | Male | 7.75 | 9.380603836 |
| GSE41273 | FXS | *FMR1* | NA | NA | CGG repeat expansion | YES | Male | 4.333333 | 7.378290152 |
| GSE41273 | FXS | *FMR1* | NA | NA | CGG repeat expansion | YES | Male | 0.083333 | 7.256745087 |
| GSE41273 | FXS | *FMR1* | NA | NA | CGG repeat expansion | YES | Male | 4.166667 | 6.582911793 |
| GSE41273 | FXS | *FMR1* | NA | NA | CGG repeat expansion | YES | Male | 21 | 32.38418863 |
| GSE41273 | FXS | *FMR1* | NA | NA | CGG repeat expansion | YES | Male | 34.58333 | 46.41126929 |
| GSE41273 | FXS | *FMR1* | NA | NA | CGG repeat expansion | YES | Male | 48 | 58.89975733 |
| May_2015 | FXS | *FMR1* | NA | NA | CGG repeat expansion | YES | Male | 27 | 32.354974 |
| May_2015 | FXS | *FMR1* | NA | NA | CGG repeat expansion | YES | Male | 12 | 11.03917455 |
| May_2015 | FXS | *FMR1* | NA | NA | CGG repeat expansion | YES | Male | 42 | 40.85689027 |
| May_2015 | FXS | *FMR1* | NA | NA | CGG repeat expansion | YES | Male | 28 | 31.89965321 |
| May_2015 | FXS | *FMR1* | NA | NA | CGG repeat expansion | YES | Male | 15 | 15.3286979 |
| May_2015 | FXS | *FMR1* | NA | NA | CGG repeat expansion | YES | Male | 17 | 13.98190146 |
| May_2015 | FXS | *FMR1* | NA | NA | CGG repeat expansion | YES | Male | 21 | 21.42017869 |
| May_2015 | FXS | *FMR1* | NA | NA | CGG repeat expansion | YES | Male | 30 | 35.16564816 |
| May_2015 | FXS | *FMR1* | NA | NA | CGG repeat expansion | YES | Male | 28 | 27.14880628 |
| May_2015 | FXS | *FMR1* | NA | NA | CGG repeat expansion | YES | Male | 21 | 24.03936596 |
| May_2015 | FXS | *FMR1* | NA | NA | CGG repeat expansion | YES | Male | 33 | 37.84060062 |
| May_2015 | FXS | *FMR1* | NA | NA | CGG repeat expansion | YES | Male | 29 | 35.17133434 |
| May_2015 | FXS | *FMR1* | NA | NA | CGG repeat expansion | YES | Male | 25 | 25.67600147 |
| May_2015 | FXS | *FMR1* | NA | NA | CGG repeat expansion | YES | Male | 17 | 14.45573451 |
| May_2015 | FXS | *FMR1* | NA | NA | CGG repeat expansion | YES | Male | 33 | 36.37082822 |
| May_2015 | FXS | *FMR1* | NA | NA | CGG repeat expansion | YES | Male | 29 | 34.45261333 |
| May_2015 | FXS | *FMR1* | NA | NA | CGG repeat expansion | YES | Male | 20 | 24.86340454 |
| May_2015 | FXS | *FMR1* | NA | NA | CGG repeat expansion | YES | Male | 41 | 46.76222649 |

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| May_2015 | FXS | *FMR1* | NA | NA | CGG repeat expansion | YES | Male | 31 | 34.61968346 |
| May_2015 | FXS | *FMR1* | NA | NA | CGG repeat expansion | YES | Male | 27 | 29.78714348 |
| May_2015 | FXS | *FMR1* | NA | NA | CGG repeat expansion | YES | Male | 17 | 19.72629863 |
| May_2015 | FXS | *FMR1* | NA | NA | CGG repeat expansion | YES | Male | 15 | 11.78896917 |
| May_2015 | FXS | *FMR1* | NA | NA | CGG repeat expansion | YES | Male | 14 | 12.80759084 |
| GSE116300 | Kabuki | *KMT2D* | NA | p.Pro443fs | Frameshift | YES | Female | 1 | 0.790826048 |
| GSE116300 | Kabuki | *KMT2D* | NA | p.Tyr2199fs | Frameshift | YES | Female | 3 | 4.448848163 |
| GSE116300 | Kabuki | *KMT2D* | NA | p.Ser5307fs | Frameshift | YES | Male | 5 | 11.49079359 |
| GSE116300 | Kabuki | *KMT2D* | NA | p.Asn4403fs | Frameshift | YES | Male | 4.33 | 6.325934863 |
| GSE116300 | Kabuki | *KMT2D* | NA | p.Gln4102* | Nonsense | YES | Male | 2 | 5.566745677 |
| GSE116300 | Kabuki | *KMT2D* | NA | p.Gln3934* | Nonsense | YES | Male | 3.75 | 4.443224079 |
| GSE116300 | Kabuki | *KMT2D* | c.14515+1G>T | NA | Splice site mutation | YES | Male | 2.5 | 16.55101592 |
| GSE116300 | Kabuki | *KMT2D* | NA | p.Gln4090* | Nonsense | YES | Female | 1.42 | 3.379081974 |
| GSE116300 | Kabuki | *KMT2D* | NA | p.Thr1708fs | Frameshift | YES | Female | 11.5 | 10.71344707 |
| GSE97362 | Kabuki | *KMT2D* | c.15061C>T | p.Arg5021* | Nonsense | YES | Female | 14 | 8.946680052 |
| GSE97362 | Kabuki | *KMT2D* | c.16318delG | p.Glu5440Argfs*16 | Frameshift | YES | Male | 1 | 0.664960442 |
| GSE97362 | Kabuki | *KMT2D* | c.15030dup | p.Glu5011Argfs*13 | Frameshift | YES | Male | 18 | 24.00757516 |
| GSE97362 | Kabuki | *KMT2D* | c.8172_8173del | p.Pro2724Glnfs*5 | Frameshift | YES | Female | 16 | 4.540501556 |
| GSE97362 | Kabuki | *KMT2D* | c.6595delT | p.Tyr2199Ilefs*65 | Frameshift | YES | Male | 15 | 6.279894046 |
| GSE97362 | Kabuki | *KMT2D* | c.14055_14056delCA | p.His4685Glnfs*4 | Frameshift | YES | Male | 11 | 9.2260079 |
| GSE97362 | Kabuki | *KMT2D* | c.6295C>T | p.Arg2099* | Nonsense | YES | Male | 14 | 6.594838599 |
| GSE97362 | Kabuki | *KMT2D* | c.4135delA | p.Met1379Valfs*52 | Frameshift | YES | Male | 20 | 10.04269734 |
| GSE97362 | Kabuki | *KMT2D* | c.12592C>T | p.Arg4198* | Nonsense | YES | Male | 18 | 9.095825776 |
| GSE97362 | Kabuki | *KMT2D* | c.4135delA | p.Met1379Valfs*52 | Frameshift | YES | Male | 6 | 8.462691919 |
| GSE97362 | Kabuki | *KMT2D* | c.11710C>T | p.Gln3904* | Nonsense | YES | Male | 16 | 12.68670209 |
| GSE97362 | Kabuki | *KMT2D* | c.15143G>A | p.Arg5048His | Missense | YES_predicted | Female | 7 | 0.627461504 |
| GSE97362 | Kabuki | *KMT2D* | c.16522-5_16522-4delTT | NA | Splice site mutation | YES_predicted | Female | 15 | 12.75508563 |
| Jun_2015 | Kabuki | *KMT2D* | c.1801_1822dup22 | NA | Frameshift | YES | Male | 7 | 6.044371299 |
| Nov_2015 | Kabuki | *KMT2D* | c.13059delG | p.Pro4353fs | Frameshift | YES | Female | 6.7 | 5.526369466 |
| Nov_2015 | Kabuki | *KMT2D* | c.839+1delG | NA | Splice site mutation | YES | Male | 1.9 | 2.51325414 |
| Nov_2015 | Kabuki | *KMT2D* | c.15844C>T | p.Arg5282* | Nonsense | YES | Female | 3.9 | 3.752004426 |
| Nov_2015 | Kabuki | *KMT2D* | c.16294C>T | p.Arg5432Trp | Missense | YES_predicted | Male | 21.6 | 30.3375233 |
| Nov_2015 | Kabuki | *KMT2D* | c.8488C>T | p.Arg2830* | Nonsense | YES | Female | 0 | -0.1055224 |
| Nov_2015 | Kabuki | *KMT2D* | c.4168dupG | p.Ala1390fs | Frameshift | YES | Female | 3.8 | 4.177253095 |
| Nov_2015 | Kabuki | *KMT2D* | c.15289C>T | p.Arg5097* | Nonsense | YES | Male | 4.3 | 6.455955113 |
| Nov_2015 | Kabuki | *KMT2D* | c.4419-2A>G | NA | Splice site mutation | YES | Male | 2.6 | 3.387623395 |
| Nov_2015 | Kabuki | *KMT2D* | c.16048A>T | p.Lys5350* | Nonsense | YES | Female | 19.1 | 19.2926115 |
| Nov_2015 | Kabuki | *KMT2D* | c.10201C>T | p.Gln3401* | Nonsense | YES | Male | 7.1 | 8.838432826 |
| Nov_2015 | Kabuki | *KMT2D* | c.16360C>T | p.Arg5454* | Nonsense | YES | Male | 3.4 | 5.199197126 |
| Nov_2015 | Kabuki | *KMT2D* | c.8692C>T | p.Gln2898* | Nonsense | YES | Male | 3.1 | 3.423420462 |

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Nov_2015 | Kabuki | *KMT2D* | c.14878C>T | p.Arg4960* | Nonsense | YES | Female | 4.1 | 4.752807097 |
| Nov_2015 | Kabuki | *KMT2D* | c.6265A>T | p.Lys2089* | Nonsense | YES | Female | 23.1 | 25.95907184 |
| Nov_2015 | Kabuki | *KMT2D* | c.10740+1G>A | NA | Splice site mutation | YES | Female | 6.9 | 6.253113479 |
| Nov_2015 | Kabuki | *KMT2D* | c.13652T>A | p.Leu4551* | Nonsense | YES | Male | 2.2 | 3.757460909 |
| Nov_2015 | Kabuki | *KMT2D* | c.11596C>T | p.Gln3866* | Nonsense | YES | Female | 1 | 1.193509229 |
| Nov_2015 | Kabuki | *KMT2D* | c.548delC | p.Pro183fs | Frameshift | YES | Female | 16.6 | 8.413539447 |
| Nov_2015 | Kabuki | *KMT2D* | c.7411C>T | p.Arg2471* | Nonsense | YES | Female | 3.3 | 3.541604601 |
| Nov_2015 | Kabuki | *KMT2D* | c.1966dupC | pLeu656fs | Frameshift | YES | Female | 24.1 | 28.78927404 |
| Nov_2015 | Kabuki | *KMT2D* | c.6200delA | p.Asn2067fs | Frameshift | YES | Female | 9.5 | 6.485224166 |
| Nov_2015 | Kabuki | *KMT2D* | c.7933C>T | p.Arg2645* | Nonsense | YES | Female | 9.3 | 8.701999271 |
| Nov_2015 | Kabuki | *KMT2D* | c.13450C>T | p.Arg4484* | Nonsense | YES | Female | 5.8 | 5.430619578 |
| Feb_2016 | Noonan | *PTPN11* | c.1403C>T | p.Thr468Met | Missense | YES | Male | 9 | 10.53231848 |
| Feb_2016 | Noonan | *PTPN11* | c.1391G>C | p.Gly464Ala | Missense | YES | Female | 28 | 25.06455423 |
| Feb_2016 | Noonan | *PTPN11* | c.1493G>T | p.Arg498Leu | Missense | YES | Male | 0.4 | 1.069462128 |
| Feb_2016 | Noonan | *PTPN11* | c.836A>G | p.Tyr279Cys | Missense | YES | Male | 0.2 | 0.145725107 |
| Feb_2016 | Noonan | *PTPN11* | c.1493G>T | p.Arg498Leu | Missense | YES | Male | 7 | 7.125930003 |
| Feb_2016 | Noonan | *PTPN11* | c.1528C>G | p.Gln510Glu | Missense | YES | Female | 2 | 4.906928458 |
| Feb_2016 | Noonan | *PTPN11* | c.228G>C | p.Glu76Asp | Missense | YES | Male | 17 | 17.52765019 |
| Feb_2016 | Noonan | *PTPN11* | c.215C>G | p.Ala72Gly | Missense | YES | Female | 13 | 9.011977393 |
| Feb_2016 | Noonan | *PTPN11* | c.1391G>C | p.Gly464Ala | Missense | YES | Female | 0.7 | 1.172244358 |
| Feb_2016 | Noonan | *PTPN11* | c.922A>G | p.Asn308Asp | Missense | YES | Male | 15 | 14.68576639 |
| Feb_2016 | Noonan | *PTPN11* | c.836A>G | p.Tyr279Cys | Missense | YES | Male | 0.3 | 0.576697185 |
| Feb_2016 | Noonan | *PTPN11* | c.214G>T | p.Ala72Ser | Missense | YES | Male | 0.9 | 1.080594238 |
| Feb_2016 | Noonan | *PTPN11* | c.178G>A | p.Gly60Ser | Missense | YES | Male | 2 | 3.079510066 |
| Feb_2016 | Noonan | *PTPN11* | c.172A>G | p.Asn58Asp | Missense | YES | Male | 37 | 42.63784241 |
| Feb_2016 | Noonan | *PTPN11* | c.174C>A | p.Asn58Lys | Missense | YES | Female | 27 | 32.19911243 |
| Feb_2016 | Noonan | *RAF1* | c.781C>T | p.Pro261Ser | Missense | YES | Male | 9 | 11.76954478 |
| Feb_2016 | Noonan | *RAF1* | c.770C>T | p.Ser257Leu | Missense | YES | Female | 4 | 6.836828788 |
| Feb_2016 | Noonan | *RAF1* | c.788T>G | p.Val263Gly | Missense | YES | Male | 8 | 10.54386119 |
| Feb_2016 | Noonan | *RAF1* | c.782C>T | p.Pro261Leu | Missense | YES | Male | 3 | 5.956377653 |
| Feb_2016 | Noonan | *RAF1* | c.786T>A | p.Asn262Lys | Missense | YES | Female | 3 | 3.603073783 |
| Feb_2016 | Noonan | *RAF1* | c.768G>T | p.Arg256Ser | Missense | YES | Male | 20 | 21.09275241 |
| Feb_2016 | Noonan | *RAF1* | c.524A>G | p.His175Arg | Missense | YES | Female | 0.7 | 0.815080545 |
| Feb_2016 | Noonan | *RAF1* | c.1837C>G | p.Leu613Val | Missense | YES | Female | 10 | 7.425274033 |
| Feb_2016 | Noonan | *RAF1* | c.775T>A | p.Ser259Thr | Missense | YES | Female | 8 | 8.883918263 |
| Feb_2016 | Noonan | *RAF1* | c.1472C>T | p.Thr491Ile | Missense | YES | Female | 26 | 29.82312626 |
| Feb_2016 | Noonan | *RAF1* | c.781C>A | p.Pro261Thr | Missense | YES | Female | 11 | 12.25565712 |
| Feb_2016 | Noonan | *SOS1* | c.2536G>A | p.Glu846Lys | Missense | YES | Female | 3 | 2.62618922 |
| Feb_2016 | Noonan | *SOS1* | c.1654A>G | p.Arg552Gly | Missense | YES | Male | 16 | 12.47288243 |
| Feb_2016 | Noonan | *SOS1* | c.1310T>C | p.Ile437Thr | Missense | YES | Female | 7 | 7.309199493 |
| Feb_2016 | Noonan | *SOS1* | c.806T>C | p.Met269Thr | Missense | YES | Female | 35 | 25.04627009 |
| Feb_2016 | Noonan | *SOS1* | c.1642A>C | p.Ser548Arg | Missense | YES | Female | 3 | 4.372134286 |
| Feb_2016 | Noonan | *SOS1* | c.925G>T | p.Asp309Tyr | Missense | YES | Female | 49 | 45.20434465 |
| Feb_2016 | Noonan | *SOS1* | c.1655G>C | p.Arg552Thr | Missense | YES | Male | 1 | 2.41372048 |
| Feb_2016 | Noonan | *SOS1* | c.508A>G | p.Lys170Glu | Missense | YES | Male | 0.3 | 0.944100935 |
| Feb_2016 | Noonan | *SOS1* | c.1294T>C | p.Trp432Arg | Missense | YES | Female | 14 | 17.03491762 |

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Feb_2016 | Noonan | *SOS1* | c.1322G>A | p.Cys441Tyr | Missense | YES | Female | 0.6 | 0.555111083 |
| Feb_2016 | Noonan | *SOS1* | c.806T>G | p.Met269Arg | Missense | YES | Female | 0.4 | 0.844087032 |
| Feb_2016 | Noonan | *SOS1* | c.797C>A | p.Thr266Lys | Missense | YES | Male | 1 | 2.133506512 |
| Feb_2016 | Noonan | *SOS1* | c.1297G>A | p.Glu433Lys | Missense | YES | Male | 1 | 1.481217449 |
| Feb_2016 | Noonan | *SOS1* | c.1300G>A | p.Gly434Arg | Missense | YES | Male | 5 | 8.558246566 |
| May_2015 | Rett | *MECP2* | NA | p.Arg106Trp | Missense | YES | Female | 1 | 1.835127123 |
| May_2015 | Rett | *MECP2* | NA | p.Arg168* | Nonsense | YES | Female | 25 | 29.34649481 |
| May_2015 | Rett | *MECP2* | NA | p.Pro302Arg | Missense | YES | Female | 34 | 35.17904908 |
| May_2015 | Rett | *MECP2* | NA | NA | Exonic deletion | YES | Female | 2 | 2.581071992 |
| May_2015 | Rett | *MECP2* | NA | p.Thr158Met | Missense | YES | Female | 1 | 2.210005617 |
| May_2015 | Rett | *MECP2* | Deletion in exon 4 | NA | Exonic deletion | YES | Female | 3 | 5.225511336 |
| May_2015 | Rett | *MECP2* | NA | p.Thr158Met | Missense | YES | Female | 1 | 2.510753024 |
| May_2015 | Rett | *MECP2* | NA | p.Pro225Arg | Missense | YES | Female | 4 | 6.160921221 |
| May_2015 | Rett | *MECP2* | c.1157_1197del41 | p.Glu374fs | Frameshift | YES | Female | 6 | 6.2636907 |
| May_2015 | Rett | *MECP2* | NA | p.Arg255* | Nonsense | YES | Female | 1.5 | 1.084382282 |
| May_2015 | Rett | *MECP2* | Deletion in exons 3 and 4 | NA | Exonic deletion | YES | Female | 6 | 6.883663479 |
| May_2015 | Rett | *MECP2* | NA | p.Arg106Trp | Missense | YES | Female | 29 | 38.83647398 |
| May_2015 | Rett | *MECP2* | NA | p.Thr158Met | Missense | YES | Female | 3 | 4.77442952 |
| May_2015 | Rett | *MECP2* | NA | p.Arg255* | Nonsense | YES | Female | 11 | 11.74653291 |
| May_2015 | Rett | *MECP2* | Partial deletion of exon 4 | NA | Exonic deletion | YES | Female | 4 | 3.072948979 |
| Jun_2015 | Saethre_Chotzen | *TWIST1* | c.385_405dup21 | NA | In-frame insertion | YES | Female | 0.003 | -0.35722332 |
| Nov_2015 | Saethre_Chotzen | *TWIST1* | c.149delC | p.Ala50fs | Frameshift | YES | Male | 0.02 | 0.16785508 |
| Nov_2015 | Saethre_Chotzen | *TWIST1* | c.149delC | p.Ala50fs | Frameshift | YES | Female | 0.1 | 13.96937513 |
| Nov_2015 | Saethre_Chotzen | *TWIST1* | c.376G>T | p.Glu126* | Nonsense | YES | Male | 38 | 41.56611411 |
| Nov_2015 | Saethre_Chotzen | *TWIST1* | c.406_407ins21 | NA | In-frame insertion | YES | Male | 30 | 29.61790422 |
| Nov_2015 | Saethre_Chotzen | *TWIST1* | c.156delC | p.Pro52fs | Frameshift | YES | Female | 33.5 | 27.76671901 |
| Nov_2015 | Saethre_Chotzen | *TWIST1* | c.418_419ins21 | NA | In-frame insertion | YES | Male | 17.7 | 15.97052177 |
| Nov_2015 | Saethre_Chotzen | *TWIST1* | c.211C>T | p.Gln71* | Nonsense | YES | Female | 20.7 | 18.347741 |
| Nov_2015 | Saethre_Chotzen | *TWIST1* | c.325C>T | p.Gln109* | Nonsense | YES_predicted | Male | 0.7 | 0.45749609 |
| Nov_2015 | Saethre_Chotzen | *TWIST1* | c.396_416dup21 | NA | In-frame insertion | YES | Male | 0.1 | 0.386967314 |
| Nov_2015 | Saethre_Chotzen | *TWIST1* | c.193G>T | p.Glu65* | Nonsense | YES | Female | 0.01 | 0.049927484 |
| Nov_2015 | Saethre_Chotzen | *TWIST1* | c.472T>C | p.Phe158Leu | Missense | YES | Female | 23.3 | 0.174364646 |
| Nov_2015 | Saethre_Chotzen | *TWIST1* | NA | NA | Full gene deletion | YES | Female | 0.35 | 0.404844597 |
| Nov_2015 | Saethre_Chotzen | *TWIST1* | NA | NA | Full gene deletion | YES | Female | 0.003 | 7.069271322 |
| Nov_2015 | Saethre_Chotzen | *TWIST1* | c.160G>T | p.Gly54* | Nonsense | YES | Female | 0.7 | 0.830512167 |
| Nov_2015 | Saethre_Chotzen | *TWIST1* | c.397_417dup21 | NA | In-frame insertion | YES_predicted | Female | 20.5 | 25.83177177 |
| Nov_2015 | Saethre_Chotzen | *TWIST1* | c.120_145del26 | NA | Frameshift | YES | Male | 0.6 | 0.491449014 |
| Nov_2015 | Saethre_Chotzen | *TWIST1* | c.149delC | p.Ala50fs | Frameshift | YES | Female | 23.5 | 18.94806941 |
| Nov_2015 | Saethre_Chotzen | *TWIST1* | c.394_414del21 | NA | In-frame deletion | YES | Female | 12.3 | 10.10722932 |
| Nov_2015 | Saethre_Chotzen | *TWIST1* | c.352C>G | p.Arg118Gly | Missense | YES_predicted | Female | 21.5 | 23.41800184 |

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Nov_2015 | Saethre_Chotzen | *TWIST1* | c.376G>T | p.Glu126* | Nonsense | YES | Female | 0.8 | 0.92117994 |
| Nov_2015 | Saethre_Chotzen | *TWIST1* | c.490C>T | p.Gln164* | Nonsense | YES | Female | 28.7 | 28.56296158 |
| GSE74432 | Sotos | *NSD1* | chr5:175,366,008-177,470,488 | NA | Long deletion | YES | Female | 9 | 8.442111023 |
| GSE74432 | Sotos | *NSD1* | chr5:175,764,262-177,059,256 | NA | Long deletion | YES | Female | 7 | 16.4840396 |
| GSE74432 | Sotos | *NSD1* | Exons 15-19 deletion | NA | Exonic deletion | YES | Male | 10 | 26.70242296 |
| GSE74432 | Sotos | *NSD1* | c.1716delC | p.Cys573Valfs*26 | Frameshift | YES | Female | 10 | 14.59121875 |
| GSE74432 | Sotos | *NSD1* | c.6454C>T | p.Arg2152* | Nonsense | YES | Female | 3.5 | 9.371834336 |
| GSE74432 | Sotos | *NSD1* | c.5445C>G | p.Tyr1815* | Nonsense | YES | Female | 13.2 | 22.67264348 |
| GSE74432 | Sotos | *NSD1* | c.4843delT | p.Tyr1615Thrfs*27 | Frameshift | YES | Male | 3 | 7.039068162 |
| GSE74432 | Sotos | *NSD1* | NA | NA | Microdeletion | YES | Male | 2.2 | 15.1797238 |
| GSE74432 | Sotos | *NSD1* | c.6349C>T | p.Arg2117* | Nonsense | YES | Female | 12 | 26.9093016 |
| GSE74432 | Sotos | *NSD1* | c.1492C>T | p.Arg498* | Nonsense | YES | Male | 2.2 | 8.399587071 |
| GSE74432 | Sotos | *NSD1* | c.6454C>T | p.Arg2152* | Nonsense | YES | Male | 18 | 32.23853498 |
| GSE74432 | Sotos | *NSD1* | c.1583delA | p.Lys528Argfs*8 | Frameshift | YES | Male | 19.7 | 27.25531484 |
| GSE74432 | Sotos | *NSD1* | c.2014_2018delACAGA | p.Thr672Glufs*9 | Frameshift | YES | Male | 8 | 26.46585423 |
| GSE74432 | Sotos | *NSD1* | c.2014_2018delACAGA | p.Thr672Glufs*9 | Frameshift | YES | Male | 41 | 67.36442178 |
| GSE74432 | Sotos | *NSD1* | c.2014_2018delACAGA | p.Thr672Glufs*9 | Frameshift | YES | Female | 2 | 11.34495985 |
| GSE74432 | Sotos | *NSD1* | c.1810C>T | p.Arg604* | Nonsense | YES | Female | 1.6 | 6.2471485 |
| GSE74432 | Sotos | *NSD1* | c.1801A>T | p.Lys601* | Nonsense | YES | Male | 10.6 | 30.82670587 |
| GSE74432 | Sotos | *NSD1* | c.4977_4978insG | p.Arg1660Alafs*13 | Frameshift | YES | Male | 20 | 41.38296452 |
| GSE74432 | Sotos | *NSD1* | c.6437G>C | p.Cys2146Ser | Missense | YES_predicted | Male | 2 | 9.83036953 |
| GSE74432 | Sotos | *NSD1* | c.6412T>C | p.Cys2138Arg | Missense | YES_predicted | Male | 7 | 29.0788673 |
| GSE74432 | Weaver | *EZH2* | c.457_459delTAT | p.Tyr153del | In-frame deletion | YES | Male | 30 | 40.6786865 |
| GSE74432 | Weaver | *EZH2* | c.2080C>T | p.His694Tyr | Missense | YES | Female | 10.9167 | 17.28626931 |
| GSE74432 | Weaver | *EZH2* | c.2050C>T | p.Arg684Cys | Missense | YES | Male | 2.5833 | 2.611103643 |
| GSE74432 | Weaver | *EZH2* | c.398A>G | p.Tyr133Cys | Missense | YES | Female | 17 | 7.870608634 |
| GSE74432 | Weaver | *EZH2* | c.553G>C | p.Asp185His | Missense | YES | Male | 15.4167 | 18.04003584 |
| GSE74432 | Weaver | *EZH2* | c.394C>T | p.Pro132Ser | Missense | YES | Female | 19.75 | 21.09459251 |
| GSE74432 | Weaver | *EZH2* | c.1876G>A | p.Val626Met | Missense | YES | Male | 43 | 42.37721085 |

**Fig. S2.1** Table showing information for the samples from individuals with developmental disorders (total $N = 367$). Mutation information was annotated for the human genome assembly *hg19*. ASD: autism spectrum disorder; ATR-X: alpha thalassemia/mental retardation X-linked syndrome; FXS: fragile X syndrome.

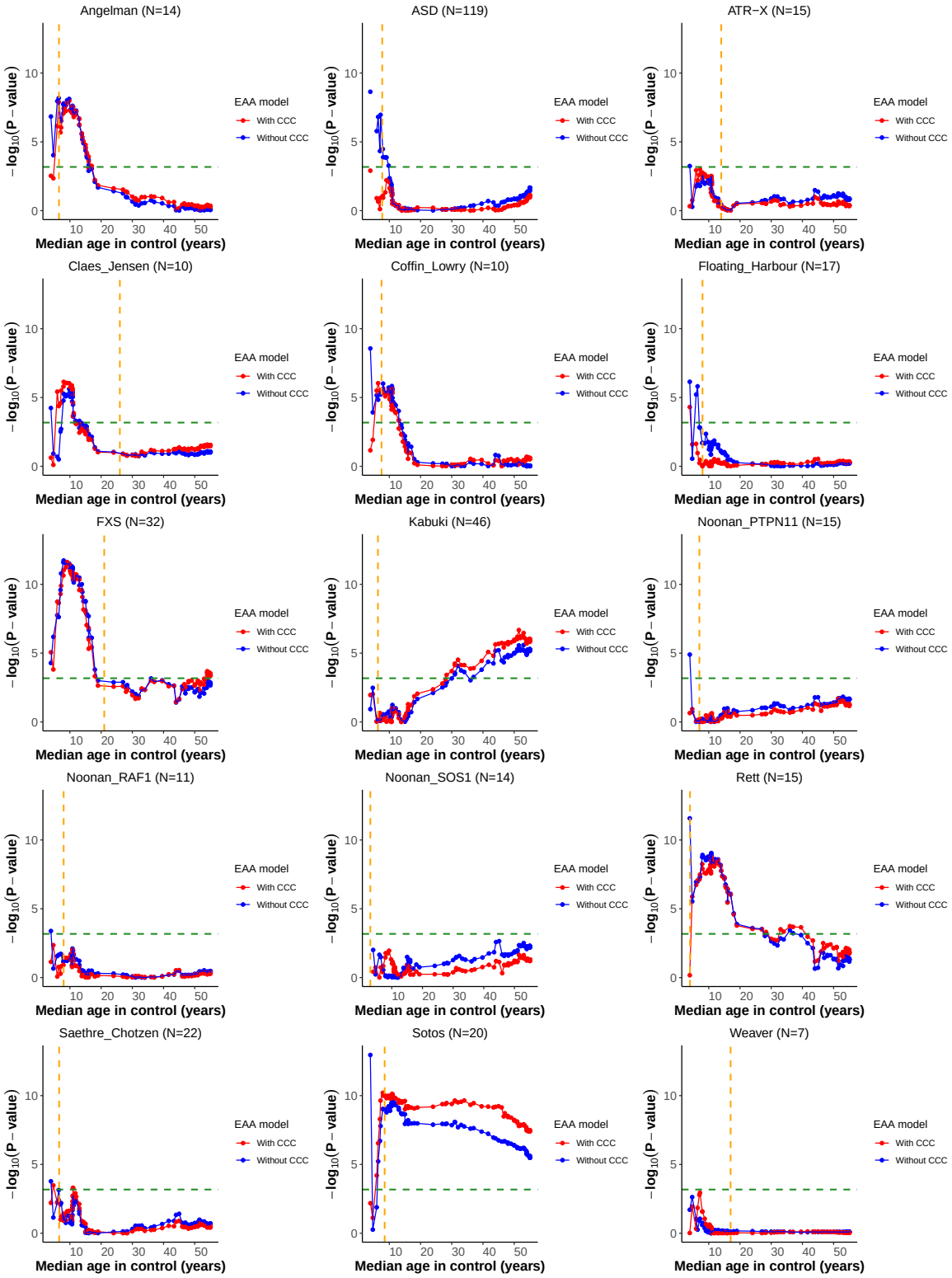
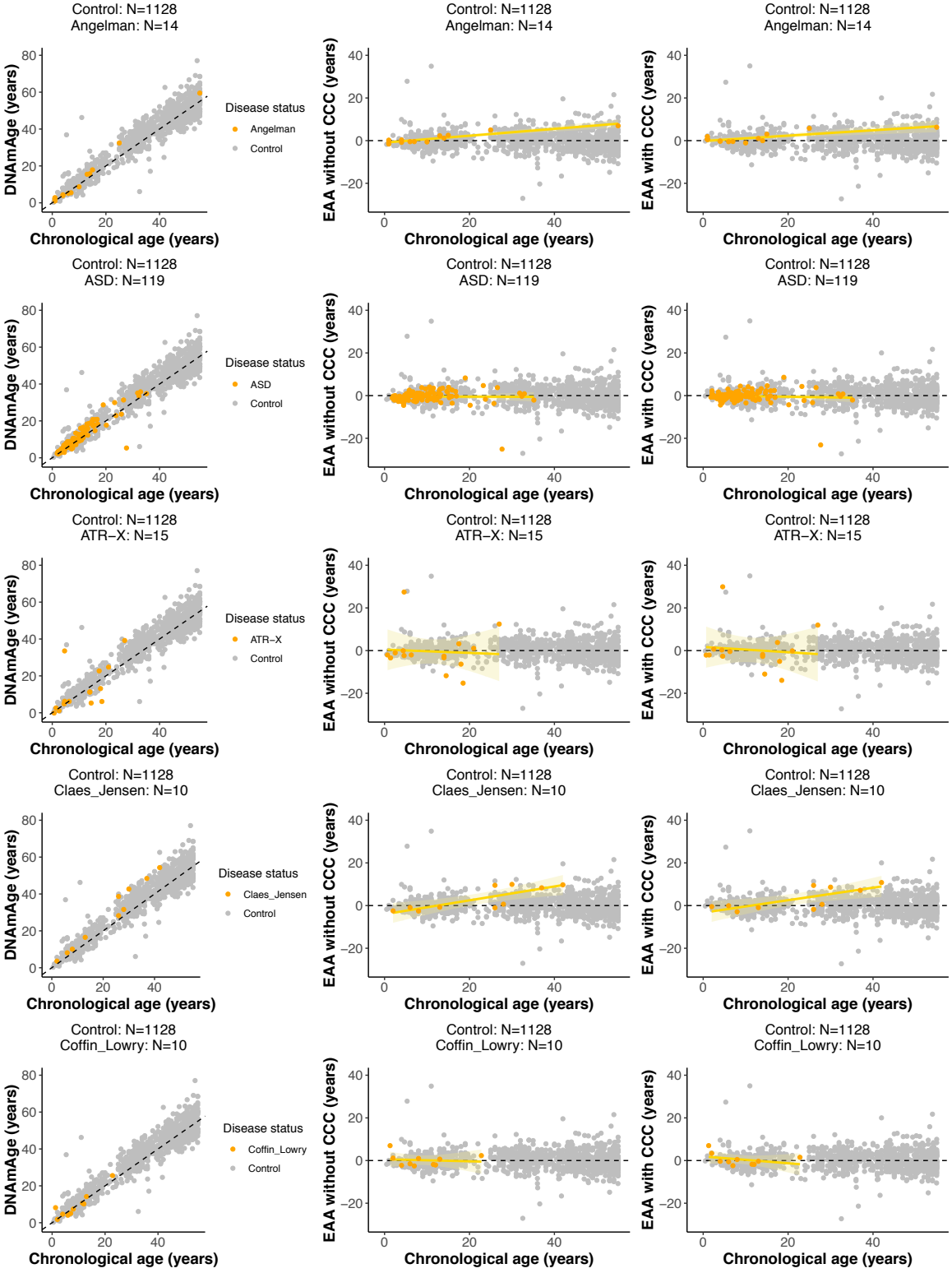
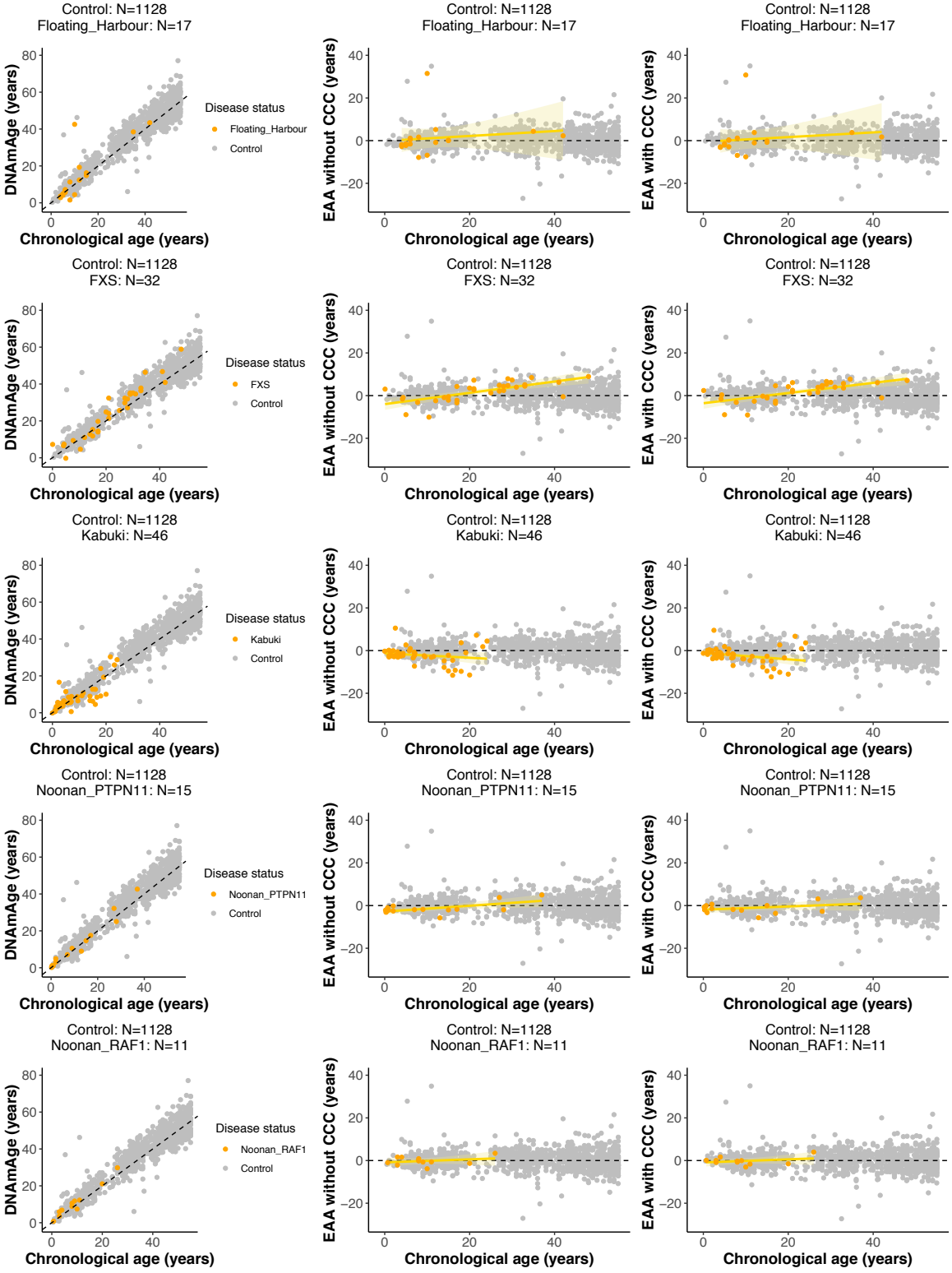



**Fig. S2.2** Effect of changing the median age of the controls when performing the screening for epigenetic age acceleration (EAA) in the different developmental disorders. The dashed green line displays the significance level of $\alpha = 0.01$ after Bonferroni correction. The dashed orange line displays the median age for the samples in the developmental disorder considered. In blue: EAA model without cell composition correction (CCC). In red: EAA model with CCC. ASD: autism spectrum disorder; ATR-X: alpha thalassemia/mental retardation X-linked syndrome; FXS: fragile X syndrome.


Control: N=1128
Angelman: N=14
Control: N=1128
ASD: N=119
Control: N=1128
ATR−X: N=15
Control: N=1128
Claes_Jensen: N=10
Control: N=1128
Coffin_Lowry: N=10
DNAmAge (years)
EAA without CCC (years)
EAA with CCC (years)
Chronological age (years)
Disease status
Angelman
ASD
ATR−X
Claes_Jensen
Coffin_Lowry
Control


Control: N=1128
Floating_Harbour: N=17
DNAmAge (years)
Chronological age (years)
Disease status
Floating_Harbour
Control
EAA without CCC (years)
EAA with CCC (years)
Control: N=1128
FXS: N=32
FXS
Control: N=1128
Kabuki: N=46
Kabuki
Control: N=1128
Noonan_PTPN11: N=15
Noonan_PTPN11
Control: N=1128
Noonan_RAF1: N=11
Noonan_RAF1

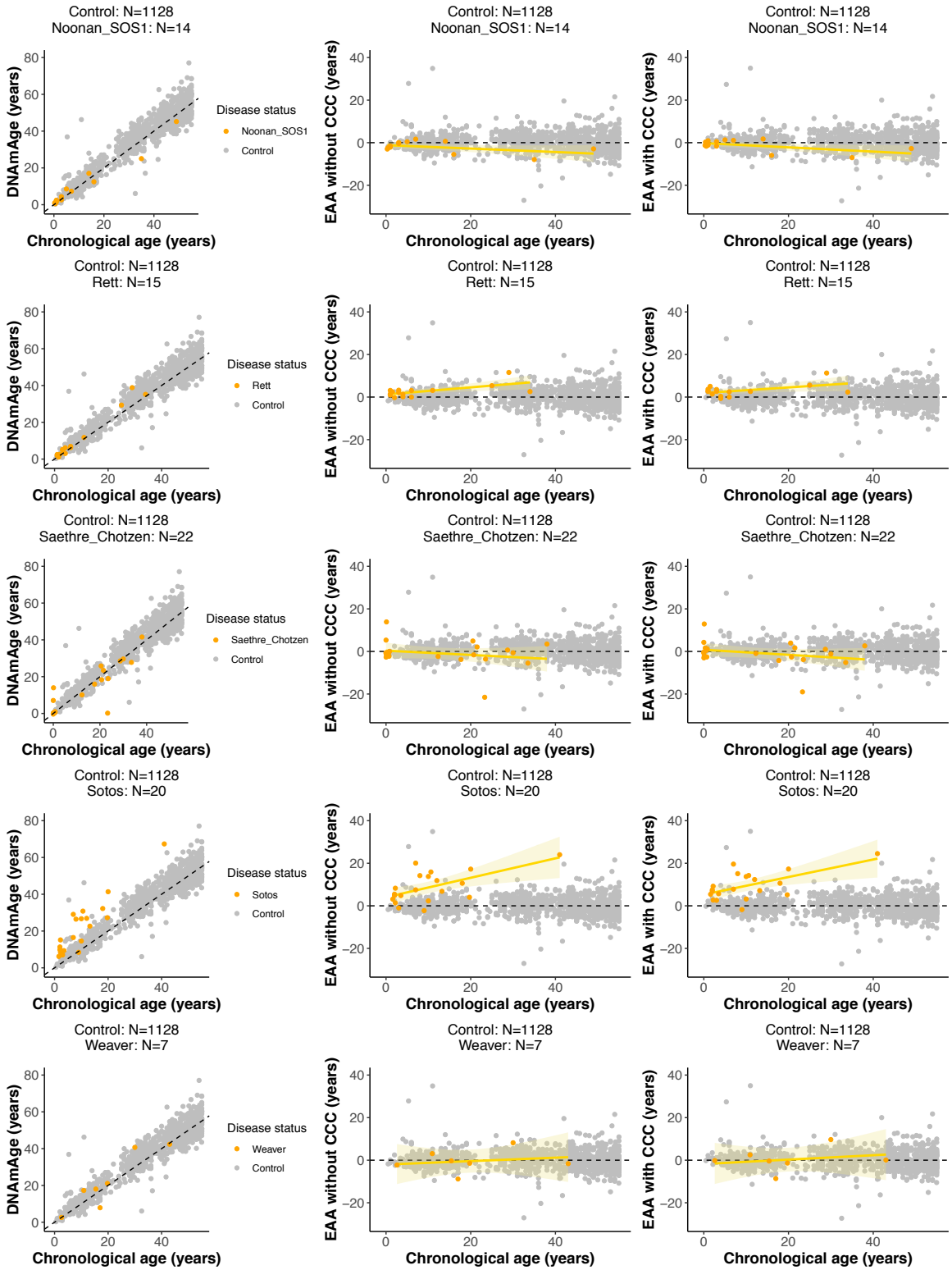



**Fig. S2.3** Screening for epigenetic age acceleration (EAA) in developmental disorders. Left panel: scatterplot showing the relation between epigenetic age (*DNAmAge*) according to Horvath's model and chronological age of the samples for a given developmental disorder (orange) and control (grey). Each sample is represented by one point. The black dashed line represents the diagonal to aid visualisation. Middle and right panels: scatterplots showing the relation between the epigenetic age acceleration (EAA) (without and with CCC respectively) and chronological age of the samples for a given developmental disorder (orange) and control (grey). Each sample is represented by one point. The yellow line represents the linear model EAA $\sim$ Age, with the standard error shown in the light yellow shade.

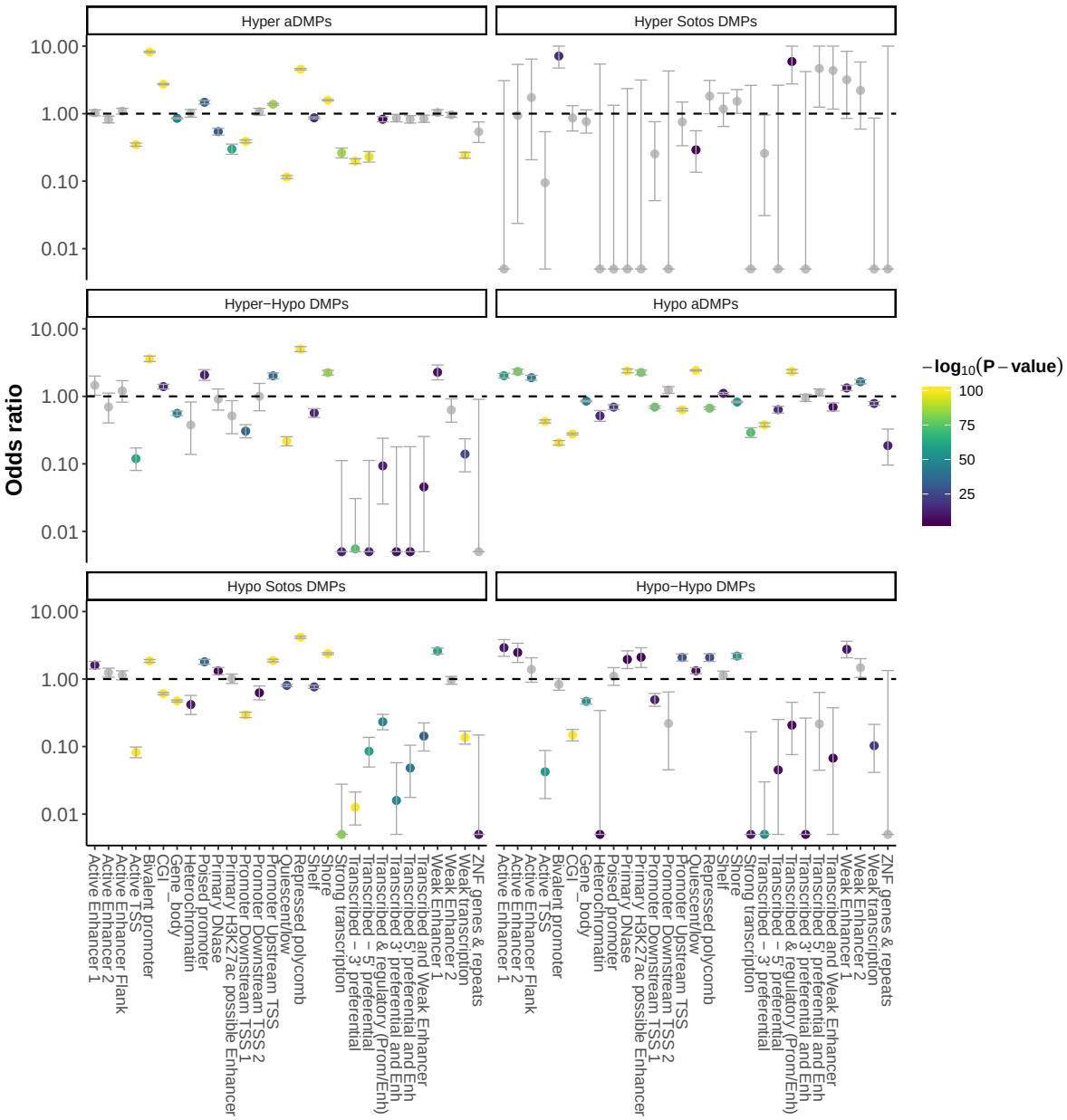



**Fig. S2.4** Enrichment for the categorical (epi)genomic features considered when comparing the different genome-wide subsets of differentially methylated positions (DMPs) in ageing and Sotos against a control (see section 3.7). The y-axis represents the odds ratio (OR), the error bars show the 95% confidence interval for the OR estimate and the colour of the points codes for $-\log_{10}$(p-value) obtained after testing for enrichment using Fisher's exact test. An OR > 1 shows that the given feature is enriched in the subset of DMPs considered, whilst an OR < 1 shows that it is found less than expected. The 'Hyper-Hypo DMPs' subset results from the intersection between the hypermethylated DMPs in ageing and the hypomethylated DMPs in Sotos. The 'Hypo-Hypo DMPs' subset results from the intersection between the hypomethylated DMPs in ageing and Sotos. In grey: features that did not reach significance using a significance level of $\alpha = 0.01$ after Bonferroni correction.

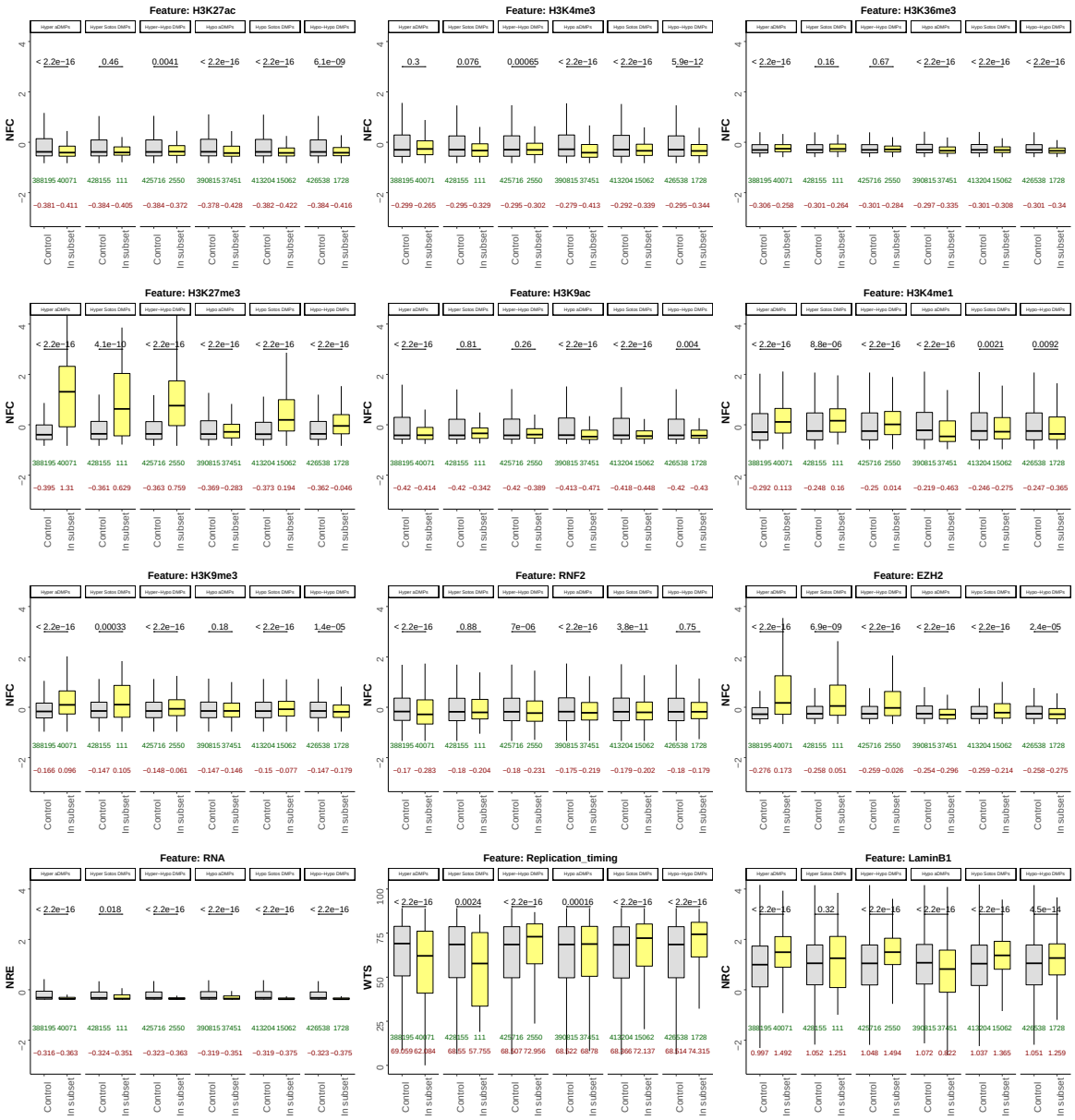



**Fig. S2.5** Boxplots showing the distributions of scores for the continuous (epi)genomic features considered when comparing the different genome-wide subsets of differentially methylated positions (DMPs) in ageing and Sotos against a control (see section 3.7). The p-values (two-sided Wilcoxon's test, before multiple testing correction) are shown above the boxplots. The number of DMPs belonging to each subset (in green) and the median value of the feature score (in dark red) are shown below the boxplots. NFC: 'normalised fold change'; NRE: 'normalised RNA expression'; WTS: 'wavelet-transformed signals'; NRC: 'normalised read counts'.

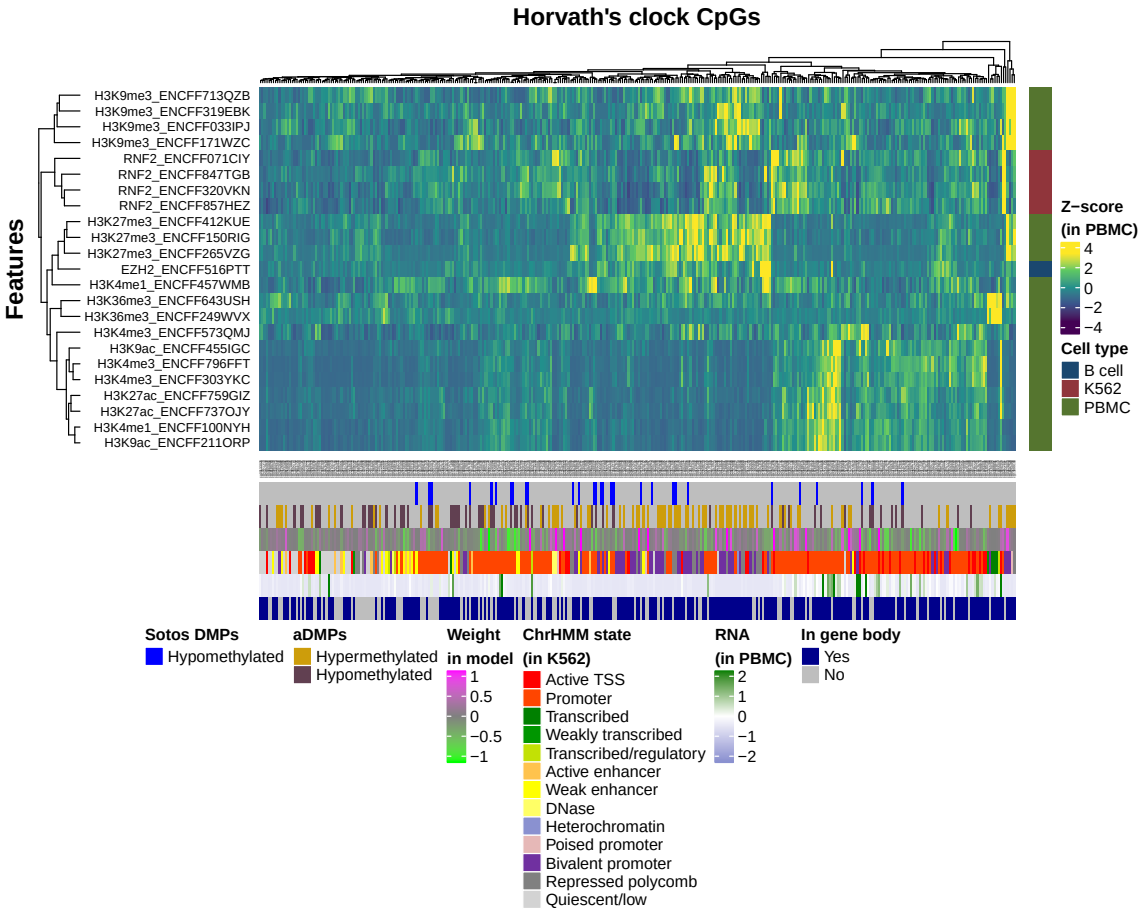



**Fig. S2.6** Heatmap displaying the scores for the different continuous (epi)genomic features (rows) in each one of the 353 Horvath's epigenetic clock CpGs (columns). The names of the features include the ENCODE ID (see Fig. S2.11). Hierarchical clustering was performed in both rows and columns. RNA refers to the 'normalised RNA expression' (NRE). aDMPs: differentially methylated positions during ageing. PBMC: peripheral blood mononuclear cells.

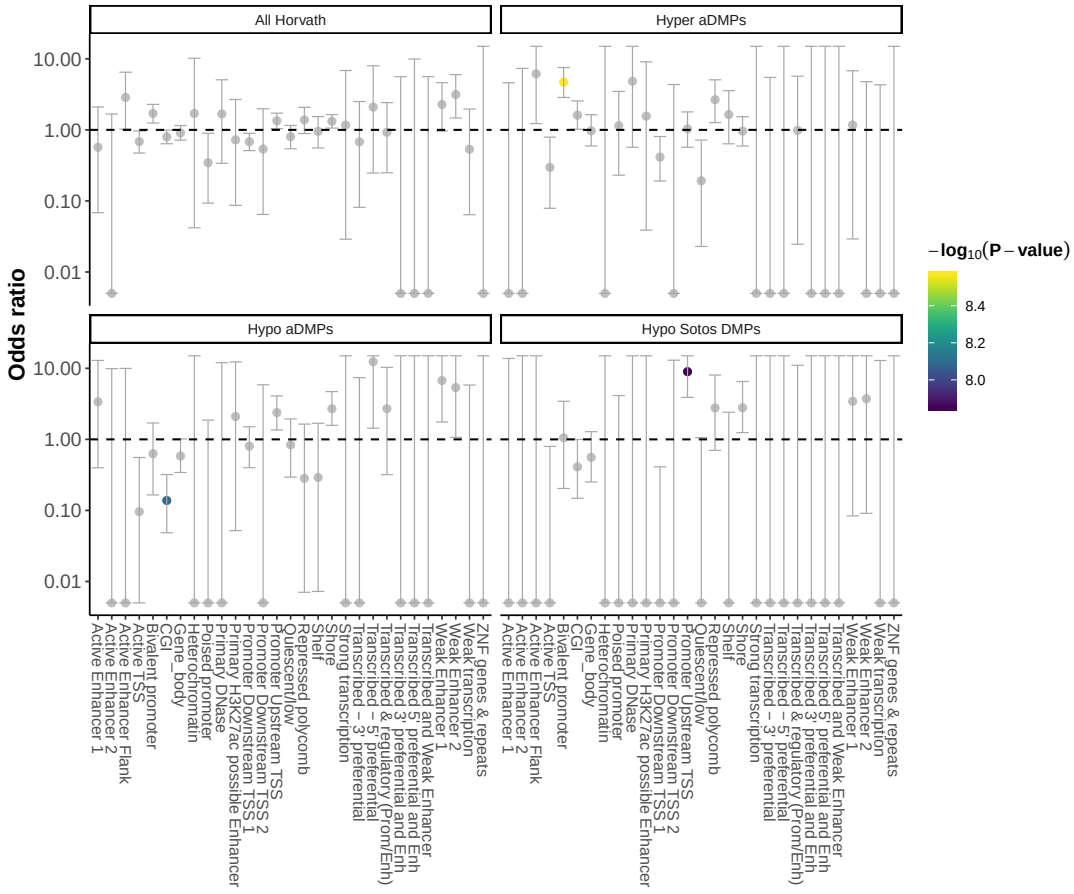



**Fig. S2.7** As in Fig. S2.4., but focused on the 353 Horvath's epigenetic clock CpG sites.

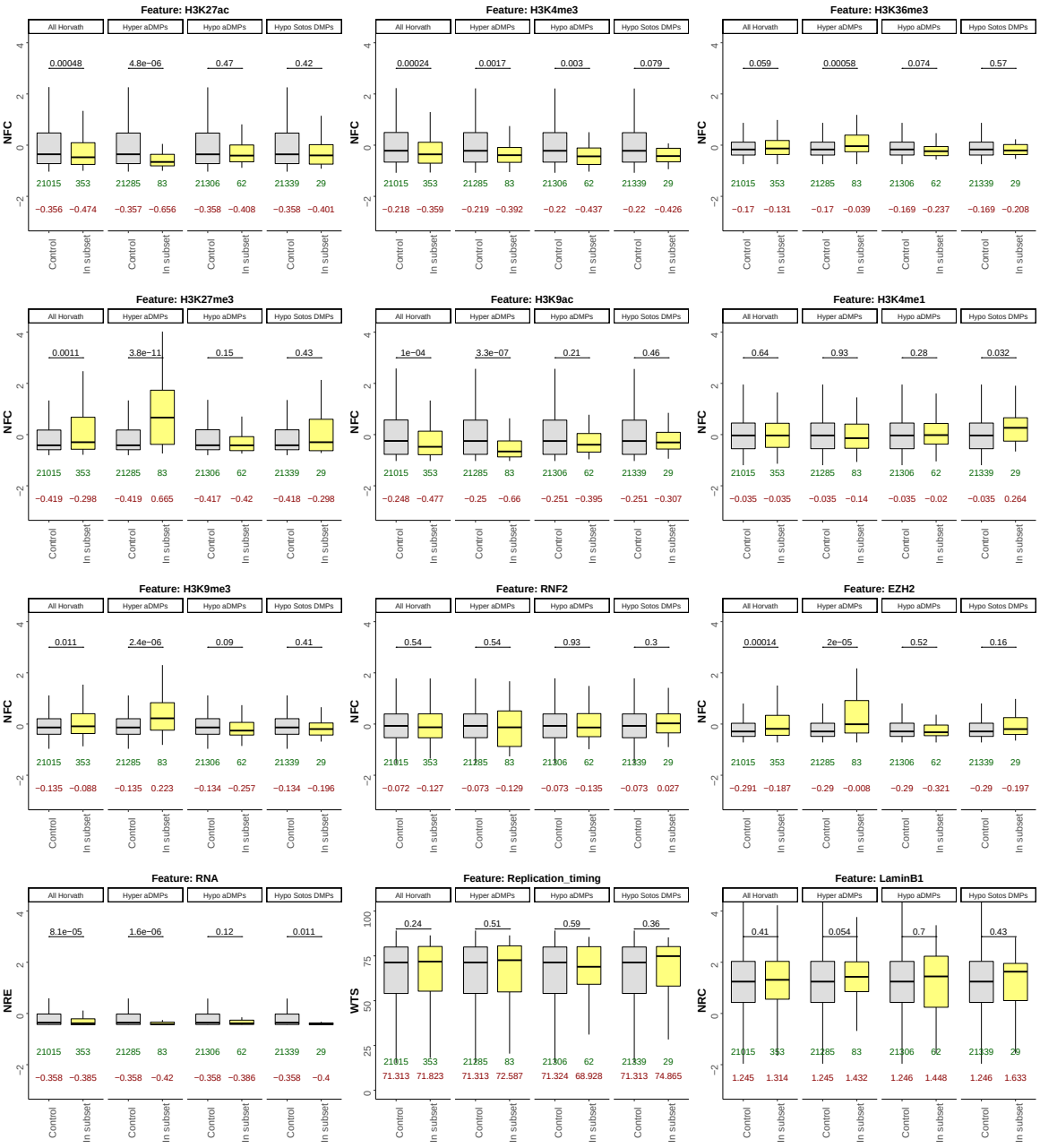



**Fig. S2.8** As in Fig. S2.5., but focused on the 353 Horvath's epigenetic clock CpG sites.

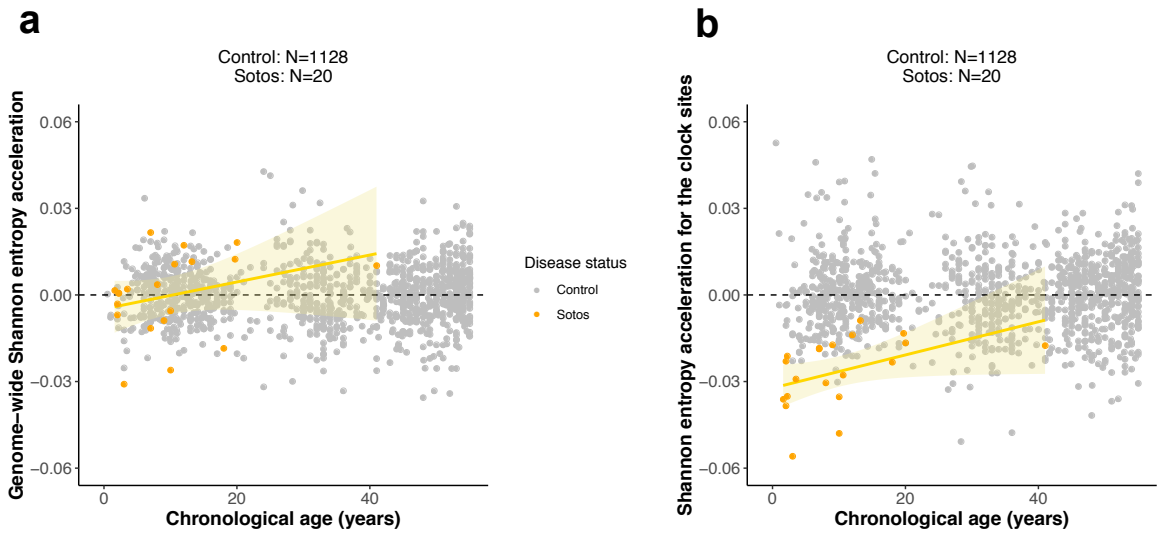



**Fig. S2.9** Methylation Shannon entropy acceleration. **a.** Scatterplot showing the relationship between the genome-wide Shannon entropy acceleration (gSEA) and chronological age of the samples for Sotos (orange) and healthy controls (grey). Each sample is represented by one point. The yellow line represents the linear model gSEA $\sim$ Age, with the standard error shown in the light yellow shade. **b.** As in a., but using the Shannon entropy acceleration calculated only for the 353 CpG sites in the Horvath's epigenetic clock (cSEA).

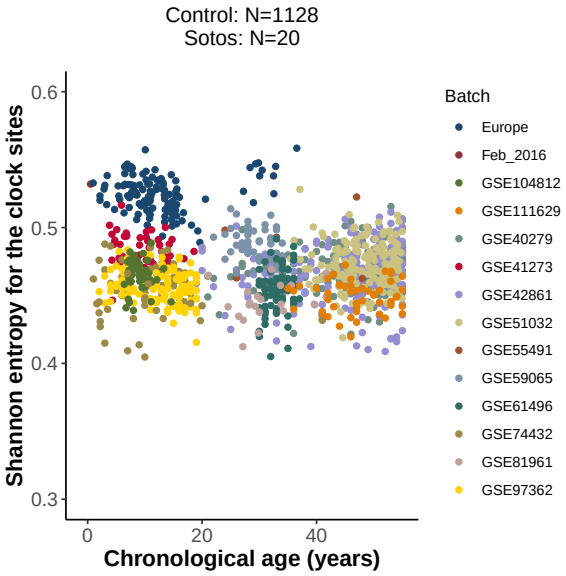



**Fig. S2.10** Scatterplot showing the effects of the different batches on the methylation Shannon entropy calculations for the 353 Horvath's epigenetic clock sites. Each sample is represented by one point and coloured according to the batch that they belong to.

| File ID | Feature type | Data type | Tissue | Age (years) | Sex | Source |
|---|---|---|---|---|---|---|
| ENCFF516PTT | EZH2 | fold change over control | B cell | 27 | Female | ENCODE |
| ENCFF071CIY | RNF2 | fold change over control | K562 | NA | NA | ENCODE |
| ENCFF857HEZ | RNF2 | fold change over control | K562 | NA | NA | ENCODE |
| ENCFF320VKN | RNF2 | fold change over control | K562 | NA | NA | ENCODE |
| ENCFF847TGB | RNF2 | fold change over control | K562 | NA | NA | ENCODE |
| ENCFF737OJY | H3K27ac | fold change over control | PBMC | 32 | Male | ENCODE |
| ENCFF303YKC | H3K4me3 | fold change over control | PBMC | 32 | Male | ENCODE |
| ENCFF643USH | H3K36me3 | fold change over control | PBMC | 32 | Male | ENCODE |
| ENCFF249WVX | H3K36me3 | fold change over control | PBMC | 28 | Male | ENCODE |
| ENCFF759GIZ | H3K27ac | fold change over control | PBMC | 28 | Female | ENCODE |
| ENCFF412KUE | H3K27me3 | fold change over control | PBMC | 32 | Male | ENCODE |
| ENCFF455IGC | H3K9ac | fold change over control | PBMC | 28 | Male | ENCODE |
| ENCFF457WMB | H3K4me1 | fold change over control | PBMC | 32 | Male | ENCODE |
| ENCFF211ORP | H3K9ac | fold change over control | PBMC | 27 | Male | ENCODE |
| ENCFF171WZC | H3K9me3 | fold change over control | PBMC | 27 | Male | ENCODE |
| ENCFF573QMJ | H3K4me3 | fold change over control | PBMC | 27 | Male | ENCODE |
| ENCFF150RIG | H3K27me3 | fold change over control | PBMC | 28 | Female | ENCODE |
| ENCFF033IPJ | H3K9me3 | fold change over control | PBMC | 28 | Female | ENCODE |
| ENCFF796FFT | H3K4me3 | fold change over control | PBMC | 28 | Female | ENCODE |
| ENCFF100NYH | H3K4me1 | fold change over control | PBMC | 27 | Male | ENCODE |
| ENCFF713QZB | H3K9me3 | fold change over control | PBMC | 32 | Male | ENCODE |
| ENCFF265VZG | H3K27me3 | fold change over control | PBMC | 28 | Male | ENCODE |
| ENCFF319EBK | H3K9me3 | fold change over control | PBMC | 28 | Male | ENCODE |
| ENCFF754LBN | RNA-seq | minus strand signal of unique reads | PBMC | 52 | Female | ENCODE |
| ENCFF398HDS | RNA-seq | plus strand signal of unique reads | PBMC | 52 | Female | ENCODE |
| GSM923447 | Replication timing | Wavelet-transformed signals | IMR90 | NA | Female | GEO |
| GSM1289416 | LaminB1 | Normalised read counts | IMR90 | NA | NA | GEO |

**Fig. S2.11** Information (including the source) about the continuous (epi)genomic features (ChIP-seq and RNA-seq data) that were included in my analysis to annotate the different sets of CpG sites. All the data were mapped to the *hg19* assembly of the human genome. PBMC: peripheral blood mononuclear cells.

# S.3 Supplementary for chapter 4

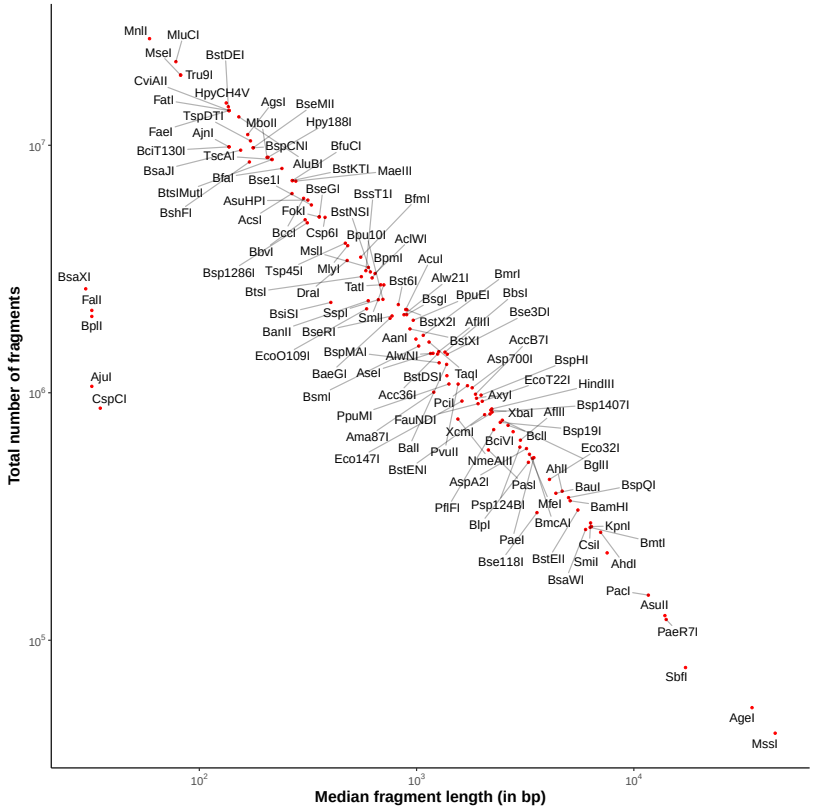



**Fig. S3.1** Scatterplot which summarises the fragment length distributions for the same isoschizomer families portrayed in Fig 4.2a. The red dots represent the actual values of median fragment length and total number of fragments for each family. The black lines assign each name label to the correspondent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

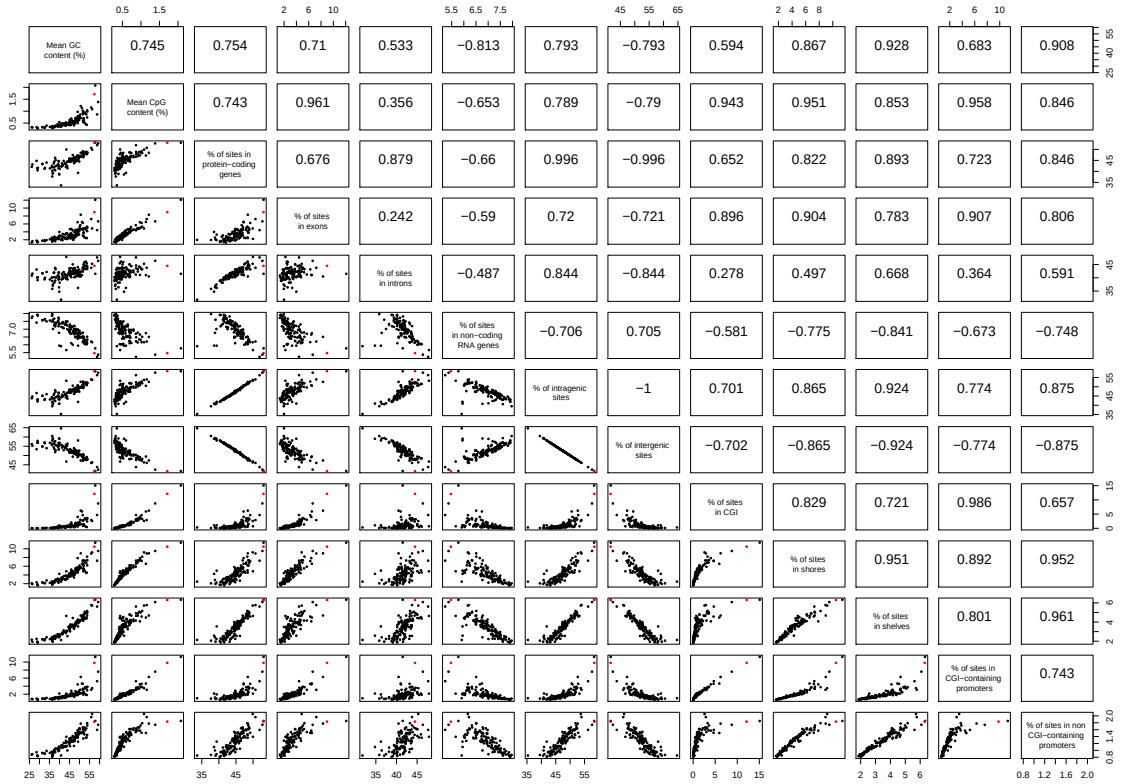



**Fig. S3.2** Matrix of scatterplots showing the percentages of cleavage sites from different restriction enzymes that overlap with several genomic features (listed on the diagonal) in the human genome (hg38). The red dot in each scatterplot represents the values for MspI. The numbers above the diagonal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

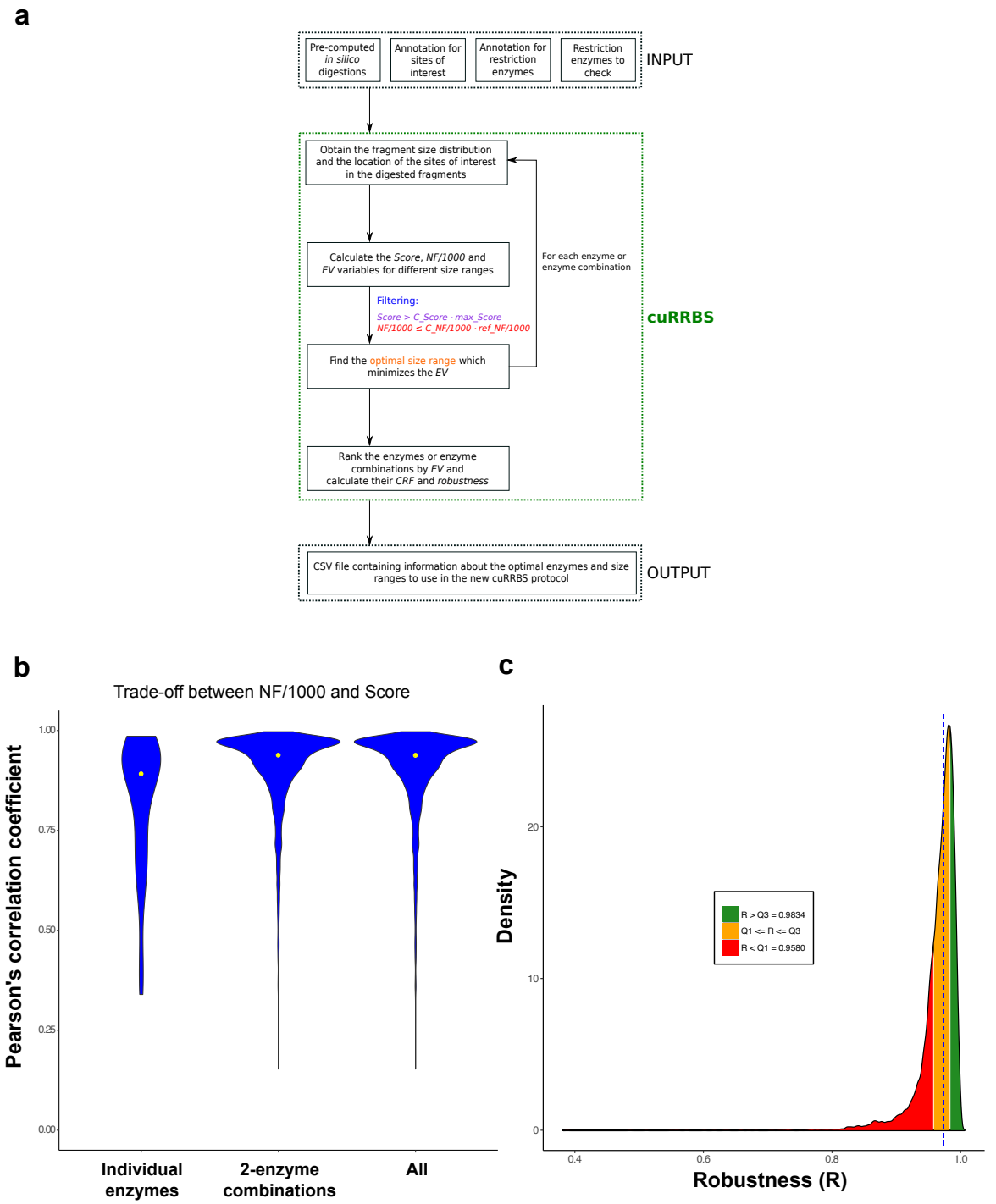



**Fig. S3.4** Additional insights into cuRRBS. **a.** Detailed flowchart showing the input, main steps in cuRRBS and the output of the software. **b.** Violin plots showing the distribution of Pearson's correlation coefficients between the number of fragments (*NF*) and the *Score* for all the different enzymes tested with cuRRBS (single-enzyme, double-enzyme, all). In this example we used the Horvath epigenetic clock system [209], 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* ($\delta$) of 20 bp. cuRRBS was run for all the biological systems under study (Fig. S3.5) [209, 407–412] 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 [209, 407–412] for which cuRRBS was run *in silico*. Some variables from the top hits in cuRRBS output are also reported.

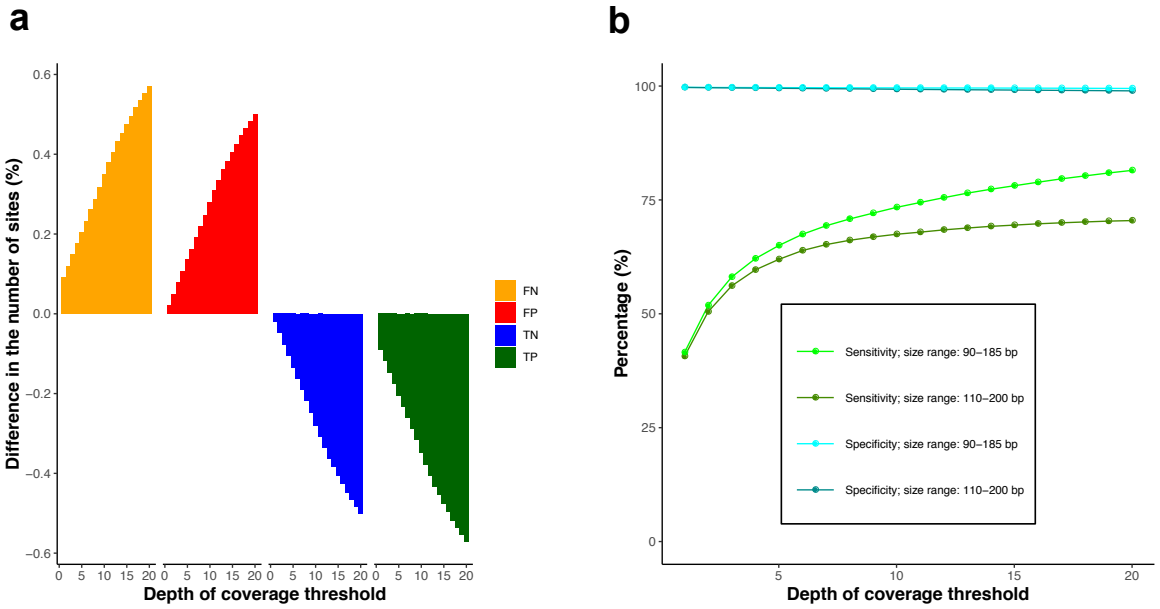



**Fig. S3.6** Effect of experimental errors during size selection in cuRRBS predictions. **a.** Barplots showing the difference in the number of true positives (TP, in green), true negatives (TN, in blue), false positives (FP, in red) and false negatives (FN, in yellow) derived from cuRRBS theoretical predictions for the XmaI-RRBS data [399] 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 [399]. 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.

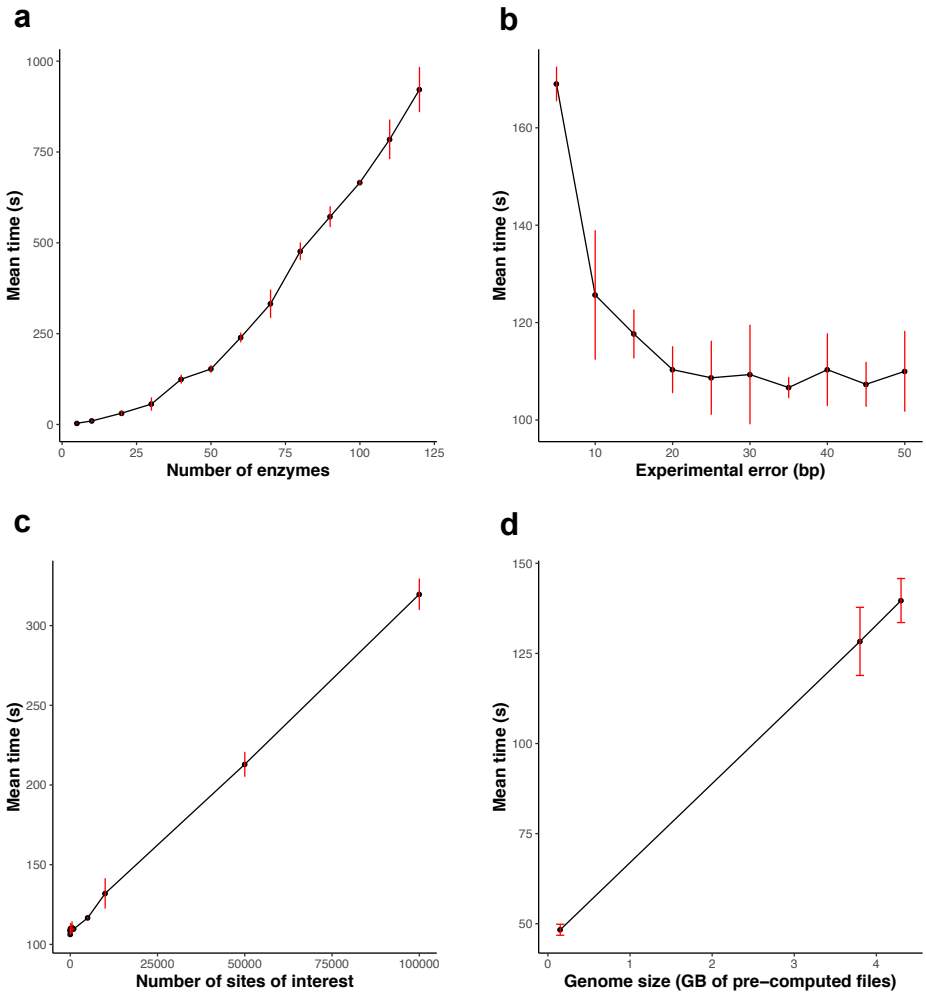



**Fig. S3.7** cuRRBS computational efficiency. **a.** Plot showing the dependency between the number of enzymes checked and the computational (real) time required by the software (mean between 3 independent runs). cuRRBS was run for the Horvath epigenetic clock system [209] with a *read length* of 75 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 [209] 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] Leonard Hayflick. Biological aging is no longer an unsolved problem. In *Annals of the New York Academy of Sciences*, volume 1100, pages 1–13, 2007.

[2] Colin Renfrew, Michael J. Boyd, and Iain Morley. *Death Rituals, Social Order and the Archeology of Immortality in the Ancient World*. 2016.

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

[4] Zhores A Medvedev. An attempt at a rational classification of theories of ageing. *Biological Reviews*, 65:375–398, 1990.

[5] Matthew Witten. Information content of biological survival curves arising in aging experiments: some further thoughts. In *Evolution of longevity in animals: a comparative approach.*, pages 295–317. 1986.

[6] Owen R Jones, Alexander Scheuerlein, Roberto Salguero-Gómez, Carlo Giovanni Camarda, Ralf Schaible, Brenda B Casper, Johan P Dahlgren, Johan Ehrlén, María B García, Eric S Menges, Pedro F Quintana-Ascencio, Hal Caswell, Annette Baudisch, and James W Vaupel. Diversity of ageing across the tree of life. *Nature*, 505:169, 2013.

[7] J P De Magalhães and J Costa. A database of vertebrate longevity records and their relation to other life-history traits. *Journal of Evolutionary Biology*, 22(8):1770–1774, 2009.

[8] C E Finch. Update on Slow Aging and Negligible Senescence – A Mini-Review. *Gerontology*, 55(3):307–313, 2009.

[9] Leonard Hayflick. Entropy Explains Aging, Genetic Determinism Explains Longevity, and Undefined Terminology Explains Misunderstanding Both. *PLOS Genetics*, 3(12):e220, dec 2007.

[10] T B L Kirkwood. Evolution of ageing. *Nature*, 270(5635):301–304, 1977.

[11] T. B.L. Kirkwood and M. R. Rose. Evolution of senescence: late survival sacrificed for reproduction. *Philosophical Transactions - Royal Society of London, B*, 332(1262):15–24, 1991.

[12] Mikhail V. Blagosklonny. Aging and immortality: Quasi-programmed senescence and its pharmacologic inhibition, 2006.

[13] Mikhail V Blagosklonny. Revisiting the antagonistic pleiotropy theory of aging: TOR-driven program and quasi-program. *Cell Cycle*, 9(16):3171–3176, aug 2010.

[14] João Pedro de Magalhães. Programmatic features of aging originating in development: aging mechanisms beyond molecular damage? *The FASEB Journal*, 26(12):4821–4826, 2012.

[15] David Gems. The aging-disease false dichotomy: understanding senescence as pathology. *Frontiers in genetics*, 6(June):212, 2015.

[16] G. C. Williams. Pleiotropy, Natural Selection, and the Evolution of Senescence. *Evolution*, 11(4):398–411, 1957.

[17] Robert E Ricklefs. Life-history connections to rates of aging in terrestrial vertebrates. *Proceedings of the National Academy of Sciences*, 107(22):10314–10319, 2010.

[18] Richard Peto and Richard Doll. There is no such thing as aging. *BMJ*, 315(7115):1030, 1997.

[19] Nicholas Stroustrup, Winston E Anthony, Zachary M Nash, Vivek Gowda, Adam Gomez, Isaac F López-Moyado, Javier Apfeld, and Walter Fontana. The temporal scaling of Caenorhabditis elegans ageing. *Nature*, 530:103–107, 2016.

[20] Adam Freund. Untangling Aging Using Dynamic, Organism-Level Phenotypic Networks. *Cell Systems*, 8(3):172–181, 2019.

[21] Cynthia Kenyon. The plasticity of aging: Insights from long-lived mutants. *Cell*, 120(4):449–460, 2005.

[22] C M McCay, L A Maynard, and Mary F Crowell. The Effect of Retarded Growth Upon the Length of Life Span and Upon the Ultimate Body Size: One Figure. *The Journal of Nutrition*, 10(1):63–79, 1935.

[23] Roger B. McDonald and Jon J. Ramsey. Honoring Clive McCay and 75 Years of Calorie Restriction Research. *The Journal of Nutrition*, 140(7):1205–1210, 2010.

[24] Luigi Fontana and Linda Partridge. Promoting health and longevity through diet: From model organisms to humans. *Cell*, 161(1):106–118, 2015.

[25] Michael Klass and David Hirsh. Non-ageing developmental variant of Caenorhabditis elegans. *Nature*, 260(5551):523–525, 1976.

[26] Thomas E Johnson. 25 years after age-1: Genes, interventions and the revolution in aging research. *Experimental Gerontology*, 48(7):640–643, 2013.

[27] Cynthia Kenyon, Jean Chang, Erin Gensch, Adam Rudner, and Ramon Tabtiang. A C. elegans mutant that lives twice as long as wild type. *Nature*, 366(6454):461–464, 1993.

[28] Jason Z Morris, Heidi A Tissenbaum, and Gary Ruvkun. A phosphatidylinositol-3-OH kinase family member regulating longevity and diapause in Caenorhabditis elegans. *Nature*, 382(6591):536–539, 1996.

[29] Cynthia J Kenyon. The genetics of ageing. *Nature*, 464(7288):504–12, 2010.

[30] Param Priya Singh, Brittany A Demmitt, Ravi D Nath, and Anne Brunet. The Genetics of Aging: A Vertebrate Perspective. *Cell*, 177(1):200–220, 2019.

[31] Eric L. Greer and Anne Brunet. Signaling networks in aging. *Journal of Cell Science*, 121:407–412, 2008.

[32] Leonard Guarente and Cynthia Kenyon. Genetic pathways that regulate ageing in model organisms. *Nature*, 408(6809):255–262, 2000.

[33] Kui Lin, Honor Hsin, Natasha Libina, and Cynthia Kenyon. Regulation of the Caenorhabditis elegans longevity protein DAF-16 by insulin/IGF-1 and germline signaling. *Nature Genetics*, 28:139–145, 2001.

[34] Rute Martins, Gordon J Lithgow, and Wolfgang Link. Long live FOXO: unraveling the role of FOXO proteins in aging and longevity. *Aging Cell*, 15(2):196–207, 2016.

[35] Ao-Lin Hsu, Coleen T Murphy, and Cynthia Kenyon. Regulation of Aging and Age-Related Disease by DAF-16 and Heat-Shock Factor. *Science*, 300(5622):1142–1145, 2003.

[36] Jennifer M A Tullet, Maren Hertweck, Jae Hyung An, Joseph Baker, Ji Yun Hwang, Shu Liu, Riva P Oliveira, Ralf Baumeister, and T Keith Blackwell. Direct Inhibition of the Longevity-Promoting Factor SKN-1 by Insulin-like Signaling in C. elegans. *Cell*, 132(6):1025–1038, 2008.

[37] Sung Hee Um, David D'Alessio, and George Thomas. Nutrient overload, insulin resistance, and ribosomal protein S6 kinase 1, S6K1. *Cell Metabolism*, 3(6):393–402, 2006.

[38] David E Harrison, Randy Strong, Zelton Dave Sharp, James F Nelson, Clinton M Astle, Kevin Flurkey, Nancy L Nadon, J Erby Wilkinson, Krystyna Frenkel, Christy S Carter, Marco Pahor, Martin a Javors, Elizabeth Fernandez, and Richard a Miller. Rapamycin fed late in life extends lifespan in genetically heterogeneous mice. *Nature*, 460(7253):392–395, 2009.

[39] Maria M Mihaylova and Reuben J Shaw. The AMPK signalling pathway coordinates cell growth, autophagy and metabolism. *Nature Cell Biology*, 13:1016–1023, 2011.

[40] Vladimir N Anisimov, Lev M Berstein, Peter A Egormin, Tatiana S Piskunova, Irina G Popovich, Mark A Zabezhinski, Margarita L Tyndyk, Maria V Yurova, Irina G Kovalenko, Tatiana E Poroshina, and Anna V Semenchenko. Metformin slows down aging and extends life span of female SHR mice. *Cell Cycle*, 7(17):2769–2773, 2008.

[41] Alejandro Martin-Montalvo, Evi M Mercken, Sarah J Mitchell, Hector H Palacios, Patricia L Mote, Morten Scheibye-Knudsen, Ana P Gomes, Theresa M Ward, Robin K Minor, Marie-José Blouin, Matthias Schwab, Michael Pollak, Yongqing Zhang, Yinbing Yu, Kevin G Becker, Vilhelm A Bohr, Donald K Ingram, David A Sinclair, Norman S Wolf, Stephen R Spindler, Michel Bernier, and Rafael de Cabo. Metformin improves healthspan and lifespan in mice. *Nature Communications*, 4:2192, 2013.

[42] Nir Barzilai, Jill P Crandall, Stephen B Kritchevsky, and Mark A Espeland. Metformin as a Tool to Target Aging. *Cell Metabolism*, 23(6):1060–1065, 2016.

[43] Michael S Bonkowski and David A Sinclair. Slowing ageing by design: the rise of NAD+ and sirtuin-activating compounds. *Nature Reviews Molecular Cell Biology*, 17:679–690, 2016.

[44] Marco Lezzerini and Yelena Budovskaya. A dual role of the Wnt signaling pathway during aging in Caenorhabditis elegans. *Aging Cell*, 13(1):8–18, 2014.

[45] Sean P Curran, Xiaoyun Wu, Christian G Riedel, and Gary Ruvkun. A soma-to-germline transformation in long-lived Caenorhabditis elegans mutants. *Nature*, 459:1079–1084, 2009.

[46] Andrew Dillin, Douglas K Crawford, and Cynthia Kenyon. Timing Requirements for Insulin/IGF-1 Signaling in C. elegans. *Science*, 298(5594):830–834, 2002.

[47] Eric L. Greer, Philip R. Oskoui, Max R. Banko, Jay M. Maniar, Melanie P. Gygi, Steven P. Gygi, and Anne Brunet. The energy sensor AMP-activated protein kinase directly regulates the mammalian FOXO3 transcription factor. *Journal of Biological Chemistry*, 282:30107–30119, 2007.

[48] Nuno Arantes-Oliveira, Jennifer R Berman, and Cynthia Kenyon. Healthy Animals with Extreme Longevity. *Science*, 302(5645):611, 2003.

[49] Srinivas Ayyadevara, Ramani Alla, John J Thaden, and Robert J Shmookler Reis. Remarkable longevity and stress resistance of nematode PI3K-null mutants. *Aging Cell*, 7(1):13–22, 2008.

[50] J Graham Ruby, Megan Smith, and Rochelle Buffenstein. Naked mole-rat mortality rates defy Gompertzian laws by not increasing with age. *eLife*, 7:e31157, 2018.

[51] Toni Fleischer, Jutta Gampe, Alexander Scheuerlein, and Gerald Kerth. Rare catastrophic events drive population dynamics in a bat species with negligible senescence. *Scientific Reports*, 7(1):7370, 2017.

[52] Iñigo Martincorena, Joanna C Fowler, Agnieszka Wabik, Andrew R J Lawson, Federico Abascal, Michael W J Hall, Alex Cagan, Kasumi Murai, Krishnaa Mahbubani, Michael R Stratton, Rebecca C Fitzgerald, Penny A Handford, Peter J Campbell, Kourosh Saeb-Parsy, and Philip H Jones. Somatic mutant clones colonize the human esophagus with age. *Science*, 362(6417):911–917, 2018.

[53] Nils-Göran Larsson. Somatic Mitochondrial DNA Mutations in Mammalian Aging. *Annual Review of Biochemistry*, 79(1):683–706, 2010.

[54] Ludmil B Alexandrov and Michael R Stratton. Mutational signatures: the patterns of somatic mutations hidden in cancer genomes. *Current Opinion in Genetics & Development*, 24:52–60, 2014.

[55] Ludmil B Alexandrov, Philip H Jones, David C Wedge, Julian E Sale, Peter J Campbell, Serena Nik-Zainal, and Michael R Stratton. Clock-like mutational processes in human somatic cells. *Nature Genetics*, 47(12):1402–1407, 2015.

[56] Philipp Oberdoerffer and David A Sinclair. The role of nuclear architecture in genomic instability and ageing. *Nature Reviews Molecular Cell Biology*, 8:692–702, 2007.

[57] William C. Orr. Tightening the connection between transposable element mobilization and aging. *Proceedings of the National Academy of Sciences*, 113(40):11069–11070, 2016.

[58] Roderick J O'Sullivan and Jan Karlseder. Telomeres: protecting chromosomes against genome instability. *Nature Reviews Molecular Cell Biology*, 11:171–181, 2010.

[59] L Hayflick and P S Moorhead. The serial cultivation of human diploid cell strains. *Experimental Cell Research*, 25(3):585–621, 1961.

[60] Leonard Hayflick. A Brief History of the Mortality and Immortality of Cultured Cells. *The Keio Journal of Medicine*, 47(3):174–182, 1998.

[61] Maria A Blasco. Telomere length, stem cells and aging. *Nature Chemical Biology*, 3:640, sep 2007.

[62] Antonia Tomás-Loba, Ignacio Flores, Pablo J Fernández-Marcos, María L Cayuela, Antonio Maraver, Agueda Tejera, Consuelo Borrás, Ander Matheu, Peter Klatt, Juana M Flores, José Viña, Manuel Serrano, and Maria A Blasco. Telomerase Reverse Transcriptase Delays Aging in Cancer-Resistant Mice. *Cell*, 135(4):609–622, 2008.

[63] Nicolás Herranz and Jesús Gil. Mechanisms and functions of cellular senescence. *The Journal of Clinical Investigation*, 128(4):1238–1246, 2018.

[64] Darren J Baker, Tobias Wijshake, Tamar Tchkonia, Nathan K LeBrasseur, Bennett G Childs, Bart van de Sluis, James L Kirkland, and Jan M van Deursen. Clearance of p16Ink4a-positive senescent cells delays ageing-associated disorders. *Nature*, 479:232–236, 2011.

[65] Darren J Baker, Bennett G Childs, Matej Durik, Melinde E Wijers, Cynthia J Sieben, Jian Zhong, Rachel A. Saltness, Karthik B Jeganathan, Grace Casaclang Verzosa, Abdulmohammad Pezeshki, Khashayarsha Khazaie, Jordan D Miller, and Jan M van Deursen. Naturally occurring p16Ink4a-positive cells shorten healthy lifespan. *Nature*, 530:184–189, 2016.

[66] Ming Xu, Tamar Pirtskhalava, Joshua N Farr, Bettina M Weigand, Allyson K Palmer, Megan M Weivoda, Christina L Inman, Mikolaj B Ogrodnik, Christine M Hachfeld, Daniel G Fraser, Jennifer L Onken, Kurt O Johnson, Grace C Verzosa, Larissa G P Langhi, Moritz Weigl, Nino Giorgadze, Nathan K LeBrasseur, Jordan D Miller, Diana Jurk, Ravinder J Singh, David B Allison, Keisuke Ejima, Gene B Hubbard, Yuji Ikeno, Hajrunisa Cubro, Vesna D Garovic, Xiaonan Hou, S John Weroha, Paul D Robbins, Laura J Niedernhofer, Sundeep Khosla, Tamara Tchkonia, and James L Kirkland. Senolytics improve physical function and increase lifespan in old age. *Nature Medicine*, 24(8):1246–1256, 2018.

[67] James L Kirkland, Tamara Tchkonia, Yi Zhu, Laura J Niedernhofer, and Paul D Robbins. The Clinical Potential of Senolytic Drugs. *Journal of the American Geriatrics Society*, 65(10):2297–2301, 2017.

[68] Marco De Cecco, Takahiro Ito, Anna P Petrashen, Amy E Elias, Nicholas J Skvir, Steven W Criscione, Alberto Caligiana, Greta Brocculi, Emily M Adney, Jef D Boeke, Oanh Le, Christian Beauséjour, Jayakrishna Ambati, Kameshwari Ambati, Matthew Simon, Andrei Seluanov, Vera Gorbunova, P Eline Slagboom, Stephen L Helfand, Nicola Neretti, and John M Sedivy. L1 drives IFN in senescent cells and promotes age-associated inflammation. *Nature*, 566(7742):73–78, 2019.

[69] Linda Partridge, Joris Deelen, and P Eline Slagboom. Facing up to the global challenges of ageing. *Nature*, 561(7721):45–56, 2018.

[70] Xiao Dong, Brandon Milholland, and Jan Vijg. Evidence for a limit to human lifespan. *Nature*, 538:257–259, 2016.

[71] Benjamin Gompertz. On the Nature of the Function Expressive of the Law of Human Mortality, and on a New Mode of Determining the Value of Life Contingencies. *Philosophical Transactions of the Royal Society of London*, 115:513–583, 1825.

[72] Elisabetta Barbi, Francesco Lagona, Marco Marsili, James W Vaupel, and Kenneth W Wachter. The plateau of human mortality: Demography of longevity pioneers. *Science*, 360(6396):1459–1461, 2018.

[73] Vasilis Kontis, James E. Bennett, Colin D. Mathers, Guangquan Li, Kyle Foreman, and Majid Ezzati. Future life expectancy in 35 industrialised countries: projections with a Bayesian model ensemble. *The Lancet*, 389(10076):1323–1335, 2017.

[74] Michael Fine. Intergenerational perspectives on ageing, economics and globalisation. *Australasian Journal on Ageing*, 33(4):220–225, 2014.

[75] Breanne L Newell Stamper, James R Cypser, Katerina Kechris, David Alan Kitzenberg, Patricia M Tedesco, and Thomas E Johnson. Movement decline across lifespan of Caenorhabditis elegans mutants in the insulin/insulin-like signaling pathway. *Aging Cell*, 17(1):e12704, 2018.

[76] Lori Feldman, Stacy L Andersen, Thomas T Perls, Daniel A Dworkis, and Paola Sebastiani. Health Span Approximates Life Span Among Many Supercentenarians: Compression of Morbidity at the Approximate Limit of Life Span. *The Journals of Gerontology: Series A*, 67A(4):395–405, 2012.

[77] Orli G Bahcall. UK Biobank — a new era in genomic medicine. *Nature Reviews Genetics*, 19(12):737, 2018.

[78] Ridho Rahmadi, Perry Groot, Marieke H C van Rijn, Jan A J G van den Brand, Marianne Heins, Hans Knoop, and Tom Heskes. Causality on longitudinal data: Stable specification search in constrained structural equation modeling. *Statistical Methods in Medical Research*, 27(12):3814–3834, 2017.

[79] Garret FitzGerald, David Botstein, Robert Califf, Rory Collins, Keith Peters, Nick Van Bruggen, and Dan Rader. The future of humans as model organisms. *Science*, 361(6402):552–553, 2018.

[80] Steven N Austad and Kathleen E Fischer. Sex Differences in Lifespan. *Cell Metabolism*, 23(6):1022–1033, 2016.

[81] J Graham Ruby, Kevin M Wright, Kristin A Rand, Amir Kermany, Keith Noto, Don Curtis, Neal Varner, Daniel Garrigan, Dmitri Slinkov, Ilya Dorfman, Julie M Granka, Jake Byrnes, Natalie Myres, and Catherine Ball. Estimates of the Heritability of Human Longevity Are Substantially Inflated due to Assortative Mating. *Genetics*, 210(3):1109–1124, 2018.

[82] Joanna Kaplanis, Assaf Gordon, Tal Shor, Omer Weissbrod, Dan Geiger, Mary Wahl, Michael Gershovits, Barak Markus, Mona Sheikh, Melissa Gymrek, Gaurav Bhatia, Daniel G MacArthur, Alkes L Price, and Yaniv Erlich. Quantitative analysis of population-scale family trees with millions of relatives. *Science*, 360(6385):171–175, 2018.

[83] Leanne M Redman, Steven R Smith, Jeffrey H Burton, Corby K Martin, Dora Il'yasova, and Eric Ravussin. Metabolic Slowing and Reduced Oxidative Damage with Sustained Caloric Restriction Support the Rate of Living and Oxidative Damage Theories of Aging. *Cell Metabolism*, 27(4):805–815.e4, 2018.

[84] Min Wei, Sebastian Brandhorst, Mahshid Shelehchi, Hamed Mirzaei, Chia Wei Cheng, Julia Budniak, Susan Groshen, Wendy J Mack, Esra Guen, Stefano Di Biase, Pinchas Cohen, Todd E Morgan, Tanya Dorff, Kurt Hong, Andreas Michalsen, Alessandro Laviano, and Valter D Longo. Fasting-mimicking diet and markers/risk factors for aging, diabetes, cancer, and cardiovascular disease. *Science Translational Medicine*, 9(377):eaai8700, 2017.

[85] Erik A. Richter and Neil B. Ruderman. AMPK and the biochemistry of exercise: implications for human health and disease. *Biochemical Journal*, 418(2):261–275, 2009.

[86] Jasper Most, Valeria Tosti, Leanne M Redman, and Luigi Fontana. Calorie restriction in humans: An update. *Ageing Research Reviews*, 39:36–45, 2017.

[87] Yang Claire Yang, Courtney Boen, Karen Gerken, Ting Li, Kristen Schorpp, and Kathleen Mullan Harris. Social relationships and physiological determinants of longevity across the human life span. *Proceedings of the National Academy of Sciences*, 113(3):578–583, 2016.

[88] Michel Poulain, Anne Herm, and Gianni Pes. The Blue Zones: areas of exceptional longevity around the world. *Vienna Yearbook of Population Research*, 11:87–108, 2013.

[89] Conrad H. Waddington. The Epigenotype. *Endeavor*, 1:18–20, 1942.

[90] Scott F Gilbert. Commentary: 'The Epigenotype' by C.H. Waddington. *International Journal of Epidemiology*, 41(1):20–23, 2011.

[91] Conrad H. Waddington. The cybernetics of development. In *The strategy of the genes*, pages 27–38. 1957.

[92] Tuuli Lappalainen and John M Greally. Associating cellular epigenetic models with human phenotypes. *Nature Reviews Genetics*, 18:441–451, 2017.

[93] The ENCODE Project Consortium, Ian Dunham, Anshul Kundaje, Shelley F Aldred, Patrick J Collins, Carrie A Davis, Francis Doyle, Charles B Epstein, Seth Frietze, Jennifer Harrow, Rajinder Kaul, Jainab Khatun, Bryan R Lajoie, Stephen G Landt, Bum-Kyu Lee, Florencia Pauli, Kate R Rosenbloom, Peter Sabo, Alexias Safi, Amartya Sanyal, Noam Shoresh, Jeremy M Simon, Lingyun Song, Nathan D Trinklein, Robert C Altshuler, Ewan Birney, James B Brown, Chao Cheng, Sarah Djebali, Xianjun Dong, Ian Dunham, Jason Ernst, Terrence S Furey, Mark Gerstein, Belinda Giardine, Melissa Greven, Ross C Hardison, Robert S Harris, Javier Herrero, Michael M Hoffman, Sowmya Iyer, Manolis Kellis, Jainab Khatun, Pouya Kheradpour, Anshul Kundaje, Timo Lassmann, Qunhua Li, Xinying Lin, Georgi K Marinov, Angelika Merkel, Ali Mortazavi, Stephen C J Parker, Timothy E Reddy, Joel Rozowsky, Felix Schlesinger, Robert E Thurman, Jie Wang, Lucas D Ward, Troy W Whitfield, Steven P Wilder, Weisheng Wu, Hualin S Xi, Kevin Y Yip, Jiali Zhuang, Bradley E Bernstein, Ewan Birney, Ian Dunham, Eric D Green, Chris Gunter, Michael Snyder, Michael J Pazin, Rebecca F Lowdon, Laura A L Dillon, Leslie B Adams, Caroline J Kelly, Julia Zhang, Judith R Wexler, Eric D Green, Peter J Good, Elise A Feingold, Bradley E Bernstein, Ewan Birney, Gregory E Crawford, Job Dekker, Laura Elnitski, Peggy J Farnham, Mark Gerstein, Morgan C Giddings, Thomas R Gingeras, Eric D Green, Roderic Guigó, Ross C Hardison, Timothy J Hubbard, Manolis Kellis, W James Kent, Jason D Lieb, Elliott H Margulies, Richard M Myers, Michael Snyder, John A Stamatoyannopoulos, Scott A Tenenbaum, Zhiping Weng, Kevin P White, Barbara Wold, Jainab Khatun, Yanbao Yu, John Wrobel, Brian A Risk, Harsha P Gunawardena, Heather C Kuiper, Christopher W Maier, Ling Xie, Xian Chen, Morgan C Giddings, Bradley E Bernstein, Charles B Epstein, Noam Shoresh, Jason Ernst, Pouya Kheradpour, Tarjei S Mikkelsen, Shawn Gillespie, Alon Goren, Oren Ram, Xiaolan Zhang, Li Wang, Robbyn Issner, Michael J Coyne, Timothy Durham, Manching Ku, Thanh Truong, Lucas D Ward, Robert C Altshuler, Matthew L Eaton, Manolis Kellis, Sarah Djebali, Carrie A Davis, Angelika Merkel, Alex Dobin, Timo Lassmann, Ali Mortazavi, Andrea Tanzer, Julien Lagarde, Wei Lin, Felix Schlesinger, Chenghai Xue, Georgi K Marinov, Jainab Khatun, Brian A Williams, Chris Zaleski, Joel Rozowsky, Maik Röder, Felix Kokocinski, Rehab F Abdelhamid, Tyler Alioto, Igor Antoshechkin, Michael T Baer, Philippe Batut, Ian Bell, Kimberly Bell, Sudipto Chakrabortty, Xian Chen, Jacqueline Chrast, Joao Curado, Thomas Derrien, Jorg Drenkow, Erica Dumais, Jackie Dumais, Radha Duttagupta, Megan Fastuca, Kata Fejes-Toth, Pedro Ferreira, Sylvain Foissac, Melissa J Fullwood, Hui Gao, David Gonzalez, Assaf Gordon, Harsha P Gunawardena, Cédric Howald, Sonali Jha, Rory Johnson, Philipp Kapranov, Brandon King, Colin Kingswood, Guoliang Li, Oscar J Luo, Eddie Park, Jonathan B Preall, Kimberly Presaud, Paolo Ribeca, Brian A Risk, Daniel Robyr, Xiaoan Ruan, Michael Sammeth, Kuljeet Singh Sandhu, Lorain Schaeffer, Lei-Hoon See, Atif Shahab, Jorgen Skancke, Ana Maria Suzuki, Hazuki Takahashi, Hagen Tilgner, Diane Trout, Nathalie Walters, Huaien Wang, John Wrobel, Yanbao Yu, Yoshihide Hayashizaki, Jennifer Harrow, Mark Gerstein, Timothy J Hubbard, Alexandre Reymond, Stylianos E Antonarakis, Gregory J Hannon, Morgan C Giddings, Yijun Ruan, Barbara Wold, Piero Carninci, Roderic Guigó, Thomas R Gingeras, Kate R Rosenbloom, Cricket A Sloan, Katrina Learned, Venkat S Malladi, Matthew C Wong, Galt P Barber, Melissa S Cline, Timothy R Dreszer, Steven G Heitner, Donna Karolchik, W James Kent, Vanessa M Kirkup, Laurence R Meyer, Jeffrey C Long, Morgan Maddren, Brian J Raney, Terrence S Furey, Lingyun Song, Linda L Grasfeder, Paul G

Giresi, Bum-Kyu Lee, Anna Battenhouse, Nathan C Sheffield, Jeremy M Simon, Kimberly A Showers, Alexias Safi, Darin London, Akshay A Bhinge, Christopher Shestak, Matthew R Schaner, Seul Ki Kim, Zhuzhu Z Zhang, Piotr A Mieczkowski, Joanna O Mieczkowska, Zheng Liu, Ryan M McDaniell, Yunyun Ni, Naim U Rashid, Min Jae Kim, Sheera Adar, Zhancheng Zhang, Tianyuan Wang, Deborah Winter, Damian Keefe, Ewan Birney, Vishwanath R Iyer, Jason D Lieb, Gregory E Crawford, Guoliang Li, Kuljeet Singh Sandhu, Meizhen Zheng, Ping Wang, Oscar J Luo, Atif Shahab, Melissa J Fullwood, Xiaoan Ruan, Yijun Ruan, Richard M Myers, Florencia Pauli, Brian A Williams, Jason Gertz, Georgi K Marinov, Timothy E Reddy, Jost Vielmetter, E Partridge, Diane Trout, Katherine E Varley, Clarke Gasper, Anita Bansal, Shirley Pepke, Preti Jain, Henry Amrhein, Kevin M Bowling, Michael Anaya, Marie K Cross, Brandon King, Michael A Muratet, Igor Antoshechkin, Kimberly M Newberry, Kenneth McCue, Amy S Nesmith, Katherine I Fisher-Aylor, Barbara Pusey, Gilberto DeSalvo, Stephanie L Parker, Sreeram Balasubramanian, Nicholas S Davis, Sarah K Meadows, Tracy Eggleston, Chris Gunter, J Scott Newberry, Shawn E Levy, Devin M Absher, Ali Mortazavi, Wing H Wong, Barbara Wold, Matthew J Blow, Axel Visel, Len A Pennachio, Laura Elnitski, Elliott H Margulies, Stephen C J Parker, Hanna M Petrykowska, Alexej Abyzov, Bronwen Aken, Daniel Barrell, Gemma Barson, Andrew Berry, Alexandra Bignell, Veronika Boychenko, Giovanni Bussotti, Jacqueline Chrast, Claire Davidson, Thomas Derrien, Gloria Despacio-Reyes, Mark Diekhans, Iakes Ezkurdia, Adam Frankish, James Gilbert, Jose Manuel Gonzalez, Ed Griffiths, Rachel Harte, David A Hendrix, Cédric Howald, Toby Hunt, Irwin Jungreis, Mike Kay, Ekta Khurana, Felix Kokocinski, Jing Leng, Michael F Lin, Jane Loveland, Zhi Lu, Deepa Manthravadi, Marco Mariotti, Jonathan Mudge, Gaurab Mukherjee, Cedric Notredame, Baikang Pei, Jose Manuel Rodriguez, Gary Saunders, Andrea Sboner, Stephen Searle, Cristina Sisu, Catherine Snow, Charlie Steward, Andrea Tanzer, Electra Tapanari, Michael L Tress, Marijke J van Baren, Nathalie Walters, Stefan Washietl, Laurens Wilming, Amonida Zadissa, Zhengdong Zhang, Michael Brent, David Haussler, Manolis Kellis, Alfonso Valencia, Mark Gerstein, Alexandre Reymond, Roderic Guigó, Jennifer Harrow, Timothy J Hubbard, Stephen G Landt, Seth Frietze, Alexej Abyzov, Nick Addleman, Roger P Alexander, Raymond K Auerbach, Suganthi Balasubramanian, Keith Bettinger, Nitin Bhardwaj, Alan P Boyle, Alina R Cao, Philip Cayting, Alexandra Charos, Yong Cheng, Chao Cheng, Catharine Eastman, Ghia Euskirchen, Joseph D Fleming, Fabian Grubert, Lukas Habegger, Manoj Hariharan, Arif Harmanci, Sushma Iyengar, Victor X Jin, Konrad J Karczewski, Maya Kasowski, Phil Lacroute, Hugo Lam, Nathan Lamarre-Vincent, Jing Leng, Jin Lian, Marianne Lindahl-Allen, Renqiang Min, Benoit Miotto, Hannah Monahan, Zarmik Moqtaderi, Xinmeng J Mu, Henriette O'Geen, Zhengqing Ouyang, Dorrelyn Patacsil, Baikang Pei, Debasish Raha, Lucia Ramirez, Brian Reed, Joel Rozowsky, Andrea Sboner, Minyi Shi, Cristina Sisu, Teri Slifer, Heather Witt, Linfeng Wu, Xiaoqin Xu, Koon-Kiu Yan, Xinqiong Yang, Kevin Y Yip, Zhengdong Zhang, Kevin Struhl, Sherman M Weissman, Mark Gerstein, Peggy J Farnham, Michael Snyder, Scott A Tenenbaum, Luiz O Penalva, Francis Doyle, Subhradip Karmakar, Stephen G Landt, Raj R Bhanvadia, Alina Choudhury, Marc Domanus, Lijia Ma, Jennifer Moran, Dorrelyn Patacsil, Teri Slifer, Alec Victorsen, Xinqiong Yang, Michael Snyder, Kevin P White, Thomas Auer, Lazaro Centanin, Michael Eichenlaub, Franziska Gruhl, Stephan Heermann, Burkhard Hoeckendorf, Daigo Inoue, Tanja Kellner, Stephan Kirchmaier, Claudia Mueller, Robert Reinhardt, Lea Schertel, Stephanie Schneider, Rebecca Sinn, Beate

Wittbrodt, Jochen Wittbrodt, Zhiping Weng, Troy W Whitfield, Jie Wang, Patrick J Collins, Shelley F Aldred, Nathan D Trinklein, E Christopher Partridge, Richard M Myers, Job Dekker, Gaurav Jain, Bryan R Lajoie, Amartya Sanyal, Gayathri Balasundaram, Daniel L Bates, Rachel Byron, Theresa K Canfield, Morgan J Diegel, Douglas Dunn, Abigail K Ebersol, Tristan Frum, Kavita Garg, Erica Gist, R Scott Hansen, Lisa Boatman, Eric Haugen, Richard Humbert, Gaurav Jain, Audra K Johnson, Ericka M Johnson, Tattyana V Kutyavin, Bryan R Lajoie, Kristen Lee, Dimitra Lotakis, Matthew T Maurano, Shane J Neph, Fiedencio V Neri, Eric D Nguyen, Hongzhu Qu, Alex P Reynolds, Vaughn Roach, Eric Rynes, Peter Sabo, Minerva E Sanchez, Richard S Sandstrom, Amartya Sanyal, Anthony O Shafer, Andrew B Stergachis, Sean Thomas, Robert E Thurman, Benjamin Vernot, Jeff Vierstra, Shinny Vong, Hao Wang, Molly A Weaver, Yongqi Yan, Miaohua Zhang, Joshua M Akey, Michael Bender, Michael O Dorschner, Mark Groudine, Michael J MacCoss, Patrick Navas, George Stamatoyannopoulos, Rajinder Kaul, Job Dekker, John A Stamatoyannopoulos, Ian Dunham, Kathryn Beal, Alvis Brazma, Paul Flicek, Javier Herrero, Nathan Johnson, Damian Keefe, Margus Lukk, Nicholas M Luscombe, Daniel Sobral, Juan M Vaquerizas, Steven P Wilder, Serafim Batzoglou, Arend Sidow, Nadine Hussami, Sofia Kyriazopoulou-Panagiotopoulou, Max W Libbrecht, Marc A Schaub, Anshul Kundaje, Ross C Hardison, Webb Miller, Belinda Giardine, Robert S Harris, Weisheng Wu, Peter J Bickel, Balazs Banfai, Nathan P Boley, James B Brown, Haiyan Huang, Qunhua Li, Jingyi Jessica Li, William Stafford Noble, Jeffrey A Bilmes, Orion J Buske, Michael M Hoffman, Avinash D Sahu, Peter V Kharchenko, Peter J Park, Dannon Baker, James Taylor, Zhiping Weng, Sowmya Iyer, Xianjun Dong, Melissa Greven, Xinying Lin, Jie Wang, Hualin S Xi, Jiali Zhuang, Mark Gerstein, Roger P Alexander, Suganthi Balasubramanian, Chao Cheng, Arif Harmanci, Lucas Lochovsky, Renqiang Min, Xinmeng J Mu, Joel Rozowsky, Koon-Kiu Yan, Kevin Y Yip, and Ewan Birney. An integrated encyclopedia of DNA elements in the human genome. *Nature*, 489:57–74, 2012.

[94] Roadmap Epigenomics Consortium, Anshul Kundaje, Wouter Meuleman, Jason Ernst, Misha Bilenky, Angela Yen, Alireza Heravi-Moussavi, Pouya Kheradpour, Zhizhuo Zhang, Jianrong Wang, Michael J Ziller, Viren Amin, John W Whitaker, Matthew D Schultz, Lucas D Ward, Abhishek Sarkar, Gerald Quon, Richard S Sandstrom, Matthew L Eaton, Yi-Chieh Wu, Andreas Pfenning, Xinchen Wang, Melina ClaussnitzerYaping Liu, Cristian Coarfa, R Alan Harris, Noam Shoresh, Charles B Epstein, Elizabeta Gjoneska, Danny Leung, Wei Xie, R David Hawkins, Ryan Lister, Chibo Hong, Philippe Gascard, Andrew J Mungall, Richard Moore, Eric Chuah, Angela Tam, Theresa K Canfield, R Scott Hansen, Rajinder Kaul, Peter J Sabo, Mukul S Bansal, Annaick Carles, Jesse R Dixon, Kai-How Farh, Soheil Feizi, Rosa Karlic, Ah-Ram Kim, Ashwinikumar Kulkarni, Daofeng Li, Rebecca Lowdon, GiNell Elliott, Tim R Mercer, Shane J Neph, Vitor Onuchic, Paz Polak, Nisha Rajagopal, Pradipta Ray, Richard C Sallari, Kyle T Siebenthall, Nicholas A Sinnott-Armstrong, Michael Stevens, Robert E Thurman, Jie Wu, Bo Zhang, Xin Zhou, Nezar Abdennur, Mazhar Adli, Martin Akerman, Luis Barrera, Jessica Antosiewicz-Bourget, Tracy Ballinger, Michael J Barnes, Daniel Bates, Robert J A Bell, David A Bennett, Katherine Bianco, Christoph Bock, Patrick Boyle, Jan Brinchmann, Pedro Caballero-Campo, Raymond Camahort, Marlene J Carrasco-Alfonso, Timothy Charnecki, Huaming Chen, Zhao Chen, Jeffrey B Cheng, Stephanie Cho, Andy Chu, Wen-Yu Chung, Chad Cowan, Qixia Athena Deng, Vikram Deshpande, Morgan Diegel, Bo Ding, Timothy Durham,

Lorigail Echipare, Lee Edsall, David Flowers, Olga Genbacev-Krtolica, Casey Gifford, Shawn Gillespie, Erika Giste, Ian A Glass, Andreas Gnirke, Matthew Gormley, Hongcang Gu, Junchen Gu, David A Hafler, Matthew J Hangauer, Manoj Hariharan, Meital Hatan, Eric Haugen, Yupeng He, Shelly Heimfeld, Sarah Herlofsen, Zhonggang Hou, Richard Humbert, Robbyn Issner, Andrew R Jackson, Haiyang Jia, Peng Jiang, Audra K Johnson, Theresa Kadlecek, Baljit Kamoh, Mirhan Kapidzic, Jim Kent, Audrey Kim, Markus Kleinewietfeld, Sarit Klugman, Jayanth Krishnan, Samantha Kuan, Tanya Kutyavin, Ah-Young Lee, Kristen Lee, Jian Li, Nan Li, Yan Li, Keith L Ligon, Shin Lin, Yiing Lin, Jie Liu, Yuxuan Liu, C John Luckey, Yussanne P Ma, Cecile Maire, Alexander Marson, John S Mattick, Michael Mayo, Michael McMaster, Hayden Metsky, Tarjei Mikkelsen, Diane Miller, Mohammad Miri, Eran Mukame, Raman P Nagarajan, Fidencio Neri, Joseph Nery, Tung Nguyen, Henriette O'Geen, Sameer Paithankar, Thalia Papayannopoulou, Mattia Pelizzola, Patrick Plettner, Nicholas E Propson, Sriram Raghuraman, Brian J Raney, Anthony Raubitschek, Alex P Reynolds, Hunter Richards, Kevin Riehle, Paolo Rinaudo, Joshua F Robinson, Nicole B Rockweiler, Evan Rosen, Eric Rynes, Jacqueline Schein, Renee Sears, Terrence Sejnowski, Anthony Shafer, Li Shen, Robert Shoemaker, Mahvash Sigaroudinia, Igor Slukvin, Sandra Stehling-Sun, Ron Stewart, Sai Lakshmi Subramanian, Kran Suknuntha, Scott Swanson, Shulan Tian, Hannah Tilden, Linus Tsai, Mark Urich, Ian Vaughn, Jeff Vierstra, Shinny Vong, Ulrich Wagner, Hao Wang, Tao Wang, Yunfei Wang, Arthur Weiss, Holly Whitton, Andre Wildberg, Heather Witt, Kyoung-Jae Won, Mingchao Xie, Xiaoyun Xing, Iris Xu, Zhenyu Xuan, Zhen Ye, Chia-an Yen, Pengzhi Yu, Xian Zhang, Xiaolan Zhang, Jianxin Zhao, Yan Zhou, Jiang Zhu, Yun Zhu, Steven Ziegler, Arthur E Beaudet, Laurie A Boyer, Philip L De Jager, Peggy J Farnham, Susan J Fisher, David Haussler, Steven J M Jones, Wei Li, Marco A Marra, Michael T McManus, Shamil Sunyaev, James A Thomson, Thea D Tlsty, Li-Huei Tsai, Wei Wang, Robert A Waterland, Michael Q Zhang, Lisa H Chadwick, Bradley E Bernstein, Joseph F Costello, Joseph R Ecker, Martin Hirst, Alexander Meissner, Aleksandar Milosavljevic, Bing Ren, John A Stamatoyannopoulos, Ting Wang, Manolis Kellis, Anshul Kundaje, Wouter Meuleman, Jason Ernst, Misha Bilenky, Angela Yen, Alireza Heravi-Moussavi, Pouya Kheradpour, Zhizhuo Zhang, Jianrong Wang, Michael J Ziller, Viren Amin, John W Whitaker, Matthew D Schultz, Lucas D Ward, Abhishek Sarkar, Gerald Quon, Richard S Sandstrom, Matthew L Eaton, Yi-Chieh Wu, Andreas R Pfenning, Xinchen Wang, Melina Claussnitzer, Yaping Liu, Cristian Coarfa, R Alan Harris, Noam Shoresh, Charles B Epstein, Elizabeta Gjoneska, Danny Leung, Wei Xie, R David Hawkins, Ryan Lister, Chibo Hong, Philippe Gascard, Andrew J Mungall, Richard Moore, Eric Chuah, Angela Tam, Theresa K Canfield, R Scott Hansen, Rajinder Kaul, Peter J Sabo, Mukul S Bansal, Annaick Carles, Jesse R Dixon, Kai-How Farh, Soheil Feizi, Rosa Karlic, Ah-Ram Kim, Ashwinikumar Kulkarni, Daofeng Li, Rebecca Lowdon, GiNell Elliott, Tim R Mercer, Shane J Neph, Vitor Onuchic, Paz Polak, Nisha Rajagopal, Pradipta Ray, Richard C Sallari, Kyle T Siebenthall, Nicholas A Sinnott-Armstrong, Michael Stevens, Robert E Thurman, Jie Wu, Bo Zhang, Xin Zhou, Arthur E Beaudet, Laurie A Boyer, Philip L De Jager, Peggy J Farnham, Susan J Fisher, David Haussler, Steven J M Jones, Wei Li, Marco A Marra, Michael T McManus, Shamil Sunyaev, James A Thomson, Thea D Tlsty, Li-Huei Tsai, Wei Wang, Robert A Waterland, Michael Q Zhang, Lisa H Chadwick, Bradley E Bernstein, Joseph F Costello, Joseph R Ecker, Martin Hirst, Alexander Meissner, Aleksandar Milosavljevic, Bing Ren, John A Stamatoyannopoulos, Ting Wang, and

Manolis Kellis. Integrative analysis of 111 reference human epigenomes. *Nature*, 518:317–330, 2015.

[95] John M Greally. A user's guide to the ambiguous word 'epigenetics'. *Nature Reviews Molecular Cell Biology*, 19:207–208, 2018.

[96] Adrian Bird. Perceptions of epigenetics. *Nature*, 447:396–398, 2007.

[97] C.-t. Wu and J R Morris. Genes, Genetics, and Epigenetics: A Correspondence. *Science*, 293(5532):1103–1105, 2001.

[98] Horng D Ou, Sébastien Phan, Thomas J Deerinck, Andrea Thor, Mark H Ellisman, and Clodagh C O'Shea. ChromEMT: Visualizing 3D chromatin structure and compaction in interphase and mitotic cells. *Science*, 357(6349):eaag0025, 2017.

[99] Moyra Lawrence, Sylvain Daujat, and Robert Schneider. Lateral Thinking: How Histone Modifications Regulate Gene Expression. *Trends in Genetics*, 32(1):42–56, 2016.

[100] Brian D Strahl and C David Allis. The language of covalent histone modifications. *Nature*, 403(6765):41–45, 2000.

[101] John S Mattick, Paulo P Amaral, Marcel E Dinger, Tim R Mercer, and Mark F Mehler. RNA regulation of epigenetic processes. *BioEssays*, 31(1):51–59, 2009.

[102] Kevin V Morris and John S Mattick. The rise of regulatory RNA. *Nature Reviews Genetics*, 15:423–437, 2014.

[103] Randal Halfmann and Susan Lindquist. Epigenetics in the Extreme: Prions and the Inheritance of Environmentally Acquired Traits. *Science*, 330(6004):629–632, 2010.

[104] C David Allis and Thomas Jenuwein. The molecular hallmarks of epigenetic control. *Nature Reviews Genetics*, 17:487–500, 2016.

[105] Danny Reinberg and Lynne D Vales. Chromatin domains rich in inheritance. *Science*, 361(6397):33–34, 2018.

[106] Patrick Trojer and Danny Reinberg. Facultative Heterochromatin: Is There a Distinctive Molecular Signature? *Molecular Cell*, 28(1):1–13, 2007.

[107] Jason Ernst and Manolis Kellis. Discovery and characterization of chromatin states for systematic annotation of the human genome. *Nature Biotechnology*, 28:817–825, 2010.

[108] Jo Peters. The role of genomic imprinting in biology and disease: an expanding view. *Nature Reviews Genetics*, 15:517–530, 2014.

[109] Anton Wutz. Gene silencing in X-chromosome inactivation: advances in understanding facultative heterochromatin formation. *Nature Reviews Genetics*, 12:542–553, 2011.

[110] Solenn Patalano, Timothy A Hore, Wolf Reik, and Seirian Sumner. Shifting behaviour: epigenetic reprogramming in eusocial insects. *Current Opinion in Cell Biology*, 24(3):367–373, 2012.

[111] Silvia C. Remolina and Kimberly A. Hughes. Evolution and mechanisms of long life and high fertility in queen honey bees. *Age*, 30(2-3):177–185, 2008.

[112] Aaron Taudt, Maria Colomé-Tatché, and Frank Johannes. Genetic sources of population epigenomic variation. *Nature Reviews Genetics*, 17:319–332, 2016.

[113] John W Whitaker, Zhao Chen, and Wei Wang. Predicting the human epigenome from DNA motifs. *Nature Methods*, 12:265–272, 2014.

[114] Juan E Castillo-Fernandez, Tim D Spector, and Jordana T Bell. Epigenetics of discordant monozygotic twins: implications for disease. *Genome Medicine*, 6(7):60, 2014.

[115] María A Sánchez-Romero, Ignacio Cota, and Josep Casadesús. DNA methylation in bacteria: from the methyl group to the methylome. *Current Opinion in Microbiology*, 25:9–16, 2015.

[116] Xiaoji Wu and Yi Zhang. TET-mediated active DNA demethylation: mechanism, function and beyond. *Nature Reviews Genetics*, 18:517–534, 2017.

[117] En Li and Yi Zhang. DNA methylation in mammals. *Cold Spring Harbor Perspectives in Biology*, 6(5):a019133, 2014.

[118] Francesco Neri, Stefania Rapelli, Anna Krepelova, Danny Incarnato, Caterina Parlato, Giulia Basile, Mara Maldotti, Francesca Anselmi, and Salvatore Oliviero. Intragenic DNA methylation prevents spurious transcription initiation. *Nature*, 543(7643):72–77, 2017.

[119] Zachary D Smith and Alexander Meissner. DNA methylation: roles in mammalian development. *Nature Reviews Genetics*, 14:204–220, 2013.

[120] Danuta M Jeziorska, Robert J S Murray, Marco De Gobbi, Ricarda Gaentzsch, David Garrick, Helena Ayyub, Taiping Chen, En Li, Jelena Telenius, Magnus Lynch, Bryony Graham, Andrew J H Smith, Jonathan N Lund, Jim R Hughes, Douglas R Higgs, and Cristina Tufarelli. DNA methylation of intragenic CpG islands depends on their transcriptional activity during differentiation and disease. *Proceedings of the National Academy of Sciences*, 114(36):E7526–E7535, 2017.

[121] Jia Zeng, Hema K Nagrajan, and Soojin V Yi. Fundamental diversity of human CpG islands at multiple biological levels. *Epigenetics*, 9(4):483–491, 2014.

[122] Yuta Takahashi, Jun Wu, Keiichiro Suzuki, Paloma Martinez-Redondo, Mo Li, Hsin-Kai Liao, Min-Zu Wu, Reyna Hernández-Benítez, Tomoaki Hishida, Maxim Nikolaievich Shokhirev, Concepcion Rodriguez Esteban, Ignacio Sancho-Martinez, and Juan Carlos Izpisua Belmonte. Integration of CpG-free DNA induces de novo methylation of CpG islands in pluripotent stem cells. *Science*, 356(6337):503–508, 2017.

[123] William A Flavahan, Elizabeth Gaskell, and Bradley E Bernstein. Epigenetic plasticity and the hallmarks of cancer. *Science*, 357(6348):eaal2380, 2017.

[124] R Holliday and J E Pugh. DNA modification mechanisms and gene activity during development. *Science*, 187(4173):226–232, 1975.

[125] A D Riggs. X inactivation, differentiation, and DNA methylation. *Cytogenetic and Genome Research*, 14(1):9–25, 1975.

[126] Matthew D Schultz, Yupeng He, John W Whitaker, Manoj Hariharan, Eran A Mukamel, Danny Leung, Nisha Rajagopal, Joseph R Nery, Mark A Urich, Huaming Chen, Shin Lin, Yiing Lin, Inkyung Jung, Anthony D Schmitt, Siddarth Selvaraj, Bing Ren, Terrence J Sejnowski, Wei Wang, and Joseph R Ecker. Human body epigenome maps reveal noncanonical DNA methylation variation. *Nature*, 523:212–216, 2015.

[127] Michael J Ziller, Fabian Müller, Jing Liao, Yingying Zhang, Hongcang Gu, Christoph Bock, Patrick Boyle, Charles B Epstein, Bradley E Bernstein, Thomas Lengauer, Andreas Gnirke, and Alexander Meissner. Genomic Distribution and Inter-Sample Variation of Non-CpG Methylation across Human Cell Types. *PLOS Genetics*, 7(12):e1002389, 2011.

[128] Yupeng He and Joseph R Ecker. Non-CG Methylation in the Human Genome. *Annual Review of Genomics and Human Genetics*, 16(1):55–77, 2015.

[129] Déborah Bourc'his, Guo-Liang Xu, Chyuan-Sheng Lin, Brooke Bollman, and Timothy H Bestor. Dnmt3L and the Establishment of Maternal Genomic Imprints. *Science*, 294(5551):2536–2539, 2001.

[130] Mary W Tomida, Sally Gaddis, Yoko Takata, Bigang Liu, Kevin Lin, Marcos R Estecio, Swanand Hardikar, Yue Lu, Nicolas Veland, Yang Zeng, Taiping Chen, Jianjun Shen, Debapriya Saha, Humaira Gowher, and Hongbo Zhao. DNMT3L facilitates DNA methylation partly by maintaining DNMT3A stability in mouse embryonic stem cells. *Nucleic Acids Research*, 47(1):152–167, 2018.

[131] Mario Iurlaro, Ferdinand von Meyenn, and Wolf Reik. DNA methylation homeostasis in human and mouse development. *Current Opinion in Genetics & Development*, 43:101–109, 2017.

[132] Skirmantas Kriaucionis and Nathaniel Heintz. The Nuclear DNA Base 5-Hydroxymethylcytosine Is Present in Purkinje Neurons and the Brain. *Science*, 324(5929):929–930, 2009.

[133] M Tahiliani, K P Koh, Y Shen, W A Pastor, H Bandukwala, and Y Brudno. Conversion of 5-methylcytosine to 5-hydroxymethylcytosine in mammalian DNA by MLL partner TET1. *Science*, 324(5929):930–935, 2009.

[134] N W Penn, R Suwalski, C O'Riley, K Bojanowski, and R Yura. The presence of 5-hydroxymethylcytosine in animal deoxyribonucleic acid. *Biochemical Journal*, 126(4):781–790, 1972.

[135] Shinsuke Ito, Ana C D'Alessio, Olena V Taranova, Kwonho Hong, Lawrence C Sowers, and Yi Zhang. Role of Tet proteins in 5mC to 5hmC conversion, ES-cell self-renewal and inner cell mass specification. *Nature*, 466:1129–1133, 2010.

[136] Tao P. Wu, Tao Wang, Matthew G. Seetin, Yongquan Lai, Shijia Zhu, Kaixuan Lin, Yifei Liu, Stephanie D. Byrum, Samuel G. Mackintosh, Mei Zhong, Alan Tackett, Guilin Wang, Lawrence S. Hon, Gang Fang, James a. Swenberg, and Andrew Z. Xiao. DNA methylation on N6-adenine in mammalian embryonic stem cells. *Nature*, 532:1–18, 2016.

[137] Chuan-Le Xiao, Song Zhu, Minghui He, De Chen, Qian Zhang, Ying Chen, Guoliang Yu, Jinbao Liu, Shang-Qian Xie, Feng Luo, Zhe Liang, De-Peng Wang, Xiao-Chen Bo, Xiao-Feng Gu, Kai Wang, and Guang-Rong Yan. N6-Methyladenine DNA Modification in the Human Genome. *Molecular Cell*, 71(2):306–318.e7, 2018.

[138] Walfred W C Tang, Toshihiro Kobayashi, Naoko Irie, Sabine Dietmann, and M Azim Surani. Specification and epigenetic programming of the human germ line. *Nature Reviews Genetics*, 17:585–600, 2016.

[139] Yaser Atlasi and Hendrik G Stunnenberg. The interplay of epigenetic marks during stem cell differentiation and development. *Nature Reviews Genetics*, 18:643–658, 2017.

[140] M Frommer, L E McDonald, D S Millar, C M Collis, F Watt, G W Grigg, P L Molloy, and C L Paul. A genomic sequencing protocol that yields a positive display of 5-methylcytosine residues in individual DNA strands. *Proceedings of the National Academy of Sciences*, 89(5):1827–1831, 1992.

[141] 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.

[142] Yibin Liu, Paulina Siejka-Zielińska, Gergana Velikova, Ying Bi, Fang Yuan, Marketa Tomkova, Chunsen Bai, Lei Chen, Benjamin Schuster-Böckler, and Chun-Xiao Song. Bisulfite-free direct detection of 5-methylcytosine and 5-hydroxymethylcytosine at base resolution. *Nature Biotechnology*, 37(4):424–429, 2019.

[143] Elizabeth J Radford, Mitsuteru Ito, Hui Shi, Jennifer A Corish, Kazuki Yamazawa, Elvira Isganaitis, Stefanie Seisenberger, Timothy A Hore, Wolf Reik, Serap Erkek, Antoine H F M Peters, Mary-Elizabeth Patti, and Anne C Ferguson-Smith. In utero undernourishment perturbs the adult sperm methylome and intergenerational metabolism. *Science*, 345(6198):1255903, 2014.

[144] Joehanes Roby, Just Allan C., Marioni Riccardo E., Pilling Luke C., Reynolds Lindsay M., Mandaviya Pooja R., Guan Weihua, Xu Tao, Elks Cathy E., Aslibekyan Stella, Moreno-Macias Hortensia, Smith Jennifer A., Brody Jennifer A., Dhingra Radhika, Yousefi Paul, Pankow James S., Kunze Sonja, Shah Sonia H., McRae Allan F., Lohman Kurt, Sha Jin, Absher Devin M., Ferrucci Luigi, Zhao Wei, Demerath Ellen W., Bressler Jan, Grove Megan L., Huan Tianxiao, Liu Chunyu, Mendelson Michael M., Yao Chen, Kiel Douglas P., Peters Annette, Wang-Sattler Rui, Visscher Peter

M., Wray Naomi R., Starr John M., Ding Jingzhong, Rodriguez Carlos J., Wareham Nicholas J., Irvin Marguerite R., Zhi Degui, Barrdahl Myrto, Vineis Paolo, Ambatipudi Srikant, Uitterlinden André G., Hofman Albert, Schwartz Joel, Colicino Elena, Hou Lifang, Vokonas Pantel S., Hernandez Dena G., Singleton Andrew B., Bandinelli Stefania, Turner Stephen T., Ware Erin B., Smith Alicia K., Klengel Torsten, Binder Elisabeth B., Psaty Bruce M., Taylor Kent D., Gharib Sina A., Swenson Brenton R., Liang Liming, DeMeo Dawn L., O'Connor George T., Herceg Zdenko, Ressler Kerry J., Conneely Karen N., Sotoodehnia Nona, Kardia Sharon L R., Melzer David, Baccarelli Andrea A., van Meurs Joyce B J., Romieu Isabelle, Arnett Donna K., Ong Ken K., Liu Yongmei, Waldenberger Melanie, Deary Ian J., Fornage Myriam, Levy Daniel, and London Stephanie J. Epigenetic Signatures of Cigarette Smoking. *Circulation: Cardiovascular Genetics*, 9(5):436–447, 2016.

[145] Andrew E Teschendorff, Zhen Yang, Andrew Wong, Christodoulos P Pipinikas, Yinming Jiao, Allison Jones, Shahzia Anjum, Rebecca Hardy, Helga B Salvesen, Christina Thirlwell, Samuel M Janes, Diana Kuh, and Martin Widschwendter. Correlation of Smoking-Associated DNA Methylation Changes in Buccal Cells With DNA Methylation Changes in Epithelial CancerSmoking and DNA Methylation Changes in Buccal Cells and Epithelial CancerSmoking and DNA Methylation Changes in Buccal Cells and Epi. *JAMA Oncology*, 1(4):476–485, 2015.

[146] 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.

[147] Martin Widschwendter, Allison Jones, Iona Evans, Daniel Reisel, Joakim Dillner, Karin Sundström, Ewout W. Steyerberg, Yvonne Vergouwe, Odette Wegwarth, Felix G. Rebitschek, Uwe Siebert, Gaby Sroczynski, Inez D. de Beaufort, Ineke Bolt, David Cibula, Michal Zikan, Line Bjørge, Nicoletta Colombo, Nadia Harbeck, Frank Dudbridge, Anne-Marie Tasse, Bartha M. Knoppers, Yann Joly, Andrew E. Teschendorff, and Nora Pashayan. Epigenome-based cancer risk prediction: rationale, opportunities and challenges. *Nature Reviews Clinical Oncology*, 15:292–309, 2018.

[148] Lauren N. Booth and Anne Brunet. The Aging Epigenome. *Molecular Cell*, 62(5):728–744, 2016.

[149] Bérénice A Benayoun, Elizabeth A Pollina, and Anne Brunet. Epigenetic regulation of ageing: linking environmental inputs to genomic stability. *Nature Reviews Molecular Cell Biology*, 16:593–610, 2015.

[150] Sangita Pal and Jessica K Tyler. Epigenetics and aging. *Science Advances*, 2(7):e1600584, 2016.

[151] Payel Sen, Parisha P. Shah, Raffaella Nativio, and Shelley L. Berger. Epigenetic Mechanisms of Longevity and Aging. *Cell*, 166(4):822–839, 2016.

[152] Bryant Villeponteau. The heterochromatin loss model of aging. *Experimental Gerontology*, 32(4):383–394, 1997.

[153] Amy Tsurumi and Willis Li. Global heterochromatin loss: A unifying theory of aging? *Epigenetics*, 7(7):680–688, 2012.

[154] Weiqi Zhang, Jingyi Li, Keiichiro Suzuki, Jing Qu, Ping Wang, Junzhi Zhou, Xiaomeng Liu, Ruotong Ren, Xiuling Xu, Alejandro Ocampo, Tingting Yuan, Jiping Yang, Ying Li, Liang Shi, Dee Guan, Huize Pan, Shunlei Duan, Zhichao Ding, Mo Li, Fei Yi, Ruijun Bai, Yayu Wang, Chang Chen, Fuquan Yang, Xiaoyu Li, Zimei Wang, Emi Aizawa, April Goebl, Rupa Devi Soligalla, Pradeep Reddy, Concepcion Rodriguez Esteban, Fuchou Tang, Guang-Hui Liu, and Juan Carlos Izpisua Belmonte. A Werner syndrome stem cell model unveils heterochromatin alterations as a driver of human aging. *Science*, 348(6239):1160– 1163, 2015.

[155] Rugang Zhang, Wei Chen, and Peter D Adams. Molecular Dissection of Formation of Senescence-Associated Heterochromatin Foci. *Molecular and Cellular Biology*, 27(6):2343–2358, 2007.

[156] Marco De Cecco, Steven W Criscione, Abigail L Peterson, Nicola Neretti, John M Sedivy, and Jill A Kreiling. Transposable elements become active and mobile in the genomes of aging mammalian somatic tissues. *Aging*, 5(12):867–883, 2013.

[157] Yariv Kanfi, Shoshana Naiman, Gail Amir, Victoria Peshti, Guy Zinman, Liat Nahum, Ziv Bar-Joseph, and Haim Y Cohen. The sirtuin SIRT6 regulates lifespan in male mice. *Nature*, 483:218221, 2012.

[158] Raul Mostoslavsky, Katrin F Chua, David B Lombard, Wendy W Pang, Miriam R Fischer, Lionel Gellon, Pingfang Liu, Gustavo Mostoslavsky, Sonia Franco, Michael M Murphy, Kevin D Mills, Parin Patel, Joyce T Hsu, Andrew L Hong, Ethan Ford, Hwei-Ling Cheng, Caitlin Kennedy, Nomeli Nunez, Roderick Bronson, David Frendewey, Wojtek Auerbach, David Valenzuela, Margaret Karow, Michael O Hottiger, Stephen Hursting, J Carl Barrett, Leonard Guarente, Richard Mulligan, Bruce Demple, George D Yancopoulos, and Frederick W Alt. Genomic Instability and Aging-like Phenotype in the Absence of Mammalian SIRT6. *Cell*, 124(2):315–329, 2006.

[159] Jason Feser, David Truong, Chandrima Das, Joshua J Carson, Jeffrey Kieft, Troy Harkness, and Jessica K Tyler. Elevated Histone Expression Promotes Life Span Extension. *Molecular Cell*, 39(5):724–735, 2010.

[160] Yuan Gao, Haiyun Gan, Zhenkun Lou, and Zhiguo Zhang. Asf1a resolves bivalent chromatin domains for the induction of lineage-specific genes during mouse embryonic stem cell differentiation. *Proceedings of the National Academy of Sciences*, 115(27):E6162–E6171, 2018.

[161] Li Tan, Zhonghe Ke, Gregory Tombline, Nicholas Macoretta, Kevin Hayes, Xiao Tian, Ruitu Lv, Julia Ablaeva, Michael Gilbert, Natarajan V Bhanu, Zuo-Fei Yuan, Benjamin A Garcia, Yujiang G Shi, Yang Shi, Andrei Seluanov, and Vera Gorbunova. Naked Mole Rat Cells Have a Stable Epigenome that Resists iPSC Reprogramming. *Stem Cell Reports*, 9(5):1721–1734, 2017.

[162] S Maegawa, G Hinkal, H S Kim, L Shen, L Zhang, and J Zhang. Widespread and tissue specific age-related DNA methylation changes in mice. *Genome Res*, 20:332–340, 2010.

[163] Dana Avrahami, Changhong Li, Jia Zhang, Jonathan Schug, Ran Avrahami, Shilpa Rao, Michael B. Stadler, Lukas Burger, Dirk Schübeler, Benjamin Glaser, and Klaus H. Kaestner. Aging-Dependent Demethylation of Regulatory Elements Correlates with Chromatin State and Improved Cell Function. *Cell Metabolism*, 22(4):619–632, 2015.

[164] 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.

[165] 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.

[166] András Sziráki, Alexander Tyshkovskiy, and Vadim N Gladyshev. Global remodeling of the mouse DNA methylome during aging and in response to calorie restriction. *Aging Cell*, 17(3):e12738, 2018.

[167] Vardhman K. Rakyan, Thomas A. Down, Siarhei Maslau, Toby Andrew, Tsun Po Yang, Huriya Beyan, Pamela Whittaker, Owen T. McCann, Sarah Finer, Ana M. Valdes, R. David Leslie, Panogiotis Deloukas, and Timothy D. Spector. Human aging-associated DNA hypermethylation occurs preferentially at bivalent chromatin domains. *Genome Research*, 20:434–439, 2010.

[168] Andrew E. Teschendorff, Usha Menon, Aleksandra Gentry-Maharaj, Susan J. Ramus, Daniel J. Weisenberger, Hui Shen, Mihaela Campan, Houtan Noushmehr, Christopher G. Bell, A. Peter Maxwell, David A. Savage, Elisabeth Mueller-Holzner, Christian Marth, Gabrijela Kocjan, Simon A. Gayther, Allison Jones, Stephan Beck, Wolfgang Wagner, Peter W. Laird, Ian J. Jacobs, and Martin Widschwendter. Age-dependent DNA methylation of genes that are suppressed in stem cells is a hallmark of cancer. *Genome Research*, 20(4):440–446, 2010.

[169] 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:R97, 2012.

[170] H Heyn, N Li, H J Ferreira, S Moran, D G Pisano, and A Gomez. Distinct DNA methylomes of newborns and centenarians. *Proc Natl Acad Sci U S A*, 109(26):10522–10527, 2012.

[171] K Day, L L Waite, A Thalacker-Mercer, A West, M M Bamman, and J D Brooks. Differential DNA methylation with age displays both common and dynamic features across human tissues that are influenced by CpG landscape. *Genome Biology*, 14:R102, 2013.

[172] Günter Raddatz, Sabine Hagemann, Dvir Aran, Jörn Söhle, Pranav P Kulkarni, Lars Kaderali, Asaf Hellman, Marc Winnefeld, and Frank Lyko. Aging is associated with

highly defined epigenetic changes in the human epidermis. *Epigenetics & Chromatin*, 6(1):36, 2013.

[173] C I Weidner, Q Lin, C M Koch, L Eisele, F Beier, and P Ziegler. Aging of blood can be tracked by DNA methylation changes at just three CpG sites. *Genome Biol*, 15:R24, 2014.

[174] Agustín F. Fernández, Gustavo F. Bayón, Rocío G. Urdinguio, Estela G. Toraño, María G. García, Antonella Carella, Sandra Petrus-Reurer, Cecilia Ferrero, Pablo Martinez-Camblor, Isabel Cubillo, Javier García-Castro, Jesús U. Delgado-Calle, Flor M. Pérez-Campo, José A. Riancho, Clara Bueno, Pablo Menéndez, Anouk Mentink, Katia Mareschi, Fabian Claire, Corrado Fagnani, Emanuela Medda, Virgilia Toccaceli, Sonia Brescianini, Sebastián Moran, Manel Esteller, Alexandra Stolzing, Jan De Boer, Lorenza Nistico, Maria A. Stazi, and Mario F. Fraga. H3K4me1 marks DNA regions hypomethylated during aging in human stem and differentiated cells. *Genome Research*, 25:27–40, 2015.

[175] Mikhail G Dozmorov. Polycomb repressive complex 2 epigenomic signature defines age-associated hypermethylation and gene expression changes. *Epigenetics*, 10(6):484–495, jun 2015.

[176] Tian Yuan, Yinming Jiao, Simone de Jong, Roel A Ophoff, Stephan Beck, and Andrew E Teschendorff. An Integrative Multi-scale Analysis of the Dynamic DNA Methylation Landscape in Aging. *PLOS Genetics*, 11(2):e1004996, 2015.

[177] 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.

[178] 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.

[179] 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.

[180] Philipp Voigt, Wee Wei Tee, and Danny Reinberg. A double take on bivalent promoters. *Genes and Development*, 27:1318–1338, 2013.

[181] B E Bernstein, T S Mikkelsen, X Xie, M Kamal, D J Huebert, and J Cuff. A bivalent chromatin structure marks key developmental genes in embryonic stem cells. *Cell*, 125(2):315–326, 2006.

[182] Deqiang Sun, Min Luo, Mira Jeong, Benjamin Rodriguez, Zheng Xia, Rebecca Hannah, Hui Wang, Thuc Le, Kym F. Faull, Rui Chen, Hongcang Gu, Christoph Bock, Alexander Meissner, Berthold Göttgens, Gretchen J. Darlington, Wei Li, and Margaret A. Goodell. Epigenomic profiling of young and aged HSCs reveals concerted changes during aging that reinforce self-renewal. *Cell Stem Cell*, 14(5):673–688, 2014.

[183] 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.

[184] Stephan H Bernhart, Helene Kretzmer, Lesca M Holdt, Frank Jühling, Ole Ammerpohl, Anke K Bergmann, Bernd H Northoff, Gero Doose, Reiner Siebert, Peter F Stadler, and Steve Hoffmann. Changes of bivalent chromatin coincide with increased expression of developmental genes in cancer. *Scientific Reports*, 6:37393, 2016.

[185] 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.

[186] Robert Lowe, Carl Barton, Christopher A Jenkins, Christina Ernst, Oliver Forman, Denise S Fernandez-Twinn, Christoph Bock, Stephen J Rossiter, Chris G Faulkes, Susan E Ozanne, Lutz Walter, Duncan T Odom, Cathryn Mellersh, and Vardhman K Rakyan. Ageing-associated DNA methylation dynamics are a molecular readout of lifespan variation among mammalian species. *Genome Biology*, 19(1):22, 2018.

[187] James West, Andrew E Teschendorff, and Stephan Beck. Age-associated epigenetic drift: implications, and a case of epigenetic thrift? *Human Molecular Genetics*, 22(R1):R7–R15, 2013.

[188] Mario F Fraga, Esteban Ballestar, Maria F Paz, Santiago Ropero, Fernando Setien, Maria L Ballestar, Damia Heine-Suñer, Juan C Cigudosa, Miguel Urioste, Javier Benitez, Manuel Boix-Chornet, Abel Sanchez-Aguilera, Charlotte Ling, Emma Carlsson, Pernille Poulsen, Allan Vaag, Zarko Stephan, Tim D Spector, Yue-Zhong Wu, Christoph Plass, and Manel Esteller. Epigenetic differences arise during the lifetime of monozygotic twins. *Proceedings of the National Academy of Sciences of the United States of America*, 102(30):10604–10609, 2005.

[189] Rudolf P Talens, Kaare Christensen, Hein Putter, Gonneke Willemsen, Lene Christiansen, Dennis Kremer, H Eka D Suchiman, P Eline Slagboom, Dorret I Boomsma, and Bastiaan T Heijmans. Epigenetic variation during the adult lifespan: cross-sectional and longitudinal data on monozygotic twin pairs. *Aging Cell*, 11(4):694–703, 2012.

[190] Irene Hernando-Herraez, Brendan Evano, Thomas Stubbs, Pierre-Henri Commere, Stephen Clark, Simon Andrews, Shahragim Tajbakhsh, and Wolf Reik. Ageing affects DNA methylation drift and transcriptional cell-to-cell variability in muscle stem cells. *bioRxiv*, page 500900, 2018.

[191] Celia Pilar Martinez-Jimenez, Nils Eling, Hung-Chang Chen, Catalina A. Vallejos, Aleksandra A. Kolodziejczyk, Frances Connor, Lovorka Stojic, Timothy F. Rayner, Michael J. T. Stubbington, Sarah A. Teichmann, Maike de la Roche, John C. Marioni,

and Duncan T. Odom. Aging increases cell-to-cell transcriptional variability upon immune stimulation. *Science*, 355(6332):1433–1436, 2017.

[192] Nicholas Stroustrup, Bryne E Ulmschneider, Zachary M Nash, Isaac F López-Moyado, Javier Apfeld, and Walter Fontana. The Caenorhabditis elegans Lifespan Machine. *Nature methods*, 10(7):665–70, 2013.

[193] Alexander Bürkle, María Moreno-Villanueva, Jürgen Bernhard, María Blasco, Gerben Zondag, Jan H J Hoeijmakers, Olivier Toussaint, Beatrix Grubeck-Loebenstein, Eugenio Mocchegiani, Sebastiano Collino, Efstathios S Gonos, Ewa Sikora, Daniela Gradinaru, Martijn Dollé, Michel Salmon, Peter Kristensen, Helen R Griffiths, Claude Libert, Tilman Grune, Nicolle Breusing, Andreas Simm, Claudio Franceschi, Miriam Capri, Duncan Talbot, Paola Caiafa, Bertrand Friguet, P Eline Slagboom, Antti Hervonen, Mikko Hurme, and Richard Aspinall. MARK-AGE biomarkers of ageing. *Mechanisms of Ageing and Development*, 151:2–12, 2015.

[194] Juulia Jylhävä, Nancy L Pedersen, and Sara Hägg. Biological Age Predictors. *EBioMedicine*, 21:29–36, 2017.

[195] Marjolein J Peters, Roby Joehanes, Luke C Pilling, Claudia Schurmann, Karen N Conneely, Joseph Powell, Eva Reinmaa, George L Sutphin, Alexandra Zhernakova, Katharina Schramm, Yana A Wilson, Sayuko Kobes, Taru Tukiainen, NABEC/UKBEC Consortium, Yolande F Ramos, Harald H H Göring, Myriam Fornage, Yongmei Liu, Sina A Gharib, Barbara E Stranger, Philip L De Jager, Abraham Aviv, Daniel Levy, Joanne M Murabito, Peter J Munson, Tianxiao Huan, Albert Hofman, André G Uitterlinden, Fernando Rivadeneira, Jeroen van Rooij, Lisette Stolk, Linda Broer, Michael M P J Verbiest, Mila Jhamai, Pascal Arp, Andres Metspalu, Liina Tserel, Lili Milani, Nilesh J Samani, Pärt Peterson, Silva Kasela, Veryan Codd, Annette Peters, Cavin K Ward-Caviness, Christian Herder, Melanie Waldenberger, Michael Roden, Paula Singmann, Sonja Zeilinger, Thomas Illig, Georg Homuth, Hans-Jörgen Grabe, Henry Völzke, Leif Steil, Thomas Kocher, Anna Murray, David Melzer, Hanieh Yaghootkar, Stefania Bandinelli, Eric K Moses, Jack W Kent, Joanne E Curran, Matthew P Johnson, Sarah Williams-Blangero, Harm-Jan Westra, Allan F McRae, Jennifer A Smith, Sharon L R Kardia, Iiris Hovatta, Markus Perola, Samuli Ripatti, Veikko Salomaa, Anjali K Henders, Nicholas G Martin, Alicia K Smith, Divya Mehta, Elisabeth B Binder, K Maria Nylocks, Elizabeth M Kennedy, Torsten Klengel, Jingzhong Ding, Astrid M Suchy-Dicey, Daniel A Enquobahrie, Jennifer Brody, Jerome I Rotter, Yii-Der I Chen, Jeanine Houwing-Duistermaat, Margreet Kloppenburg, P Eline Slagboom, Quinta Helmer, Wouter den Hollander, Shannon Bean, Towfique Raj, Noman Bakhshi, Qiao Ping Wang, Lisa J Oyston, Bruce M Psaty, Russell P Tracy, Grant W Montgomery, Stephen T Turner, John Blangero, Ingrid Meulenbelt, Kerry J Ressler, Jian Yang, Lude Franke, Johannes Kettunen, Peter M Visscher, G Gregory Neely, Ron Korstanje, Robert L Hanson, Holger Prokisch, Luigi Ferrucci, Tonu Esko, Alexander Teumer, Joyce B J van Meurs, and Andrew D Johnson. The transcriptional landscape of age in human peripheral blood. *Nature communications*, 6:8570, 2015.

[196] Toshiko Tanaka, Angelique Biancotto, Ruin Moaddel, Ann Zenobia Moore, Marta Gonzalez-Freire, Miguel A Aon, Julián Candia, Pingbo Zhang, Foo Cheung, Giovanna Fantoni, C H I Consortium, Richard D Semba, and Luigi Ferrucci. Plasma proteomic signature of age in healthy humans. *Aging Cell*, 17(5):e12799, 2018.

[197] Johannes Hertel, Nele Friedrich, Katharina Wittfeld, Maik Pietzner, Kathrin Budde, Sandra Van der Auwera, Tobias Lohmann, Alexander Teumer, Henry Völzke, Matthias Nauck, and Hans Jörgen Grabe. Measuring Biological Age via Metabonomics: The Metabolic Age Score. *Journal of Proteome Research*, 15(2):400–410, 2016.

[198] Fedor Galkin, Alexander Aliper, Evgeny Putin, Igor Kuznetsov, Vadim N Gladyshev, and Alex Zhavoronkov. Human microbiome aging clocks based on deep learning and tandem of permutation feature importance and accumulated local effects. *bioRxiv*, page 507780, 2018.

[199] J H Cole, S J Ritchie, M E Bastin, M C Valdés Hernández, S Muñoz Maniega, N Royle, J Corley, A Pattie, S E Harris, Q Zhang, N R Wray, P Redmond, R E Marioni, J M Starr, S R Cox, J M Wardlaw, D J Sharp, and I J Deary. Brain age predicts mortality. *Molecular Psychiatry*, 23:1385–1392, 2017.

[200] Sadiya S Khan, Benjamin D Singer, and Douglas E Vaughan. Molecular and physiological manifestations and measurement of aging in humans. *Aging Cell*, 16(4):624–633, 2017.

[201] Evgeny Putin, Polina Mamoshina, Alexander Aliper, Mikhail Korzinkin, Alexey Moskalev, Alexey Kolosov, Alexander Ostrovskiy, Charles Cantor, Jan Vijg, and Alex Zhavoronkov. Deep biomarkers of human aging: Application of deep neural networks to biomarker development. *Aging*, 8(5):1021–1033, 2016.

[202] Anne B Newman and Jason L Sanders. Telomere Length in Epidemiology: A Biomarker of Aging, Age-Related Disease, Both, or Neither? *Epidemiologic Reviews*, 35(1):112–131, 2013.

[203] 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.

[204] Simone Bork, Stefan Pfister, Hendrik Witt, Patrick Horn, Bernhard Korn, Anthony D Ho, and Wolfgang Wagner. DNA methylation pattern changes upon long-term culture and aging of human mesenchymal stromal cells. *Aging Cell*, 9(1):54–63, 2010.

[205] Elke Grönniger, Barbara Weber, Oliver Heil, Nils Peters, Franz Stäb, Horst Wenck, Bernhard Korn, Marc Winnefeld, and Frank Lyko. Aging and Chronic Sun Exposure

Cause Distinct Epigenetic Changes in Human Skin. *PLOS Genetics*, 6(5):e1000971, 2010.

[206] S Bocklandt, W Lin, M E Sehl, F J Sánchez, J S Sinsheimer, S Horvath, and E Vilain. Epigenetic predictor of age. *PLoS One*, 6(6):e14821, 2011.

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

[208] 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.

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

[210] 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.

[211] 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.

[212] 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.

[213] 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.

[214] Andrea M Polanowski, Jooke Robbins, David Chandler, and Simon N Jarman. Epigenetic estimation of age in humpback whales. *Molecular Ecology Resources*, 14(5):976–987, 2014.

[215] 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.

[216] Steve Horvath and Andrew J Levine. HIV-1 Infection Accelerates Age According to the Epigenetic Clock. *The Journal of infectious diseases*, 212(10):1563–73, 2015.

[217] Steve Horvath, Paolo Garagnani, Maria Giulia Bacalini, Chiara Pirazzini, Stefano Salvioli, Davide Gentilini, Anna Maria Di Blasio, Cristina Giuliani, Spencer Tung, Harry V. Vinters, and Claudio Franceschi. Accelerated epigenetic aging in Down syndrome. *Aging Cell*, 14(3):491–495, 2015.

[218] Steve Horvath, Wiebke Erhart, Mario Brosch, Ole Ammerpohl, Witigo von Schönfels, Markus Ahrens, Nils Heits, Jordana T. Bell, Pei-Chien Tsai, Tim D. Spector, Panos Deloukas, Reiner Siebert, Bence Sipos, Thomas Becker, Christoph Röcken, Clemens Schafmayer, and Jochen Hampe. Obesity accelerates epigenetic aging of human liver. *Proceedings of the National Academy of Sciences*, page 201412759, 2014.

[219] Morgan E Levine, Ake T Lu, Brian H Chen, Dena G Hernandez, Andrew B Singleton, Luigi Ferrucci, Stefania Bandinelli, Elias Salfati, JoAnn E Manson, Austin Quach, Cynthia D J Kusters, Diana Kuh, Andrew Wong, Andrew E Teschendorff, Martin Widschwendter, Beate R Ritz, Devin Absher, Themistocles L Assimes, and Steve Horvath. Menopause accelerates biological aging. *Proceedings of the National Academy of Sciences*, 113(33):9327–9332, 2016.

[220] Jacob K Kresovich, Zongli Xu, Katie M O'Brien, Clarice R Weinberg, Dale P Sandler, and Jack A Taylor. Methylation-Based Biological Age and Breast Cancer Risk. *JNCI: Journal of the National Cancer Institute*, page djz020, 2019.

[221] Anna Maierhofer, Julia Flunkert, Junko Oshima, George M. Martin, Thomas Haaf, and Steve Horvath. Accelerated epigenetic aging in Werner syndrome. *Aging*, 9(4):1143–1152, 2017.

[222] 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.

[223] Steve Horvath and Kenneth Raj. DNA methylation-based biomarkers and the epigenetic clock theory of ageing. *Nature Reviews Genetics*, 19(6):371–384, 2018.

[224] 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.

[225] 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.

[226] 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.

[227] Morgan E Levine, Ake T Lu, Austin Quach, Brian H Chen, Themistocles L Assimes, Stefania Bandinelli, Lifang Hou, Andrea A Baccarelli, James D Stewart, Yun Li, Eric A Whitsel, James G Wilson, Alex P Reiner, Abraham Aviv, Kurt Lohman, Yongmei Liu, Luigi Ferrucci, and Steve Horvath. An epigenetic biomarker of aging for lifespan and healthspan. *Aging*, 10(4):573–591, 2018.

[228] Ake T Lu, Austin Quach, James G Wilson, Alex P Reiner, Abraham Aviv, Kenneth Raj, Lifang Hou, Andrea A Baccarelli, Yun Li, James D Stewart, Eric A Whitsel, Themistocles L Assimes, Luigi Ferrucci, and Steve Horvath. DNA methylation GrimAge strongly predicts lifespan and healthspan. *Aging*, 11(2):303–327, 2019.

[229] Jerome Friedman, Trevor Hastie, and Rob Tibshirani. Regularization Paths for Generalized Linear Models via Coordinate Descent. *Journal of statistical software*, 33(1):1–22, 2010.

[230] Adam E. Field, Neil A. Robertson, Tina Wang, Aaron Havas, Trey Ideker, and Peter D. Adams. DNA Methylation Clocks in Aging: Categories, Causes, and Consequences. *Molecular Cell*, 71(6):882–895, 2018.

[231] 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.

[232] Steve Horvath, Vei Mah, Ake T Lu, Jennifer S Woo, Oi-Wa Choi, Anna J Jasinska, José A Riancho, Spencer Tung, Natalie S Coles, Jonathan Braun, Harry V Vinters, and L Stephen Coles. The cerebellum ages slowly according to the epigenetic clock. *Aging*, 7(5):294–306, 2015.

[233] Akina Hoshino, Steve Horvath, Akshayalakshmi Sridhar, Alex Chitsazan, and Thomas A Reh. Synchrony and asynchrony between an epigenetic clock and developmental timing. *Scientific Reports*, 9(1):3770, 2019.

[234] Arne Søraas, Mieko Matsuyama, Marcos de Lima, David Wald, Jochen Buechner, Tobias Gedde-Dahl, Camilla Lund Søraas, Brian Chen, Luigi Ferrucci, John Arne Dahl, Steve Horvath, and Shigemi Matsuyama. Epigenetic age is a cell-intrinsic property in transplanted human hematopoietic cells. *Aging Cell*, 18(2):e12897, 2019.

[235] Irina M Conboy, Michael J Conboy, Amy J Wagers, Eric R Girma, Irving L Weissman, and Thomas A Rando. Rejuvenation of aged progenitor cells by exposure to a young systemic environment. *Nature*, 433(7027):760–764, 2005.

[236] Christine J Huh, Bo Zhang, Matheus B Victor, Sonika Dahiya, Luis F Z Batista, Steve Horvath, and Andrew S Yoo. Maintenance of age in human neurons generated by microRNA-based neuronal conversion of fibroblasts. *eLife*, 5:e18648, 2016.

[237] 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.

[238] Donna Lowe, Steve Horvath, and Kenneth Raj. Epigenetic clock analyses of cellular senescence and ageing. *Oncotarget*, 7(8):8524–8531, 2016.

[239] Alejandro Ocampo, Pradeep Reddy, Paloma Martinez-Redondo, Aida Platero-Luengo, Fumiyuki Hatanaka, Tomoaki Hishida, Mo Li, David Lam, Masakazu Kurita, Ergin Beyret, Toshikazu Araoka, Eric Vazquez-Ferrer, David Donoso, Jose Luis Roman, Jinna Xu, Concepcion Rodriguez Esteban, Gabriel Nuñez, Estrella Nuñez Delicado, Josep M Campistol, Isabel Guillen, Pedro Guillen, and Juan Carlos Izpisua Belmonte. In Vivo Amelioration of Age-Associated Hallmarks by Partial Reprogramming. *Cell*, 167(7):1719–1733.e12, 2016.

[240] Thomas A. Rando and Howard Y. Chang. Aging, Rejuvenation, and Epigenetic Reprogramming: Resetting the Aging Clock. *Cell*, 148(1):46–57, 2012.

[241] Nelly Olova, Daniel J Simpson, Riccardo E Marioni, and Tamir Chandra. Partial reprogramming induces a steady decline in epigenetic age before loss of somatic identity. *Aging Cell*, 18(1):e12877, 2019.

[242] Salah Mahmoudi, Lucy Xu, and Anne Brunet. Turning back time with emerging rejuvenation strategies. *Nature Cell Biology*, 21(1):32–43, 2019.

[243] Tapash Jay Sarkar, Marco Quarta, Shravani Mukherjee, Alex Colville, Patrick Paine, Linda Doan, Christopher M Tran, Constance R Chu, Steve Horvath, Nidhi Bhutani, Thomas A Rando, and Vittorio Sebastiano. Transient non-integrative nuclear reprogramming promotes multifaceted reversal of aging in human cells. *bioRxiv*, page 573386, 2019.

[244] William Thomnson. *Popular lectures and addresses*. London Macmillan, 1889.

[245] 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.

[246] 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.

[247] 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.

[248] 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.

[249] 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.

[250] 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.

[251] 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.

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

[253] 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.

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

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

[256] 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.

[257] 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.

[258] 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.

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

[260] 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.

[261] 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.

[262] 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.

[263] 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.

[264] 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.

[265] 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.

[266] 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.

[267] 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.

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

[269] 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.

[270] 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.

[271] 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.

[272] 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.

[273] 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.

[274] 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.

[275] 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.

[276] 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.

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

[278] 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.

[279] 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.

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

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

[282] 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.

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

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

[285] 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.

[286] 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.

[287] 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.

[288] 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.

[289] 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.

[290] 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.

[291] 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.

[292] 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.

[293] 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.

[294] 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.

[295] 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.

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

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

[298] 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.

[299] 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.

[300] 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.

[301] 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.

[302] 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.

[303] 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.

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

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

[306] 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.

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

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

[309] 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.

[310] 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.

[311] 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.

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

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

[314] 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. *Genome Biology*, In review, 2019.

[315] Ake T Lu, Eilis Hannon, Morgan E Levine, Ke Hao, Eileen M Crimmins, Katie Lunnon, Alexey Kozlenkov, Jonathan Mill, Stella Dracheva, and Steve Horvath. Genetic variants near MLST8 and DHX57 affect the epigenetic age of the cerebellum. *Nature Communications*, 7:10561, 2016.

[316] Hans Tomas Bjornsson. The Mendelian disorders of the epigenetic machinery. *Genome Research*, 25(10):1473–1481, 2015.

[317] Erfan Aref-Eshghi, Laila C Schenkel, Hanxin Lin, Cindy Skinner, Peter Ainsworth, Guillaume Paré, David Rodenhiser, Charles Schwartz, and Bekim Sadikovic. The defining DNA methylation signature of Kabuki syndrome enables functional assessment of genetic variants of unknown clinical significance. *Epigenetics*, 12(11):923–933, 2017.

[318] Erfan Aref-Eshghi, Eric G Bend, Rebecca L Hood, Laila C Schenkel, Deanna Alexis Carere, Rana Chakrabarti, Sandesh C S Nagamani, Sau Wai Cheung, Philippe M Campeau, Chitra Prasad, Victoria Mok Siu, Lauren Brady, Mark A Tarnopolsky, David J Callen, A Micheil Innes, Susan M White, Wendy S Meschino, Andrew Y Shuen, Guillaume Paré, Dennis E Bulman, Peter J Ainsworth, Hanxin Lin, David I Rodenhiser, Raoul C Hennekam, Kym M Boycott, Charles E Schwartz, and Bekim Sadikovic. BAFopathies' DNA methylation epi-signatures demonstrate diagnostic utility and functional continuum of Coffin–Siris and Nicolaides–Baraitser syndromes. *Nature Communications*, 9(1):4885, 2018.

[319] Darci T Butcher, Cheryl Cytrynbaum, Andrei L Turinsky, Michelle T Siu, Michal Inbar-Feigenberg, Roberto Mendoza-Londono, David Chitayat, Susan Walker, Jerry Machado, Oana Caluseriu, Lucie Dupuis, Daria Grafodatskaya, William Reardon, Brigitte Gilbert-Dussardier, Alain Verloes, Frederic Bilan, Jeff M Milunsky, Raveen Basran, Blake Papsin, Tracy L Stockley, Stephen W Scherer, Sanaa Choufani, Michael Brudno, and Rosanna Weksberg. CHARGE and Kabuki Syndromes: Gene-Specific DNA Methylation Signatures Identify Epigenetic Mechanisms Linking These Clinically Overlapping Conditions. *The American Journal of Human Genetics*, 100(5):773–788, 2017.

[320] S Choufani, C Cytrynbaum, B H Y Chung, A L Turinsky, D Grafodatskaya, Y A Chen, A S A Cohen, L Dupuis, D T Butcher, M T Siu, H M Luk, I F M Lo, S T S Lam, O Caluseriu, D J Stavropoulos, W Reardon, R Mendoza-Londono, M Brudno, W T Gibson, D Chitayat, and R Weksberg. NSD1 mutations generate a genome-wide DNA methylation signature. *Nature Communications*, 6:10207, 2015.

[321] Laila C Schenkel, Charles Schwartz, Cindy Skinner, David I Rodenhiser, Peter J Ainsworth, Guillaume Pare, and Bekim Sadikovic. Clinical Validation of Fragile X Syndrome Screening by DNA Methylation Array. *The Journal of Molecular Diagnostics*, 18(6):834–841, 2016.

[322] Reid S Alisch, Tao Wang, Pankaj Chopra, Jeannie Visootsak, Karen N Conneely, and Stephen T Warren. Genome-wide analysis validates aberrant methylation in fragile X syndrome is specific to the FMR1locus. *BMC Medical Genetics*, 14(1):18, 2013.

[323] Laila C Schenkel, Kristin D Kernohan, Arran McBride, Ditta Reina, Amanda Hodge, Peter J Ainsworth, David I Rodenhiser, Guillaume Pare, Nathalie G Bérubé, Cindy Skinner, Kym M Boycott, Charles Schwartz, and Bekim Sadikovic. Identification of epigenetic signature associated with alpha thalassemia/mental retardation X-linked syndrome. *Epigenetics & Chromatin*, 10(1):10, 2017.

[324] Rebecca L Hood, Laila C Schenkel, Sarah M Nikkel, Peter J Ainsworth, Guillaume Pare, Kym M Boycott, Dennis E Bulman, and Bekim Sadikovic. The defining DNA methylation signature of Floating-Harbor Syndrome. *Scientific Reports*, 6:38803, 2016.

[325] Kimberly A Aldinger, Jasmine T Plummer, and Pat Levitt. Comparative DNA methylation among females with neurodevelopmental disorders and seizures identifies TAC1 as a MeCP2 target gene. *Journal of Neurodevelopmental Disorders*, 5(1):15, 2013.

[326] Daria Grafodatskaya, Barian H Y Chung, Darci T Butcher, Andrei L Turinsky, Sarah J Goodman, Sana Choufani, Yi-An Chen, Youliang Lou, Chunhua Zhao, Rageen Rajendram, Fatima E Abidi, Cindy Skinner, James Stavropoulos, Carolyn A Bondy, Jill Hamilton, Shoshana Wodak, Stephen W Scherer, Charles E Schwartz, and Rosanna Weksberg. Multilocus loss of DNA methylation in individuals with mutations in the histone H3 Lysine 4 Demethylase KDM5C. *BMC Medical Genomics*, 6(1):1, 2013.

[327] Kristin D Kernohan, Laila Cigana Schenkel, Lijia Huang, Amanda Smith, Guillaume Pare, Peter Ainsworth, Kym M Boycott, Jodi Warman-Chardon, Bekim Sadikovic, and Care4Rare Canada Consortium. Identification of a methylation profile for DNMT1-associated autosomal dominant cerebellar ataxia, deafness, and narcolepsy. *Clinical Epigenetics*, 8(1):91, 2016.

[328] George Leventopoulos, Sophia Kitsiou-Tzeli, Konstantinos Kritikos, Stavroula Psoni, Ariadni Mavrou, Emmanuel Kanavakis, and Helen Fryssira. A Clinical Study of Sotos Syndrome Patients With Review of the Literature. *Pediatric Neurology*, 40(5):357–364, 2009.

[329] Naohiro Kurotaki, Kiyoshi Imaizumi, Naoki Harada, Mitsuo Masuno, Tatsuro Kondoh, Toshiro Nagai, Hirofumi Ohashi, Kenji Naritomi, Masato Tsukahara, Yoshio Makita, Tateo Sugimoto, Tohru Sonoda, Tomoko Hasegawa, Yasuaki Chinen, Hiro-aki Tomita, Akira Kinoshita, Tsuyoshi Mizuguchi, Koh-ichiro Yoshiura, Tohru Ohta, Tatsuya Kishino, Yoshimitsu Fukushima, Norio Niikawa, and Naomichi Matsumoto. Haploinsufficiency of NSD1 causes Sotos syndrome. *Nature Genetics*, 30:365–366, 2002.

[330] Lorenzo Rinaldi, Debayan Datta, Judit Serrat, Lluis Morey, Guiomar Solanas, Alexandra Avgustinova, Enrique Blanco, José Ignacio Pons, David Matallanas, Alex Von Kriegsheim, Luciano Di Croce, and Salvador Aznar Benitah. Dnmt3a and Dnmt3b Associate with Enhancers to Regulate Human Epidermal Stem Cell Homeostasis. *Cell Stem Cell*, 19(4):491–501, 2016.

[331] Eric J Wagner and Phillip B Carpenter. Understanding the language of Lys36 methylation at histone H3. *Nature Reviews Molecular Cell Biology*, 13:115–126, 2012.

[332] Armelle Luscan, Ingrid Laurendeau, Valérie Malan, Christine Francannet, Sylvie Odent, Fabienne Giuliano, Didier Lacombe, Renaud Touraine, Michel Vidaud, Eric Pasmant, and Valérie Cormier-Daire. Mutations in SETD2 cause a novel overgrowth condition. *Journal of Medical Genetics*, 51(8):512–517, 2014.

[333] Stephen L McDaniel, Austin J Hepperla, Jie Huang, Raghuvar Dronamraju, Alexander T Adams, Vidyadhar G Kulkarni, Ian J Davis, and Brian D Strahl. H3K36 Methylation Regulates Nutrient Stress Response in Saccharomyces cerevisiae by Enforcing Transcriptional Fidelity. *Cell Reports*, 19(11):2371–2382, 2017.

[334] Zhuoyu Ni, Atsushi Ebata, Elham Alipanahiramandi, and Siu Sylvia Lee. Two SET domain containing genes link epigenetic changes and aging in Caenorhabditis elegans. *Aging Cell*, 11(2):315–325, 2012.

[335] Payel Sen, Weiwei Dang, Greg Donahue, Junbiao Dai, Jean Dorsey, Xiaohua Cao, Wei Liu, Kajia Cao, Rocco Perry, Jun Yeop Lee, Brian M. Wasko, Daniel T. Carr,

Chong He, Brett Robison, John Wagner, Brian D. Gregory, Matt Kaeberlein, Brian K. Kennedy, Jef D. Boeke, and Shelley L. Berger. H3K36 methylation promotes longevity by enhancing transcriptional fidelity. *Genes and Development*, 29(13):1362–1376, 2015.

[336] Mintie Pu, Zhuoyu Ni, Minghui Wang, Xiujuan Wang, Jason G. Wood, Stephen L. Helfand, Haiyuan Yu, and Siu Sylvia Lee. Trimethylation of Lys36 on H3 restricts gene expression change during aging and impacts life span. *Genes and Development*, 29(7):718–731, 2015.

[337] Dirk Schübeler. Function and information content of DNA methylation. *Nature*, 517(7534):321–326, 2015.

[338] Arunkumar Dhayalan, Arumugam Rajavelu, Philipp Rathert, Raluca Tamas, Renata Z. Jurkowska, Sergey Ragozin, and Albert Jeltsch. The Dnmt3a PWWP domain reads histone 3 lysine 36 trimethylation and guides DNA methylation. *Journal of Biological Chemistry*, 285:26114–26120, 2010.

[339] Tuncay Baubec, Daniele F. Colombo, Christiane Wirbelauer, Juliane Schmidt, Lukas Burger, Arnaud R. Krebs, Altuna Akalin, and Dirk Schübeler. Genomic profiling of DNA methyltransferases reveals a role for DNMT3B in genic methylation. *Nature*, 520(7546):243–247, 2015.

[340] Patricia Heyn, Clare V Logan, Adeline Fluteau, Rachel C Challis, Tatsiana Auchynnikava, Carol-Anne Martin, Joseph A Marsh, Francesca Taglini, Fiona Kilanowski, David A Parry, Valerie Cormier-Daire, Chin-To Fong, Kate Gibson, Vivian Hwa, Lourdes Ibáñez, Stephen P Robertson, Giorgia Sebastiani, Juri Rappsilber, Robin C Allshire, Martin A M Reijns, Andrew Dauber, Duncan Sproul, and Andrew P Jackson. Gain-of-function DNMT3A mutations cause microcephalic dwarfism and hypermethylation of Polycomb-regulated regions. *Nature Genetics*, 51(1):96–105, 2019.

[341] Wei Xie, Matthew D. Schultz, Ryan Lister, Zhonggang Hou, Nisha Rajagopal, Pradipta Ray, John W. Whitaker, Shulan Tian, R. David Hawkins, Danny Leung, Hongbo Yang, Tao Wang, Ah Young Lee, Scott A. Swanson, Jiuchun Zhang, Yun Zhu, Audrey Kim, Joseph R. Nery, Mark A. Urich, Samantha Kuan, Chia-an Yen, Sarit Klugman, Pengzhi Yu, Kran Suknuntha, Nicholas E. Propson, Huaming Chen, Lee E. Edsall, Ulrich Wagner, Yan Li, Zhen Ye, Ashwinikumar Kulkarni, Zhenyu Xuan, Wen-Yu Chung, Neil C. Chi, Jessica E. Antosiewicz-Bourget, Igor Slukvin, Ron Stewart, Michael Q. Zhang, Wei Wang, James A. Thomson, Joseph R. Ecker, and Bing Ren. Epigenomic Analysis of Multilineage Differentiation of Human Embryonic Stem Cells. *Cell*, 153(5):1134–1148, 2013.

[342] Hannah K Long, David Sims, Andreas Heger, Neil P Blackledge, Claudia Kutter, Megan L Wright, Frank Grützner, Duncan T Odom, Roger Patient, Chris P Ponting, and Robert J Klose. Epigenetic conservation at gene regulatory elements revealed by non-methylated DNA profiling in seven vertebrates. *eLife*, 2:e00348, 2013.

[343] Mira Jeong, Deqiang Sun, Min Luo, Yun Huang, Grant A Challen, Benjamin Rodriguez, Xiaotian Zhang, Lukas Chavez, Hui Wang, Rebecca Hannah, Sang-Bae Kim, Liubin Yang, Myunggon Ko, Rui Chen, Berthold Göttgens, Ju-Seog Lee, Preethi Gunaratne, Lucy A Godley, Gretchen J Darlington, Anjana Rao, Wei Li, and Margaret A

Goodell. Large conserved domains of low DNA methylation maintained by Dnmt3a. *Nature Genetics*, 46:17–23, 2013.

[344] Yuanyuan Li, Hui Zheng, Qiujun Wang, Chen Zhou, Lei Wei, Xuehui Liu, Wenhao Zhang, Yu Zhang, Zhenhai Du, Xiaowo Wang, and Wei Xie. Genome-wide analyses reveal a role of Polycomb in promoting hypomethylation of DNA methylation valleys. *Genome Biology*, 19(1):18, 2018.

[345] Ling Cai, Scott B. Rothbart, Rui Lu, Bowen Xu, Wei-Yi Chen, Ashutosh Tripathy, Shira Rockowitz, Deyou Zheng, Dinshaw J. Patel, C. David Allis, Brian D. Strahl, Jikui Song, and Gang Greg Wang. An H3K36 Methylation-Engaging Tudor Motif of Polycomb-like Proteins Mediates PRC2 Complex Targeting. *Molecular Cell*, 49(3):571–582, 2013.

[346] Haojie Li, Robert Liefke, Junyi Jiang, Jesse Vigoda Kurland, Wei Tian, Pujuan Deng, Weidi Zhang, Qian He, Dinshaw J. Patel, Martha L. Bulyk, Yang Shi, and Zhanxin Wang. Polycomb-like proteins link the PRC2 complex to CpG islands. *Nature*, 549(7671):287–291, 2017.

[347] Sophie Chantalat, Arnaud Depaux, Patrick Héry, Sophie Barral, Jean Yves Thuret, Stefan Dimitrov, and Matthieu Gérard. Histone H3 trimethylation at lysine 36 is associated with constitutive and facultative heterochromatin. *Genome Research*, 21:1426–1437, 2011.

[348] Taiping Chen, Naomi Tsujimoto, and En Li. The PWWP Domain of Dnmt3a and Dnmt3b Is Required for Directing DNA Methylation to the Major Satellite Repeats at Pericentric Heterochromatin. *Molecular and Cellular Biology*, 24(20):9048–9058, 2004.

[349] Aaron R Jeffries, Reza Maroofian, Claire G Salter, Barry A Chioza, Harold E Cross, Michael A Patton, I Karen Temple, Deborah Mackay, Faisal I Rezwan, Lise Aksglaede, Diana Baralle, Tabir Dabir, Matthew Frank Hunter, Arveen Kamath, Ajith Kumar, Ruth Newbury-Ecob, Angelo Selicorni, Amanda Springer, Lionel van Maldergem, Vinod Varghese, Naomi Yachelevich, Katrina Tatton-Brown, Jonathan Mill, Andrew H Crosby, and Emma Baple. Growth disrupting mutations in epigenetic regulatory molecules are associated with abnormalities of epigenetic aging. *bioRxiv*, page 477356, 2018.

[350] Huilin Huang, Hengyou Weng, Keren Zhou, Tong Wu, Boxuan Simen Zhao, Mingli Sun, Zhenhua Chen, Xiaolan Deng, Gang Xiao, Franziska Auer, Lars Klemm, Huizhe Wu, Zhixiang Zuo, Xi Qin, Yunzhu Dong, Yile Zhou, Hanjun Qin, Shu Tao, Juan Du, Jun Liu, Zhike Lu, Hang Yin, Ana Mesquita, Celvie L Yuan, Yueh-Chiang Hu, Wenju Sun, Rui Su, Lei Dong, Chao Shen, Chenying Li, Ying Qing, Xi Jiang, Xiwei Wu, Miao Sun, Jun-Lin Guan, Lianghu Qu, Minjie Wei, Markus Müschen, Gang Huang, Chuan He, Jianhua Yang, and Jianjun Chen. Histone H3 trimethylation at lysine 36 guides m6A RNA modification co-transcriptionally. *Nature*, 567(7748):414–419, 2019.

[351] Kyung-Won Min, Richard W Zealy, Sylvia Davila, Mikhail Fomin, James C Cummings, Daniel Makowsky, Catherine H Mcdowell, Haley Thigpen, Markus Hafner,

Sang-Ho Kwon, Constantin Georgescu, Jonathan D Wren, and Je-Hyun Yoon. Profiling of m6A RNA modifications identified an age-associated regulation of AGO2 mRNA stability. *Aging Cell*, 17(3):e12753, 2018.

[352] Gundula Streubel, Ariane Watson, Sri Ganesh Jammula, Andrea Scelfo, Darren J Fitzpatrick, Giorgio Oliviero, Rachel McCole, Eric Conway, Eleanor Glancy, Gian Luca Negri, Eugene Dillon, Kieran Wynne, Diego Pasini, Nevan J Krogan, Adrian P Bracken, and Gerard Cagney. The H3K36me2 Methyltransferase Nsd1 Demarcates PRC2-Mediated H3K27me2 and H3K27me3 Domains in Embryonic Stem Cells. *Molecular Cell*, 70(2):371–379.e5, 2018.

[353] Eugene Froimchuk, Younghoon Jang, and Kai Ge. Histone H3 lysine 4 methyltransferase KMT2D. *Gene*, 627:337–342, 2017.

[354] Zymo Research. EZ DNA methylation-Direct™ Kit. Technical report.

[355] Illumina. Infinium® HD Assay Methylation Protocol Guide. Technical report, 2015.

[356] William McLaren, Laurent Gil, Sarah E Hunt, Harpreet Singh Riat, Graham R S Ritchie, Anja Thormann, Paul Flicek, and Fiona Cunningham. The Ensembl Variant Effect Predictor. *Genome Biology*, 17(1):122, 2016.

[357] Andreas S Richter, Devon P Ryan, Fabian Kilpert, Fidel Ramírez, Steffen Heyne, and Thomas Manke. pyBigWig GitHub Repository.

[358] 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.

[359] Wanding Zhou, Huy Q Dinh, Zachary Ramjan, Daniel J Weisenberger, Charles M Nicolet, Hui Shen, Peter W Laird, and Benjamin P Berman. DNA methylation loss in late-replicating domains is linked to mitotic cell division. *Nature Genetics*, 50(4):591–602, 2018.

[360] 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.

[361] Adam Frankish, Alexandra Bignell, Andrew Berry, Andrew Yates, Anne Parker, Bianca M Schmitt, Bronwen Aken, Carlos García Girón, Daniel Zerbino, Eloise Stapleton, Fergal J Martin, Fiona Cunningham, If Barnes, Irina Sycheva, Jane Loveland, Jonathan M Mudge, Jose Manuel Gonzalez, Magali Ruffier, Marie-Marthe Suner, Matthew Hardy, Osagie G Izuogu, Sarah Donaldson, Shamika Mohanan, Thibaut Hourlier, Tiago Grego, Toby Hunt, Paul Flicek, James Wright, Jyoti S Choudhary, Julien Lagarde, Silvia Carbonell Sala, Roderic Guigó, Fernando Pozo, Laura Martínez, Michael L Tress, Tomás Di Domenico, Paul Muir, Barbara Uszczynska-Ratajczak, Benedict Paten, Ian T Fiddes, Joel Armstrong, Mark Diekhans, Tim J P Hubbard, Alexandre Reymond, Anne-Maud Ferreira, Jacqueline Chrast, Rory Johnson, Irwin Jungreis, Manolis Kellis, Baikang Pei, Fabio C P Navarro, Jinuri Xu, Yan Zhang, Mark Gerstein, and Cristina Sisu. GENCODE reference annotation for the human and mouse genomes. *Nucleic Acids Research*, 47(D1):D766–D773, 2018.

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

[363] Daniel E Martin-Herranz, António J. M. Ribeiro, Felix Krueger, Janet M Thornton, Wolf Reik, and Thomas M Stubbs. cuRRBS: simple and robust evaluation of enzyme combinations for reduced representation approaches. *Nucleic Acids Research*, 45(20):11559–11569, 2017.

[364] NIH Roadmap Epigenomics Mapping Consortium. Roadmap Epigenomics Chromatin State Model: raw data. https://egg2.wustl.edu/roadmap/data/byFileType/chromhmmSegmentations/ChmmModels/imputed12marks/jointModel/final/catMat/hg19_chromHMM_imputed25.gz.

[365] NIH Roadmap Epigenomics Mapping Consortium. Roadmap Epigenomics Chromatin State Model: emission parameters. https://egg2.wustl.edu/roadmap/data/byFileType/chromhmmSegmentations/ChmmModels/imputed12marks/jointModel/final/emissions_25_imputed12marks.png.

[366] Søren Kierkegaard. *Journals*. 1843.

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

[368] 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.

[369] 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.

[370] 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.

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

[372] 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.

[373] 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.

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

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

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

[377] 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.

[378] 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.

[379] 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.

[380] 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.

[381] 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.

[382] 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.

[383] 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.

[384] 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.

[385] 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.

[386] 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.

[387] 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.

[388] 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.

[389] 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.

[390] 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.

[391] 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.

[392] 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.

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

[394] 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.

[395] 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.

[396] 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.

[397] 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.

[398] 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.

[399] 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.

[400] 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.

[401] 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.

[402] 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.

[403] 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.

[404] 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.

[405] 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.

[406] 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.

[407] 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.

[408] 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.

[409] 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.

[410] 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.

[411] 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.

[412] 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.

[413] 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 & Molecular Biology*, 21:64–72, 2013.

[414] 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.

[415] 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.

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

[417] 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.

[418] 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.

[419] 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.

[420] 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.

[421] 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.

[422] 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.

[423] 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.

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

[425] 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.

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

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

[428] 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.

[429] 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.

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

[431] Antonio Machado. Proverbios y cantares XXIX. In *Campos de Castilla*. 1912.

[432] Anders Boeck Jensen, Pope L Moseley, Tudor I Oprea, Sabrina Gade Ellesøe, Robert Eriksson, Henriette Schmock, Peter Bjødstrup Jensen, Lars Juhl Jensen, and Søren Brunak. Temporal disease trajectories condensed from population-wide registry data covering 6.2 million patients. *Nature Communications*, 5:4022, 2014.

[433] Chantriolnt-Andreas Kapourani and Guido Sanguinetti. Melissa: Bayesian clustering and imputation of single-cell methylomes. *Genome Biology*, 20(1):61, 2019.

[434] J P Reddington, S M Perricone, C E Nestor, J Reichmann, N A Youngson, and M Suzuki. Redistribution of H3K27me3 upon DNA hypomethylation results in de-repression of Polycomb target genes. *Genome Biol*, 14:R25, 2013.

[435] Shijie C Zheng, Martin Widschwendter, and Andrew E Teschendorff. Epigenetic drift, epigenetic clocks and cancer risk. *Epigenomics*, 8(5):705–719, 2016.

[436] María Berdasco, Santiago Ropero, Fernando Setien, Mario F Fraga, Pablo Lapunzina, Régine Losson, Miguel Alaminos, Nai-Kong Cheung, Nazneen Rahman, and Manel Esteller. Epigenetic inactivation of the Sotos overgrowth syndrome gene histone methyltransferase NSD1 in human neuroblastoma and glioma. *Proceedings of the National Academy of Sciences*, 106(51):21830–21835, 2009.

[437] Srikanth Kudithipudi, Cristiana Lungu, Philipp Rathert, Nicole Happel, and Albert Jeltsch. Substrate Specificity Analysis and Novel Substrates of the Protein Lysine Methyltransferase NSD1. *Chemistry & Biology*, 21(2):226–237, 2014.

[438] Alicia Enguix, María D Cubiles, Sonia Barroso, Andrés Aguilera, María I Vaquero-Sedas, and Miguel A Vega-Palas. Epigenetic features of human telomeres. *Nucleic Acids Research*, 46(5):2347–2355, 2018.

[439] Claus Storgaard Sørensen, Gunnar Schotta, and Stine Jørgensen. Histone H4 Lysine 20 methylation: key player in epigenetic regulation of genomic integrity. *Nucleic Acids Research*, 41(5):2797–2806, 2013.

[440] Zohar Shipony, Zohar Mukamel, Netta Mendelson Cohen, Gilad Landan, Elad Chomsky, Shlomit Reich Zeliger, Yael Chagit Fried, Elena Ainbinder, Nir Friedman, and Amos Tanay. Dynamic and static maintenance of epigenetic memory in pluripotent and somatic cells. *Nature*, 513(7516):115–119, 2014.

[441] Steffen Rulands, Heather J Lee, Stephen J Clark, Christof Angermueller, Sébastien A Smallwood, Felix Krueger, Hisham Mohammed, Wendy Dean, Jennifer Nichols, Peter Rugg-Gunn, Gavin Kelsey, Oliver Stegle, Benjamin D Simons, and Wolf Reik. Genome-Scale Oscillations in DNA Methylation during Exit from Pluripotency. *Cell Systems*, 7(1):63–76.e12, 2018.

[442] X Shawn Liu, Hao Wu, Xiong Ji, Yonatan Stelzer, Xuebing Wu, Szymon Czauderna, Jian Shu, Daniel Dadon, Richard A Young, and Rudolf Jaenisch. Editing DNA Methylation in the Mammalian Genome. *Cell*, 167(1):233–247.e17, 2016.

[443] Gavin Kelsey, Oliver Stegle, and Wolf Reik. Single-cell epigenomics: Recording the past and predicting the future. *Science*, 358(6359):69–75, 2017.