ARCHIVIO ISTITUZIONALE
DELLA RICERCA

Alma Mater Studiorum Università di Bologna
Archivio istituzionale della ricerca

Evaluation of different computational methods for DNA methylation-based biological age

This is the final peer-reviewed author's accepted manuscript (postprint) of the following publication:

Published Version:
Di Lena, P., Sala, C., Nardini, C. (2022). Evaluation of different computational methods for DNA methylation-based biological age. BRIEFINGS IN BIOINFORMATICS, 23(4), 1-19 [10.1093/bib/bbac274].

Availability:
This version is available at: https://hdl.handle.net/11585/890186 since: 2023-05-31

Published:

DOI: http://doi.org/10.1093/bib/bbac274

Terms of use:

Some rights reserved. The terms and conditions for the reuse of this version of the manuscript are specified in the publishing policy. For all terms of use and more information see the publisher's website.

This item was downloaded from IRIS Università di Bologna (https://cris.unibo.it/).
When citing, please refer to the published version.

(Article begins on next page)

06 May 2026

Evaluation of different computational methods for DNA methylation-based biological age

Pietro Di Lena[1,*] and Claudia Sala[2] and Christine Nardini[3]

[1] Department of Computer Science and Engineering, University of Bologna, Mura Anteo Zamboni 7, 40126, Bologna, Italy
[2] Department of Experimental, Diagnostic and Specialty Medicine, University of Bologna, Via Massarenti 9, 40138, Bologna, Italy
[3] CNR IAC "Mauro Picone", CNR, Via dei Taurini 19, 00185, Rome, Italy
[*] Corresponding author. email:pietro.dilena@unibo.it

Abstract

In recent years there has been a widespread interest in researching biomarkers of ageing that could predict physiological vulnerability better than chronological age. Ageing, in fact, is one of the most relevant risk factors for a wide range of maladies, and molecular surrogates of this phenotype could enable better patients stratification. Among the most promising of such biomarkers is DNA methylation-based biological age. Given the potential and variety of computational implementations (epigenetic clocks), we here present a systematic review of such clocks. Furthermore, we provide a large-scale performance comparison across different tissues and diseases in terms of age prediction accuracy and age acceleration, a measure of deviance from physiology. Our analysis offers both a state-of-the-art overview of the computational techniques developed so far and a heterogeneous picture of performances, which can be helpful in orienting future research.

Keywords: methylation, epigenetic clock, regression, age acceleration

Introduction

Ageing is one of the most relevant risk factor for a wide range of diseases [1]. Given the extreme heterogeneity of ageing in humans, especially in advanced stages of life [2], chronological age alone has long been recognized insufficient to characterize the mechanisms behind this processes. Based on this rationale, identification of biomarkers of *biological ageing*, intended as an organism's increased risk of death while progressing throughout its life cycle, has become an active research field, whose dissemination initiated with the involvement of telomeres' lengths in the process of senescence (for an overview see [3]). Among several potential biomarkers of ageing [2, 4], DNA methylation is currently considered one of the most promising in a clinical translational perspective, given the stability of methylation and its high correlation with chronological age [5].

DNA methylation (DNAm) is an epigenetic modification involving the covalent addition of a methyl group to the 5'-carbon of cytosine in a CpG dinucleotide. Such alterations play crucial roles in numerous cellular processes, including gene expression [6], genomic imprinting [7] and embriogenesis [8].

DNAm-based epigenetic clocks predict chronological age from the methylation values of ten to hundreds of CpGs identified with statistical and machine learning approaches. The very first epigenetic clocks based on DNAm used statistical approaches to exploit possible correlations between methylation values and chronological age [9, 10]. Following the influential work of Horvath [11], which used a machine-learning approach to develop a DNAm clock across a large number of tissues and cell types, a plethora of clocks have been developed in recent years [12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34].

The success of Horvath's clock depends on the impressive accucy in estimating chronological age and, most importantly, in detecting *age accelerations* i.e., generally speaking, the difference between estimated and chronological age, a measure of deviance from physiology. To date, numerous studies, using a variety of implementations of age acceleration have shown that this parameter is associated with numerous health-related afflictions, such as cancer and neurodegenerative diseases [35].

Recently [36, 37], such surrogate measures of biological age have been incorporated beyond the aim of early DNAm clocks (primarily developed to predict chronological age) with additional lifestyle-associated indicators (for example smoking pack-years) and the ambition to predicting mortality and healthspan [38, 39, 40, 41]. Given the fundamental role of DNAm clocks *per se* and in this new generation of predictors, we deem relevant to revise the landscape of computational approaches exploited for DNAm-based epigenetic clocks for the estimate of chronological age in Humans. In particular, to highlight the predictive capabilities of DNAm clocks, we present a large-scale performance evaluation of several methods on heterogeneous benchmark sets, in terms of chronological age accuracy and age acceleration detection.

Such performance comparison highlights strengths and limits of the current approaches in particular with respect to the lack of standardization of age acceleration, hopefully contributing to shape the research exploiting this important biomarker.

DNA methylation

A variety of technologies are available to explore CpGs' methylation at smaller (EpiTYPER [42]) or larger (methylation profiling arrays [43], DNA bisulphite sequencing[44]) scale [45]. The Illumina Infinium profiling arrays are currently the most widespread assays.

From the earliest Illumina Infinium HumanMethylation27 BeadChip with 27K probes, the following Infinium HumanMethylation450 BeadChip contains over 450K probes, and nowadays the Infinium MethylationEPIC BeadChip microarray covers over 850K CpG methylation sites. The Infinium 450K and 850K platforms use two different chemical assays, type I and type II, while the older 27K platform uses only Type I. The type I Infinium assays use a pair of probes to measure the intensities of the methylated (M) and unmethylated (U) alleles at each CpG site. The Infinium type II assay requires only one probe per locus allowing detection of both alleles. The standard protocol quantifies the methylation level by the β *value* metric, calculated from the intensity of the M and U alleles from all cells in the tissue, as the ratio of fluorescent signals:

$$\beta = \max(M, 0)/[\max(M, 0) + \max(U, 0) + \alpha]$$

where typically $\alpha = 100$ [46]. Thus, for each CpG, β values range from 0 (indicating a tissue where none of the cells is methylated at such position on the DNA, CpG) to 1 (indicating a tissue where all cells are methylated at this position).

It has been noted [47] that the β-values generated by the two assays are not completely compatible and that those obtained with Infinium II are less accurate and reproducible than those obtained with Infinium I, which requires some bias correction to rescale the type II signals on the basis of type I. Furthermore, among the numerous aspects that impact on the

robustness of data there are also all sample preparation steps, including tissue conservation, nucleic acids extraction, library preparation, to name a few [48, 49]. Currently, a variety of pre-processing methods exist to transform raw methylation data into β values [50, 51], perform normalization, background, bias and batch correction [52, 53, 54, 55, 56].

Computational approaches

DNAm-based epigenetic clocks considered in this review are listed in Table 1 and include only epigenetic clocks for Homo Sapiens based on the Illumina Infinium assays. They are listed using their name, or, when unavailable, with the name of the first author in the corresponding publication. A graphical representation of the main features of the pre-trained clocks (i.e., all but CPFNN and EPM) in Table 1 is shown in Fig. 1.

For the implementation of such clocks, numerous strategies have been adopted regarding the data pre-processing and computation of DNAm age but all require as input two sets of samples for which the subjects' DNA methylation levels and age are available. The first *control (or training)* set includes samples from apparently healthy individuals. This set acts as reference, on the assumption that in such individuals biological and chronological age are identical. The second *case (or test)* set includes individuals affected by a disease, whose altered biological age is questioned.

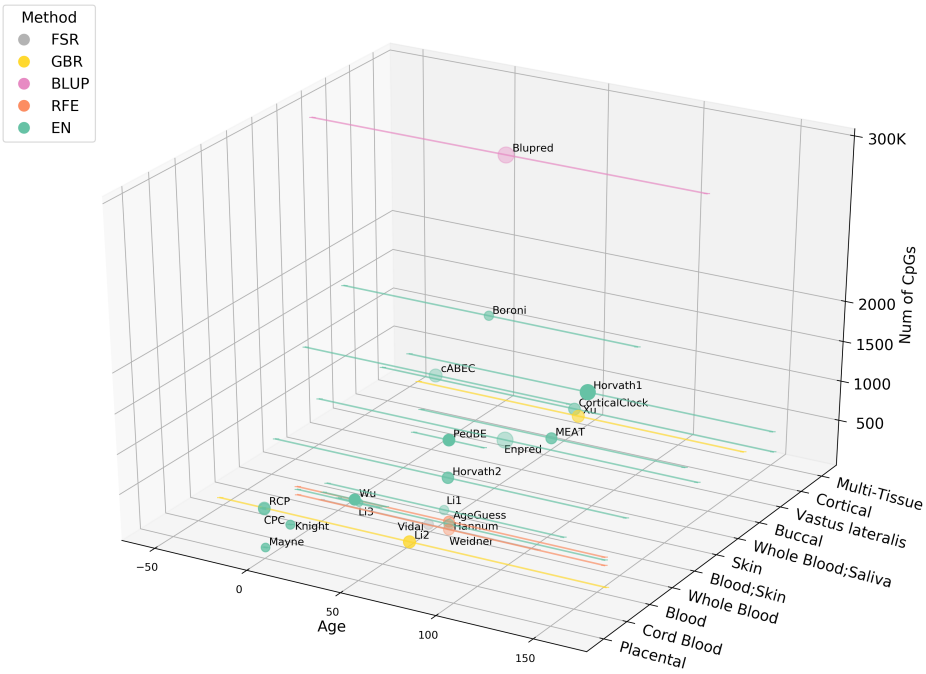

Figure 1: Graphical representation of the main features of the pre-trained DNA methylation-based aging clocks assessed in this work and detailed in Table 1. Each point represents a clock and its position in the 3D space is determined by the Tissue from which the training samples were derived (x-axis), the mean age (point position) and age range (error bar) of such samples (y-axis), and the number of CpGs used in the clock (z-axis). The point size is proportional to the logarithm of the number of training samples. The color of each point /error bar is related to the regression method used to train the clock, as detailed in Table 1. The point transparency is related to the number of pre-processing steps carried out before training the clock so that more pre-processing steps correspond to less transparency (see Table 1 for further details).

Table 1: List of DNA methylation-based aging clocks

			Data pre-processing						
Name	Tissue[1]	Age Range[2]	Filtering	Batch correction	Reduction	Age rescale	Method[3]	#Train[4]	#GpGs[5]
AgeGuess* [12]	Whole Blood	19–101 (y)	✓		✓		RFE	656	107
Blupred [13]	Whole Blood, Saliva	2–104 (y)	✓				BLUP	13,661	319,607
Bohlin* [14]	Cord Blood	NA (w)	✓		✓		LASSO	1068	96
Boroni* [15]	Skin	18–95 (y)	✓		✓		EN	249	2,266
CorticalClock [16]	Cortical	1–108 (y)	✓			✓	EN	1047	347
cABEC [17]	Whole Blood	19–88 (y)	✓				EN	2,227	1,892
CPC* [18]	Placental	5–42 (w)	✓		✓		EN	963	546
CPFNN [19]			✓	✓	✓		ANN		
Enpred [13]	Whole Blood, Saliva	2–104 (y)	✓				EN	13,661	514
EPM [20]			✓		✓		UPM		
Hannum [21]	Whole Blood	19–101 (y)	✓				EN	482	71
Horvath1 [11]	Multi-Tissue	0–100 (y)	✓	✓	✓	✓	EN	7,844	353
Horvath2 [22]	Blood, Skin	0–94 (y)	✓		✓	✓	EN	896	391
Knight* [23]	Cord Blood	24-42 (w)	✓	✓	✓		EN	207	148
Li1* [24]	Whole Blood	26–89 (y)	✓				EN	258	239
Li2* [25]	Blood	0–103 (y)	✓	✓	✓		GBR	1322	6
Li3* [26]	Whole Blood	6–17 (y)	✓		✓		EN	90	83
Mayne* [27]	Placental	8–42 (w)	✓	✓	✓		EN	170	62
MEAT [28]	Vastus lateralis	18–90 (y)	✓	✓		✓	EN	682	156
PedBE [29]	Buccal	0.2–19.5 (y)	✓	✓	✓	✓	EN	1,032	94
RCP* [18]	Placental	5–42 (w)	✓		✓		EN	1,102	558
Vidal [30]	Whole Blood	20–78 (y)			✓		FSR	390	8
Weidner [31]	Whole Blood	19–101 (y)			✓		RFE	656	3
Wu* [32]	Whole Blood	9–212 (m)	✓	✓	✓	✓	EN	716	111
Xu* [33]	Multi-Tissue	0–90 (y)	✓		✓		GBR	1280	13

* Not included in performance comparison due to unavailability of software implementation and/or independent test data (i.e. gestational clocks)

[1] Boroni: Dermis, epidermis, whole skin; PedBe: Buccal epithelial cells; Horvath1: 51 different tissues/cell types including blood, brain, muscle; Horvath2: Fibroblasts, keratinocytes, buccal cells, endothelial cells, lymphoblastoid, skin, blood, saliva; Xu: 15 different tissues including blood, brain, muscle.

[2] Age range in training data. (y) = years, (m) = months, (w) = weeks

[3] Regression methods. ANN: Artificial Neural Network, BLUP: Best Linear Unbiased Predictor, EN: Elastic Net, GBE: Gradient Booster Regressor, RFE: Recursive Feature Elimination, UPM: Universal PaceMaker

[4] Number of training samples (when available).

[5] Number of CpGs used in the clock (when available).

Data pre-processing: quality filtering

Quality filtering is typically applied on the training set for removing low-quality samples and probes. There is no universally agreed upon protocol. Here we review the three most common:

i Pre-processing tools that transform raw methylation data into β values provide *detection p-values* for every genomic position (a.k.a. probe) in each sample. Positions with non-significant detection p-value are often discarded as *missing values*. Samples and probes containing too many missing values are typically erased from the whole dataset, with exclusion threshold varying greatly. The remaining missing values are either imputed or entirely discarded.

ii Some probes are removed *a priori* to avoid gender bias or spurious signals [57, 58, 59]. In particular, most of the approaches filter probes related to sex chromosomes, as well as cross-hybridizing probes and/or probes found to contain SNPs (Bohlin [14], Boroni [15], AgeGuess [12], Hannum [21], CPC/RCP [18], cABEC [17], Mayne [27], PedBE [29], CorticalClock [16], MEAT [28], Wu [32], Blupred and Enpred [13]).

iii The detection of *outlier samples* is generally performed by means of Principal Component Analysis (PCA), with filtering procedures and cutoffs varying greatly from publication to publication (Hannum [21], Knight [23], cABEC [17], Li2 [25], CorticalClock [16], Wu [32], Xu [33], Blupred and Enpred [13]). The most stringent approach for outliers' detection is used in Horvath [11], where a sample is discarded when (i) the average value of its correlation with all other samples of the set or (ii) its maximum methylation level compared across all samples, are lower than a given (relatively high) threshold. Such filtering is not applied to cancer samples, which typically present severe alterations of the methylation levels [60]. A similar approach to detect outliers is used in CPC/RCP [18], where a gold standard methylation profile is built by taking the inter-sample median methylation value. Outliers are then detected by computing their Pearson correlation against the gold standard.

Data pre-processing: batch correction

To address the variability in the data caused by different experimental non-biological factors (*batch effect*) several approaches have been specifically designed for correction in methylation data from the Infinium platforms [61, 55]. The drawback lies in the possibility to correct biologically interesting factors unintentionally attributed to batch effects, such as methylation aberration in cancer samples [60]. This caveat may be the reason for the limited application of batch correction in the methods of Table 1.

In Horvath [11] the author creates a gold standard by taking the mean methylation values from the largest study in the training set. Next, a modified version of the BMIQ (Beta MIxture Quantile dilation of beta-values of Type II into Type I distribution) algorithm [62] is used to rescale the methylation values of each other study in the training set so that their distribution matches the gold standard. Such normalization to the gold standard is also performed by default in the prediction phase (although it can be omitted). The same approach has been adopted in MEAT [28], Knight [23], Mayne [27], PedBE [29] and Wu [32], although in Mayne, PedBE and Wu the normalization is not reported to be performed in the prediction phase. Similarly, in CPFNN [19] the authors use the *ComBat* function [56] (empirical Bayesian method to standardize mean and variance of methylation levels across studies) to reduce the batch effect on their training sets but not on the test data.

We further report that Li2 [25], Blupred and Enpred [13] perform sample standardization (remove mean and divide by standard deviation of the methylation levels sample-wise) to partially overcome the batch effect both in the training and in the prediction phase.

Data pre-processing: dimensionality reduction

Regression analysis in a high-dimensional space is often undesirable since data are sparse (curse of dimensionality, i.e. computational difficulties in data characterized by few observation and numerous features, typical of omics). In fact, in the absence of dimensionality reduction that limits the size of the feature space, an enormous amount of observations would be required to compensate and ensure that there are enough training examples.

Except for Elastic Net (EN) and other penalized regression models, most methods suffer from the curse of dimensionality. Most approaches in Table 1 perform dimensionality reduction by means of correlation measures, used to select the top-k methylomic probes having the largest correlation with chronological age. The major differences are in the correlation metrics and thresholds adopted. In AgeGuess [12], the correlation is computed by means of the *Maximal Information Coefficient* (MIC) [63]. In CPFNN [19], the authors use the Spearman's correlation. EPM [20], Li2 [25], Vidal [30], Weidner [31] and Xu [33] use Pearson's correlation. In this latter approach, the set of probes is further filtered by using a stepwise forward strategy, i.e. the variables are gradually included in the model according to their p-values.

Pearson's correlation for dimensionality reduction is used also in Horvath2 [22], although the regression method used to compute the biological age includes an EN penalization. In this case, the authors select the most statistically significant probes with high absolute correlation with chronological age, as well as 500 probes with the least significant correlation with age. In Boroni [15] the authors use different algorithmic approaches (generalized linear model via penalized maximum likelihood, boosted trees, random forests and Pearson's correlation) to select a top-ranking set of probes according to their importance in predicting the sample's age. Such set is further reduced by removing probes found to be highly correlated with each other (a common feature of CpGs [64]). In Bohlin [14], where the authors use an MM-type robust linear regression [65] to find the set of CpGs more strongly associated to (gestational) age. The regression model has been further adjusted for a set of covariates believed to be potential confounders: cell type composition estimates, child's sex, maternal smoking, maternal age, asthma and caesarian section. Finally, in the EN-based approach [32] the authors use Sure Independence Screening (SIS) [66] for dimensionality reduction.

It is finally worth mentioning that methods trained on datasets produced with different Illumina technologies typically consider only probes at the intersection between those represented in two or more Illumina arrays (Boroni [15], Horvath1 [11], Horvath2 [22], Knight [23], CPC/RPC [18], cABEC [17], Mayne [27], PedBE [29], Wu [32]).

Data pre-processing: age rescaling

Most DNA methylation-based epigenetic clocks assume a linear correlation between methylation levels and chronological age. However, there are evidences suggesting that methylation changes are more rapid early in life and progressively slow down [67]. Such non-linear correlations, while irrelevant for non-linear regressors, may affect the performances of linear models. In fact, in [11] a logarithmic relationship between predicted and chronological age can be observed between 0 and 20 years of age, and a linear relationship from 20 on. Thus, the author finds convenient, in terms of prediction performance, to transform the chronological age before regression with the following continuous function:

$$F(age) = \begin{cases} \log(age+1) - \log(x+1) & \text{if } age \leq x \\ (age - x)/(x+1) & \text{if } age > x \end{cases}$$

which is inverted on the model's output to get the age estimate:

$$F^{-1}(out) = \begin{cases} (x+1) \cdot e^{out} - 1 & \text{if } out < 0 \\ (x+1) \cdot out + x & \text{if } out \geq 0 \end{cases}$$

where the baseline age x is set to 20. This is the only approach reported in literature to deal with non-linear correlation between methylation levels and chronological age, and it has been adopted by numerous linear models (Horvath2 [22], Horvath1 [11], PedBE [29], CorticalClock [16], MEAT [28] and Wu [32], where the baseline age x is set to 48 months).

Models used for the epigenetic clocks

We here describe the regression models that have been exploited for the DNAm age estimation methods listed in Table 1.

Elastic Net (EN) and Least Absolute Shrinkage and Selection Operator (LASSO)

Elastic Net [68] is a penalized multivariate regression method that linearly combines the L_1 and L_2 penalties of the LASSO and ridge regression methods. Given a matrix $X \in \mathbb{R}^{m \times n}$ of methylation levels (β-values) and a vector $Y \in \mathbb{R}^m$ of phenotypic values (chronological age), where m is the number of observations (samples) and n the number of variables (probes), EN seeks to find the $b \in \mathbb{R}^n$ vector minimizing the objective function

$$\underset{b}{\operatorname{argmin}} \|Y - Xb\|^2 + \lambda \left(\frac{1-\alpha}{2} \|b\|_2^2 + \alpha \|b\|_1 \right)$$

where

$$\|Y - Xb\|^2 = \sum_{i=1}^{m} (Y_i - X_i \cdot b)^2 \text{ (sum of squared residuals)}$$

and

$$\|b\|_1 = \sum_{i=1}^{n} |b_i| \ (L_1 \text{ penalty})$$

$$\|b\|_2^2 = \sum_{i=1}^{n} b_i^2 \ (L_2 \text{ penalty})$$

The *mixing parameter* α linearly combines the L_1 and L_2 penalties: for $\alpha = 0$ the EN is equivalent to ridge regression, for $\alpha = 1$ to LASSO regression. The *regularization parameter* λ gives the contribution of the L_1 and L_2 penalties in the objective function. EN regression is the most appropriate when the data dimensionality is much larger than the sample size, since it performs simultaneously regression (minimization of the sum of squared residuals) and variable selection (penalize the size of parameter estimates through L_1 and L_2).

All the EN-based regression methods in Table 1 basically use the same training strategies: α is set to 0.5 and λ is estimated through a 10 fold cross validation on the training set, as implemented in the R package or python library *glmnet* [69]. In Bohlin [14] the same approach is taken with alpha set to 1 (i.e. LASSO regression).

In order to improve the robustness of the model, in Hannum [21] the authors sample the dataset with replacement a high number of times (500) and build a model for each bootstrap cohort. The final model includes only probes that are present in more than half of the cohorts. Furthermore, other covariates, such as gender and ethnicity, are included into the model and exempted from penalization (regularization).

Recursive Feature Elimination (RFE) and Forward Stepwise Regression (FSR)

Recursive feature elimination [70] is a feature selection method that iteratively builds a model and removes the less important features until a minimum number is left:

1. Train a regression model using all features.
2. Rank the features w.r.t. their importance in the model.

3. Remove the less important feature(s).

4. Repeat from 1. until a minimum number of features is left.

RFE automatically removes dependencies and redundancies that may exist in the model yielding to a more compact regression model. It is usually combined with a pre-filtering strategy to reduce the feature space and speed-up the computation.

In AgeGuess [12], the authors use a linear Support Vector Regressor (SVR) to build the model(s) and apply the RFE algorithm until all the features are removed. The features subset with best regression performance is then processed with a further redundancy removal step, performed with the iterative backward feature selection (BackFS) algorithm [71]. In Weidner [31], the authors use a linear regression model and fix to five the minimum number of features to be retained by RFE iterative filtering. Such five features (probes) were subsequently considered for locus-specific DNAm analysis by pyrosequencing, which led to select only three probes.

Forward Stepwise Regression is a regression approach that starts from an empty model and iteratively adds the variable that gives the best improvement to the model. Differently from RFE, which performs iterative elimination, FSR iteratively increases the set of variables:

1. Begin with an empty model.

2. Add the most significant variable.

3. Repeat from 2. until a stopping criterion is met.

In Vidal [30], the authors use a linear stepwise regressor.

Gradient Boosting Regression (GBR)

The Gradient Boosting Regressor is a Machine Learning technique, which defines a prediction model as an ensemble of *weak* learners [72]. The *boosting idea* is to train weak learners sequentially, each trying to correct its predecessor. Given a set of input X and output Y, the goal is to find a function F that minimizes the loss function L:

$$\text{argmin}_F \, L(Y, F(X))$$

where F is in the form of a weighted sum of functions f_i (weak learners), taken from some class of simple functions:

$$F(X) = \sum_{i=1}^{N} \gamma_i f_i(X) + const$$

The number of weak learners N is a parameter of the GBR method. Given a set of training examples $\{(X_1, Y_1), .., (X_m, Y_m)\}$ the GBR method tries to find a function F that minimizes the loss function by starting with a constant function f_1 and incrementally expands it with new weak learners.

Both Xu [33] and Li2 [25] use the Least Absolute Deviation (LAD) loss function

$$L(Y, F(X)) = \frac{1}{m} \sum_{i=1}^{M} |Y_i - F(X_i)|$$

and a high number of weak learners (400 in Xu and 300 in Li2).

Best Linear Unbiased Prediction (BLUP)

The Best Linear Unbiased Prediction is a linear mixed model for the estimation of random effects. Blupred [13] makes use of an approximate ridge regression analysis based on the BLUP model as developed in [73] and implemented in the R package *rrBLUP*. Given a matrix of methylation levels (β values) X and a vector of phenotypic data (chronological age) Y, the random effect model can be written as:

$$Y = Xb + \epsilon$$

where

$$b \sim N(0, I\sigma_b^2)$$

is a normally distributed vector of effects of β values on chronological age, and

$$\epsilon \sim N(0, I\sigma_\epsilon^2)$$

is a vector of normally distributed error terms (random effect). In a ridge regression analysis, b can be estimated as

$$b = (X^T X + I\lambda)^{-1} X^T Y$$

where $\lambda = \sigma_\epsilon^2/\sigma_b^2$ is the ratio between the (estimated) variances of residual and the effect sizes of the probe set, respectively.

Blupred does not use pre-filtering strategies, resulting in a regression model on a large set of probes.

Universal Pacemaker (UPM)

The Universal PaceMaker (UPM) is a statistical model for genome evolution [74]. In the UPM statistical framework the relative evolutionary rates of genes remain nearly constant but the absolute rates can change arbitrarily under the effect of various factors affecting the lineage. The UPM framework applied to epigenetic ageing gives rise to the Epigenetic Pacemaker (EPM) model [67] where, given a matrix $X \in \mathbb{R}^{m \times n}$ of methylation levels on m individuals (samples) and n methylation sites (probes), the methylation level X_{ij} is given by:

$$X_{ij} = x_j^0 + r_j s_i + \epsilon_{ij}$$

where x_j^0 is the initial methylation value of the j-th site, r_j is the rate of methylation change of the j-the site, s_i is the epigenetic state of the i-th individual and ϵ_{ij} is a normally distributed error term. Given a methylation matrix X the goal is to find the values of x^0, r and s that minimize the error between measured and predicted methylation values. This is equivalent to minimizing a quantity denoted as *residual sum of squares* (RSS)

$$\text{RSS} = \sum_{i=1}^{m}\sum_{j=1}^{n} \epsilon_{ij}^2 = \sum_{i=1}^{m}\sum_{j=1}^{n} (X_{ij} - x_j^0 - r_j s_i)^2$$

In EPM [20], the RSS minimization is achieved through a fast Conditional Expectation Maximization (CEM) approach [67]. The chronological age is used to provide an initial guess for the epigenetic state s_i of individual i. Such state is then updated at each iteration of the CEM algorithm. Differently from most approaches, EPM models methylation levels and not the chronological age, allowing for a non-linear relationship between the epigenetic state s_i and age.

Artificial Neural Network (ANN)

Artificial Neural Networks are a non-linear model widely used in Machine Learning [75] consisting of a collection of *artificial neurons* organized and connected only between contiguous layers. If we denote with n_i the neurons in the i-th layer (n_1 being the input layer) and with w_i the weights of the links that connect the i-the layer to the $(i+1)$-th layer, the values of the neurons in the $(i+1)$-th layer are given by the relation:

$$n_{i+1} = f(w_i \cdot n_i + b_i)$$

where f is called *activation function* and b_i is the *bias vector* (typically a constant vector of ones). There exist several types of activation functions, linear or non-linear. Traditionally the ANNs use the Mean Squared Error (MSE) as objective function. Given an input data matrix $X \in \mathbb{R}^{m \times n}$ (methylation levels) and an output vector $Y \in \mathbb{R}^m$ (chronological age), the ANN regression model searches the weights w that minimize the MSE:

$$\underset{w}{\operatorname{argmin}} \frac{1}{m} \sum_{i=1}^{m} (Y_i - F(X_i))^2$$

where F denotes the global ANN function. In their most common topology, ANNs have three layers: input, hidden and output. The input and output layer sizes depend on the size of the input and output data, respectively. The hidden layer size is chosen arbitrarily, knowing that too many neurons may result in overfitting on the training data. The ANN model suffers from overfitting also when the input size is too large in comparison to the number of training examples (curse of dimensionality).

In CPFNN [19], the ANN model consists of three layers. To prevent overfitting, the number of neurons in the hidden layer is kept small and dimensionality reduction is achieved by selecting only the probes with highest Spearman correlation with age. Furthermore, although the ANN can perform non-linear regression, CPFNN uses the linear Leaky ReLU activation function

$$f(x) = \begin{cases} x & x > 0 \\ ax & x \leq 0 \end{cases}$$

where $0 \leq a \leq 1$.

Performance evaluation

Benchmark datasets

Table 2: Benchmark dataset

		Samples		Age range			
GEO ID	Platform[1]	#Ctrls	#Cases	Ctrls	Cases	Tissue(s)	Study description
GSE105123[2]	450K	107	0	19-23	-	Peripheral Blood	Acclimatization to high altitude
GSE106648	450K	139	140	20-65	16-66	Peripheral Blood	Multiple Sclerosis
GSE157252	850K	50	84	20-56	19-58	Peripheral Blood	Recent onset of psycosis
GSE51032[3]	450K	424	421	34.4-71.9	34.9-72	Peripheral Blood	Breast, colorectal and other primary cancers
GSE67705	450K	44	142	27-66	25-67	Whole Blood	HIV-infected white males
GSE112179	850K	33	67	31-68	23-77	Frontal Cortex	Schizophrenia and Bipolar disorder
GSE66351	450K	84	106	18-88	62-97	Brain[4]	Alzheimer's disease
GSE89707	450K	75	67	21-96	24-87	Brain[5]	Schizophrenia disorder
GSE113018	450K	92	46	0.02-18.3	1-18.4	Peripheral Blood	Fetal Alcohol Spectrum Disorder (FASD)
GSE124366	450K	215	0	4-13	-	Blood, Mouth Mucosa	Healthy children (pediatric age)
GSE171140[6]	850K	195	0	18.26-48.79	-	Muscle	High Intensity Interval Training
GSE141441	450K	64	102	33-92	32-83	Breast	Breast cancer
GSE70977	450K	71	152	31-88	23-86	Mouth mucosa	Oral and pharyngeal carcinoma
GSE73832	450K	30	70	27-83	27-83	Digestive system[7]	Small Intestinal neuroendocrine tumors
GSE90124	450K	322	0	38.73-83.14	-	Skin	Punch skin biopsy of healthy female

[1] 450K: Illumina Infinum HumanMethylation450 BeadChip; 850K: Illumina Infinium MethylationEPIC

[2] 19 healthy subjects at 7 time points (not all subjects represented at all time points)

[3] Case samples developed some form of cancer over a follow-up period up to 15 years..

[4] Frontal, Occipital, Temporal cortex

[5] Cerebellum, Hippocampus, Striatum (putamen)

[6] 53 healthy subjects at 4 time points (not all subjects represented at all time points).

[7] Appendix, Colon, Liver, Mesomentum, Small intestine, Soft tissue

The list of benchmark datasets was obtained from GEO [76], selecting *Homo sapiens*, *Illumina methylation*, and containing at least 100 samples (Table 2). Such list has been further filtered by removing GEO studies that: i) were used in test or training in any of the methods in Table 1 that were further included in the performance comparison; ii) contain no control samples; iii) have no information about samples' chronological age and/or tissue.

We used methyLImp [64] to impute missing values separately on control and case samples for each dataset [77]. Upon data imputation, we further filtered all datasets still presenting missing values or CpGs preventing a fair performance comparison between the benchmarked clocks (which removed all the Illumina HumanMethylation27 BeadChip).

Control sample outliers were identified by PCA and inter-sample Person's correlation (as performed by some of the methods in Table 1 described above). Given the limited number of outliers (on average, less than one sample per dataset) and the acceptable accuracy when such samples were included, we retained all samples.

In the largest benchmark set, GSE51032, we separated the 421 individuals that developed some form of cancer from the 424 healthy controls.

Evaluation metrics

To validate the predicting ability of the methylation clocks and their prognostic power, we assessed two outcomes: 1) age estimation accuracy on controls, by measuring the *error* between the known (chronological) and predicted (biological) age of controls; 2) age acceleration of cases, by measuring the *difference* between the known (chronological) and predicted (biological) age of cases.

In particular, accuracy in age estimation on controls has been evaluated with two complementary metrics.

Pearson correlation (r), is a measure of goodness of fit between observed and predicted values, and quantifies the strength of the linear relationship between the two variables. It is defined as the ratio between the covariance of two variables and the product of their standard deviations

$$r(x,y) = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \overline{y})^2}}$$

where $\overline{x} = \frac{1}{n}\sum_i^n x_i$. It ranges from -1 to 1, where an absolute value close to 1 implies that a linear equation describes (almost) perfectly the relationship between the two variables. The sign determines the regression slope, where a negative value implies a negative correlation (anti-correlation) where one variable increases as the other decreases. A correlation close to 0 implies no linear relationship between the two variables. Pearson's r is invariant to constant shifts, i.e. $r(x,y) = r(x+a, y+b)$ for every constant a, b, which implies that the r metric does not tell how close, in absolute terms, the predicted values are to the observed ones.

Mean Absolute Error (MAE) is the arithmetic average of the absolute error between observed and predicted values:

$$\mathrm{MAE}(x,y) = \frac{1}{n}\sum_{i=1}^{n}|x_i - y_i|$$

Complementary to r, the MAE metric tells what is the prediction average error.

Regarding *Age acceleration*, which is broadly defined as the difference between DNAm and chronological age, different measures have been proposed in literature [78]. *Age acceleration* is uncorrelated with chronological age [78] and depends on the control dataset adopted for the analysis. In particular, case and control samples' chronological ages need to have approximately the same distributions. We here adopt the broadly used *Age Acceleration Residual* (AAR) on case samples, defined as the residuals between DNAm age of cases and the regression line obtained by regressing controls' DNAm age on chronological age. Other variants of age acceleration exist [78].

In more detail, if a and c are the slope and intercept, respectively, of the regression line

$$Y_{DNAm} = a \cdot X_{Age} + c$$

obtained by regressing DNAm age of controls Y_{DNAm} on chronological age X_{Age}, the AAR of case x with DNAm Age x_{DNAm} and chronological age x_{Age} is given by

$$x_{DNAm} - (a \cdot x_{Age} + c)$$

Experimental setup

Performances comparison has been carried out for all methods in Table 1 for which: i) software implementation is publicly available; ii) the list of CpGs together with their trained weights is provided. In some cases, the list of weights does not include the intercept (such as AgeGuess [12], Li3 [26]) or the performance in age estimation appeared much worse than expected (such as Li1 [24]). In both cases, the approaches were excluded from performance comparison; iii) a large and independent dataset for testing is available in GEO (see Table 2). This excluded all clocks whose age unit is in weeks (w) or months (m).

Two of the reviewed methods, i.e. CPFNN [19] and EPM [20], do not provide a pre-trained clock but a software implementation that can be used for training. We included such methods by performing training on the available data. In particular, following the approach described in [79], for each dataset in Table 2 we performed stratified sampling by age dividing the set of control samples in two subsets of approximately the same size (with all samples associated to a single individual, such as in GSE105123 and GSE171140, belonging to the same subset), eliminating the need for batch correction. The two methods have been trained in turn on one of the two subsets and tested on the other. Accuracy has been assessed only on the test samples. On the contrary, to perform age acceleration analysis on cases samples, the training has been performed on all control samples.

The two clocks have been run by keeping their original setting (software package). The EPM default threshold for CpGs selection (0.85 correlation with age) was modified to guarantee the best performance in training while avoiding an empty selection. CPFNN uniformly selects the top 3000 CpGs having the largest absolute Spearman's correlation with age.

Since Elastic Net is by far the most used regressing approach for DNAm Age estimation, for comparison we trained our own Elastic Net clock (ENC), on the same data used to train and test CPFNN and EPM. We used the R *glmnet* package with standard parameters: $\alpha = 0.5$ and λ estimated through 10-fold cross validation. We also performed age rescaling on training and test data, which enhanced performances on pediatric benchmarks. Although we did not use pre-filtering strategies for dimensionality reduction, the number of selected CpGs by the ENC clock varies from few tens to few hundreds.

Results and Discussions

Age prediction

Performances are evaluated separately on different tissues and age ranges and summarized in Tables 3-7. Each table (except Table 6, containing a single benchmark set) provides results on each GEO benchmark as well as on all samples

(*All*) analyzed jointly. Only epigenetic clocks compatible with the specific tissue and age ranges are shown in Tables 3-7 (complete results in supplementary data), with the exception of multiple-tissue clocks (Bluepred, Enpred, Horvath1, Horvath2) and clocks that have been trained on the benchmark data (CPFNN, ENC, EPM), whose performances are shown in all tables.

In Tables 3-7 we highlight in bold for each dataset the best performance(average r and MAE), as well as those (if any) for which no statistically significant difference with the best one is detectable (considering a significance threshold of 0.05 for the Benjamini-Hochberg adjusted p-values). The statistical significance has been assessed by means of the (two-sided) t-test for the MAE metric and with the paired correlation statistical test (as computed by the *paired.r* function in R) for the r metric.

Overall, we can observe that on the Blood sets age estimation accuracy is generally quite high, independently from the approach. Two exceptions are the clocks that use a small number of pre-trained CpGs, in particular Weidner (3 CpGs) and Vidal (8 CpGs), not surprisingly affected by small changes in the methylation values (experimental errors or batch effect) and thus less stable than clocks using a medium-high number of CpGs. In fact, in Tables 3-7 it is possible to observe that the performances of Blupred, Enpred, Horvath1 and Horvath2 are generally not dramatically affected by missing CpGs.

Except for set GSE105123, age estimation accuracy on the Blood benchmark sets is relevantly higher in comparison to other tissues (for most of the approaches). The overall low performances with respect to the r metric on GSE105123 can be referred to the small range of ages in such dataset, consisting of young individuals. The MAE metric, on the other hand, indicates that the absolute error in age prediction is still comparable to the average performances. This is in line with the widespread observation that DNAm-based clocks performances are dependent from the training sample size [80], and that blood is an accessible tissue.

On tissues different from Blood, tissue-specific clocks perform better than multi-tissue clocks, see CorticalClock (brain) and MEAT (muscle). Unexpectedly, PedBE, trained only on pediatric samples, does not perform better than general-purpose clocks on Pediatric sets (we note that it was trained on Mouth mucosa, and tested on buccal mucosa).

It is even more evident in Table 7 that multiple-tissue clocks do not have good performances, especially on Breast (GSE141441) and Digestive system (GSE73832) [11]. On the contrary, the overall performances on Mouth Mucosa (GSE70977) and Skin (GSE90124) are in line with those observed on the Muscle and Brain sets (r metric).

The performances of clocks trained on the benchmark data (CPFNN, ENC, EPM) are overall quite good and comparable or better than those of the pre-trained clocks. Both CPFNN and ENC seem to be slightly better performing than the EPM. Although such methods do not suffer from batch effect (training and testing belong to the same study), the number of training examples is generally quite small in comparison to the training data of the pre-trained clocks (cf. Table 1). The effect of a limited number of training samples is particularly evident in the benchmark with a small number of controls, such as GSE73832 (Digestive system, 30 controls implying training on 15 samples) and GSE112179 (Brain, 33 controls). However, we also point out the accuracy gap between the GSE157252 (Blood, 50 controls) and the GSE141441 (Breast, 64 controls). While the performances on GSE157252 are still acceptable, those on GSE141441 are quite poor. This may be a suggestion that the methylation signal is much more stable and stronger in Blood than in other tissues, which implies that DNAm-blood clocks are generally more robust.

To conclude, we observe that all approaches use a different number and type of CpGs for age estimation (see Supplementary Materials). The set of CpGs shared by two or more approaches is typically quite small, although there are some CpGs that are selected more often than others [81, 82]. Only few CpGs are shared by more than 5 clocks and no CpG is shared by all clocks (the largest overlap involves 7 clocks). All such CpGs and the related genes have been already reported to be aging-associated in independent studies (see Supplementary Materials). An alternative metric of similarity between clocks is in terms of distance between their estimates. That is, for a pair of clocks we use as a measure of similarity the mean pairwise absolute difference (MAE) between their estimated ages (on control samples only).The resulting hierarchical clustering in Fig. 2 shows that the training samples affect the output similarity of two methods more than the set of CpGs selected in training. In particular, CPFNN, ENC and EPM, trained on the same set have a

high similarity in output: MAE(ENC,CPFNN)=2.17, MAE(CPFNN,EPM)=3.67, MAE(ENC,EPM)=3.95. This happens despite different computational approaches and different number of CpGs. The same holds for Blupred (∼320k CpGs) and Enpred (514 CpGs): MAE(Blupred,Enpred)=4.89. The obvious implication is that, the main factor influencing the accuracy is the quality and quantity of the training data, as observed in [80].

Table 3: Performance comparison on blood benchmark sets (control samples only)

		cABEC		Blupred		Enpred		Hannum		Horvath1		Horvath2		Vidal		Weidner		CPFNN		ENC		EPM	
GEO ID	#	r	MAE	r	MAE	r	MAE	r	MAE	r	MAE	r	MAE	r	MAE	r	MAE	r	MAE	r	MAE	r	MAE
GSE105123	107	**0.48**	6.41	**0.43**	4.43	0.34	2.33	0.16	5.20	0.21	3.89	**0.50**	2.83	0.18	13.03	0.15	18.49	-0.17	**1.21**	-0.30	1.34	-0.34	**1.28**
GSE106648	139	**0.98**	3.05	**0.981.73**		**0.98**	2.06	0.93	7.50	0.93	5.81	**0.98**	2.25	0.77	12.00	0.63	7.13	0.93	3.27	0.94	3.13	0.90	3.78
GSE157252[1]	50	**0.96**	4.19	**0.97**	4.64	**0.96**	6.68	0.84	9.13	0.88	3.65	**0.952.15**		0.78	12.29	0.63	18.31	0.78	4.94	0.79	4.78	0.74	4.83
GSE51032	424	**0.912.65**		**0.922.47**		**0.922.58**		0.84	6.13	0.79	3.91	0.89	**2.53**	0.61	4.67	0.46	7.26	0.86	**2.73**	0.86	2.88	0.78	3.78
GSE67705	440.95	**6.48**		**0.98**9.92		0.97	8.49	0.89	29.61	0.92	10.11	0.96	8.80	0.82	**7.32**	0.41	34.36	0.80	**7.18**	0.73	**6.59**	0.83	**6.39**
All	764	**0.97**	3.57	0.96	3.18	**0.97**	3.06	0.88	7.80	0.92	4.59	**0.962.86**		0.85	7.82	0.43	11.09	0.95	**3.01**	0.96	**3.05**	0.94	3.65

[1] Blupred: 319299/319607, CpGs available in input data, Hannum: 65/71, Horvath1: 334/353

Table 4: Performance comparison on brain benchmark sets (control samples only)

		CorticalClock		Blupred		Enpred		Horvath1		Horvath2		CPFNN		ENC		EPM	
GEO ID	#	r	MAE	r	MAE	r	MAE	r	MAE	r	MAE	r	MAE	r	MAE	r	MAE
GSE112179[1]	33	**0.87**	8.59	0.65	**4.94**	0.74	23.73	0.66	33.95	0.45	36.46	0.06	**6.43**	0.43	**5.66**	-0.06	7.11
GSE66351[2]	84	**0.93**	6.50	0.79	21.98	0.58	20.61	0.88	11.39	0.50	22.47	**0.95**	**4.50**	0.93	**4.96**	0.90	6.42
GSE89707[3]	75	**0.95**	**5.38**	0.78	13.39	0.79	15.94	**0.92**	**6.64**	0.77	22.43	0.89	7.20	0.86	7.54	0.68	11.79
All	192	**0.94**	**6.42**	0.66	15.70	0.72	19.32	0.49	13.41	0.67	24.86	0.91	**5.89**	0.90	**6.09**	0.79	8.64

[1] Bluepred: 317563/319607 CpGs available in input data, Enpred: 513/514, Horvath1: 335/353, Horvath2: 389/391. [2]Blupred: 319202/319607. [3]CorticalClock: 345/347, Blupred: 316692/319607, Enpred: 513/514, Horvath1: 347/353, Horvath2: 378/391

Table 5: Performance comparison on pediatric benchmark set (control samples only)

		PedBE		Blupred		Enpred		Horvath1		Horvath2		CPFNN		ENC		EPM	
GEO ID	#	r	MAE	r	MAE	r	MAE	r	MAE	r	MAE	r	MAE	r	MAE	r	MAE
GSE113018	92	0.87	4.56	**0.96**	2.68	**0.94**	5.66	0.90	2.37	**0.96**	1.94	**0.95**	**1.23**	**0.96**	**1.23**	0.92	1.53
GSE124366[a]	110	0.38	2.64	0.16	6.08	0.32	5.80	0.27	1.98	0.34	2.75	**0.65**	**1.46**	**0.58**	**1.55**	**0.63**	1.79
GSE124366[b]	105	0.61	3.56	0.71	1.91	0.75	3.11	0.61	2.06	0.74	1.56	**0.83**	**1.28**	**0.83**	**1.35**	**0.82**	**1.43**
All	307	0.15	3.53	0.52	3.64	0.44	4.84	0.75	2.13	0.72	2.10	**0.91**	**1.33**	**0.91**	**1.39**	0.88	1.59

[a] Buccal [b]Peripheral Blood

Table 6: Performance comparison on muscle benchmark set (control samples only)

		MEAT		Blupred		Enpred		Horvath1		Horvath2		CPFNN		ENC		EPM	
GEO ID	#	r	MAE	r	MAE	r	MAE	r	MAE	r	MAE	r	MAE	r	MAE	r	MAE
GSE171140	195	**0.92**	**2.48**	0.63	7.45	0.80	5.76	0.66	6.92	0.71	17.05	0.84	3.82	0.86	3.75	0.86	3.39

[1] Bluepred: 305461/319607 CpGs available in input data, Enpred: 501/514, Horvath1: 325/353, Horvath2: 380/391

Table 7: Performance comparison on multiple-tissue benchmark sets (control samples only)

		Blupred		Enpred		Horvath1		Horvath2		CPFNN		ENC		EPM	
GEO ID	#	r	MAE	r	MAE	r	MAE	r	MAE	r	MAE	r	MAE	r	MAE
GSE141441	64	**0.38**	**16.01**	**0.38**	**17.89**	**0.30**	**18.03**	**0.18**	19.78	**0.24**	**13.45**	**0.11**	**14.64**	**0.31**	**13.66**
GSE70977[1]	71	**0.92**	**2.90**	**0.92**	**3.33**	0.83	18.25	**0.91**	3.81	0.82	4.58	0.52	6.72	0.65	7.09
GSE73832	30	**0.50**	**13.13**	**0.68**	**11.33**	**0.73**	**10.39**	**0.49**	**12.91**	**0.38**	**11.58**	0.22	**12.92**	0.39	**11.54**
GSE90124[2]	322	0.78	5.81	0.80	9.28	0.79	5.41	0.84	6.04	**0.91**	**3.15**	0.89	3.53	0.66	6.26
All	487	0.59	7.18	0.53	9.67	0.53	9.25	0.50	7.94	**0.71**	**5.23**	0.63	6.03	0.57	7.68

[1] Bluepred: 303300/319607 CpGs available in input data, Enpred: 491/514, Horvath1: 324/353, Horvath2: 355/391. [2]Bluepred: 309340/319607, Enpred: 504/514, Horvath1: 352/353, Horvath2: 389/391

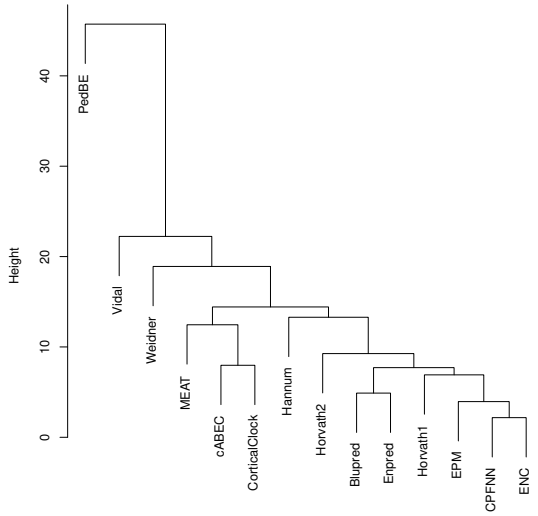

Figure 2: Hierarchical clustering of clock similarity on control samples built with MAE distance and complete clustering

Age Acceleration

Age acceleration has been analyzed on the datasets in Table 2 containing case samples, and diseases-wise for datasets containing more than one disease. Furthermore, in some datasets controls' and cases ages distribution is not perfectly comparable (see GSE66351). In such datasets we sampled the controls to achieve distributions with no statistically significant difference (t-test). We further used the (one-sided) Mann-Whitney-Wilcoxon test (with Benjamini-Hochberg correction) to detect whether there is a statistically significant acceleration (or deceleration) in cases versus controls. All results are shown in Fig. 3-6. Only epigenetic clocks compatible with the specific tissue and age ranges are shown in Fig. 3-6 (complete results in supplementary data).

Beyond the computational results obtained, AAR is meaningful for its potential clinical application (early diagnosis). We therefore search in this section to briefly make sense of these findings against the relevant literature.

Some diseases show a relatively coherent and straightforward interpretation. This includes infection with human immunodeficiency virus type 1 (HIV), known to be associated with accelerated ageing [83, 84, 85], confirmed in our analysis (GSE67705 in Fig 3) by nearly all clocks with statistically significant age acceleration. Even more promising is the association with cancer [86], confirmed in our tests (Fig 3), in most cancer tissues. Nevertheless, the situation is heterogenous for breast cancer (breast tissue), where we observe weak acceleration only for the methods that have been trained on the controls from the same dataset (CPFNN,ENC,EPM), and oral carcinoma (mouth mucosa), where one method (Horvath2) actually shows significant deceleration in cases vs controls. Such a scenario may be interpreted as the direct consequence of phenotypic and functional heterogeneity of human cancers. Furthermore, it is worth noticing that on dataset GSE73832 case samples show strong signs of age acceleration in comparison to control samples, although all epigenetic clocks achieve low age prediction performances on controls (cf. Table 7). We do not observe such strong trends on other datasets where the age prediction accuracy on controls is much higher. This phenomenon has already been observed in [80].

More relevant in a translational perspective is the hypothesis that age acceleration observed in blood may alert on

increased risk of solid cancer, as it has already been reported [87, 88, 89]. The GSE51032 benchmark set allows us to perform some comparison tests in this direction. This dataset contains DNA methylation profiles from peripheral blood samples of 845 individuals enrolled in the EPIC study [90]. Although all participants where apparently healthy at the time of blood sampling, 421 of them developed some form of cancer over a follow-up period covering 15 years. Age acceleration analysis on four types of cancer in this dataset are shown in Fig. 4. As above for tissue-specific clocks, results are mixed and overall, only colon cancer tissue shows significant signs of age acceleration for some of the clocks. For other experiments a trend of age acceleration can be observed in the Cancer subset, although conclusions on this dataset are unlikely to be robust since it contains only 20 samples, including non-specified type of cancer, prostate, bladder and hematological cancer. For the remaining Breast and Rectal cancer tissue, there is no significant difference between cases and controls. This is consistent with an earlier analysis [87] on the same dataset that confirms trends of age acceleration for colorectal cancer, while only single probe methylage markers (ELOVL2 and FHL2, not included in our analysis due to unavailability of implementation) confirm breast cancer, a limited result in line with [11].

A whole different clinical area, related to mental disorders, has also been interested by the potential impact of methylation ageing. In this case, results 5 are of more difficult interpretation in light of the existing literature. Schizophrenia (GSE157252,GSE112179,GSE89707) on different tissues (blood, brain and prefrontal cortex) presents contradictory results with a contradictory literature (accelerated ageing [91], decelerated ageing [92]), furthermore results on different tissues also give opposite results. We observe significant acceleration in Psychotic disorder (peripheral blood), none on Delusional disorder (peripheral blood) nor Bipolar disorder (frontal cortex), in which all methods are discordant [93, 94].

A similar landscape is shown in Fig 6 where we collect a miscelleaneous set of benchmarks referring to: Fetal Alchool Disorder (FASD), Alzheimer and Sclerosis. For FASD we can observe discordant results while evidences of age acceleration has been reported in [95], although, differently from our analysis, on blood tissue and only with the mortality clock GrimAge (mortality clocks are not included in our analysis). Age acceleration in Alzheimer has been reported in [96, 97] both with Horvath1 clock and on brain samples once corrected by sex and neuron count. Finally, age acceleration in sclerosis has been reported [98] on whole blood samples, although the authors report statistically significant acceleration only with the mortality clock PhenoAge (not included in our analysis).

Overall, it is difficult to conclude which of the parameters is responsible for the variability of the results and the seeming inconsistencies with what already reported in literature. Certainly, the specific computational methods have some impact on the AAR outcomes. For instance, although the three methods CPFNN, ENC and EPM (trained on the same data) are highly similar in terms of prediction output (see section "Age prediction"), their outcome is not always consistent in the AAR analysis (see, for example, Delusional and Psychotic disorders in Fig 5 or Colon cancer in Fig 3). However, we remark that the largest differences are observed mainly for EPM (which does not perform regression on chronological age), while CPFNN and ENC performances are relatively more consistent. Although less marked (see Delusional disorder and Bipolar in Fig 5), this is true also for Blupred and Enpred (again, trained on the same data). We can only explain this outcome by remarking that DNAm clocks are typically trained on healthy control samples only, thus their specific objective functions may fail to identify (and properly weight) disease-specific CpGs for all possible disorders. This may be the reason why different methods, trained on the same data, have more similar performances on healthy samples than on disease-related samples. The highly variable behaviour of similar methods, trained on different datasets, is even more complex to analyze (compare, for example, cABEC and Horvath1, both EN-based). In such case, we can only speculate that both training data and tissue specificity are the two main factors influencing performances. For what we can observe from our analysis, it is difficult to identify one single computational method or pre-trained clock which is absolutely reliable in terms of age acceleration detection and, consequently, a robust age acceleration analysis cannot rely on a single clock but it should be performed with different clocks (all compatible with the tissue of interest). To conclude, we further remark that the term acceleration itself refers to a variety of measures. This makes comparison of age acceleration analyses from different studies at least challenging, since also the specific acceleration metric introduces variability in the analysis.

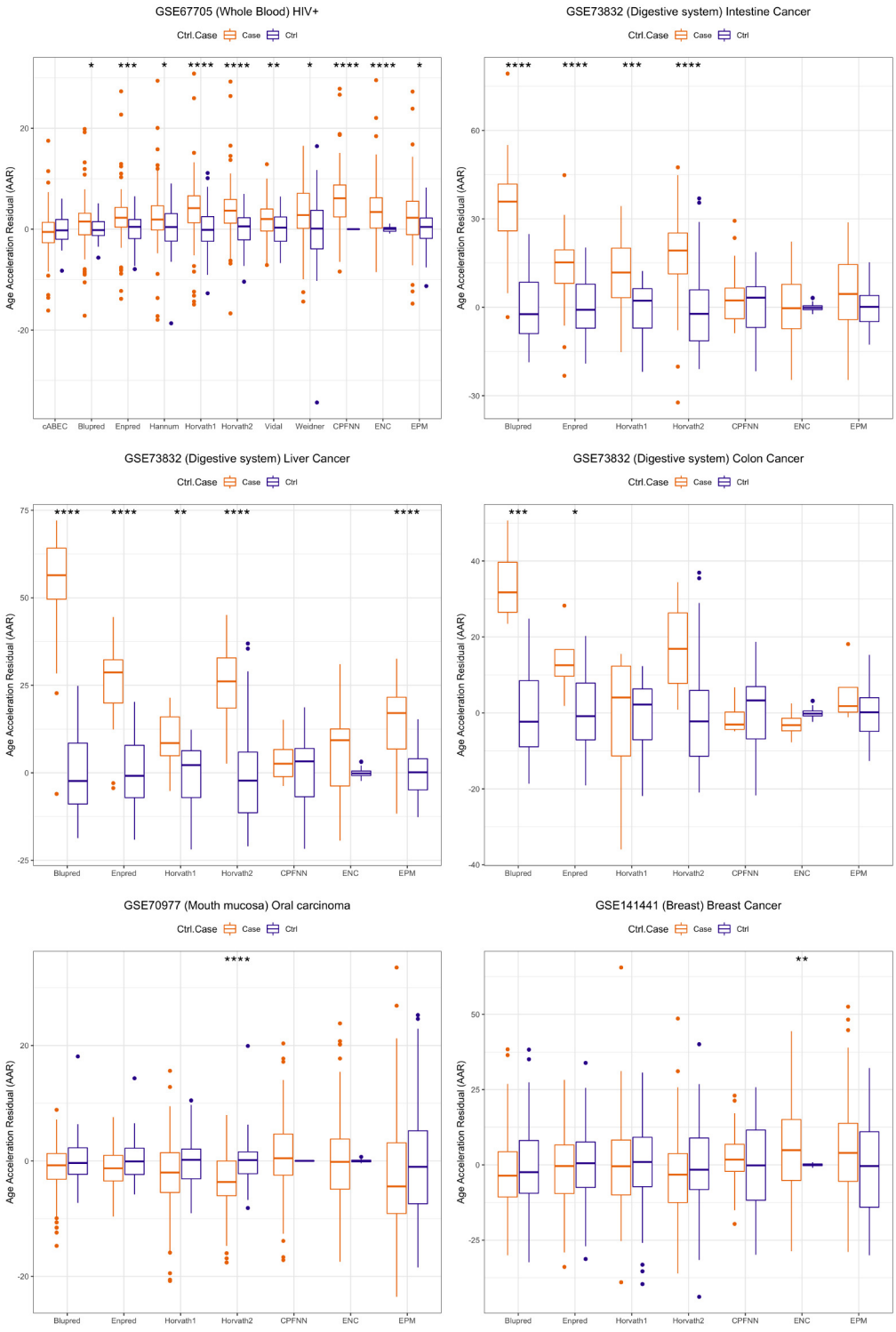

Figure 3: Age Acceleration Residual analysis of HIV+ and cancer patients. Statistically significantly different distribution of control vs case residuals: $**** \leq 0.0001, *** \leq 0.001, ** \leq 0.01, * \leq 0.05$

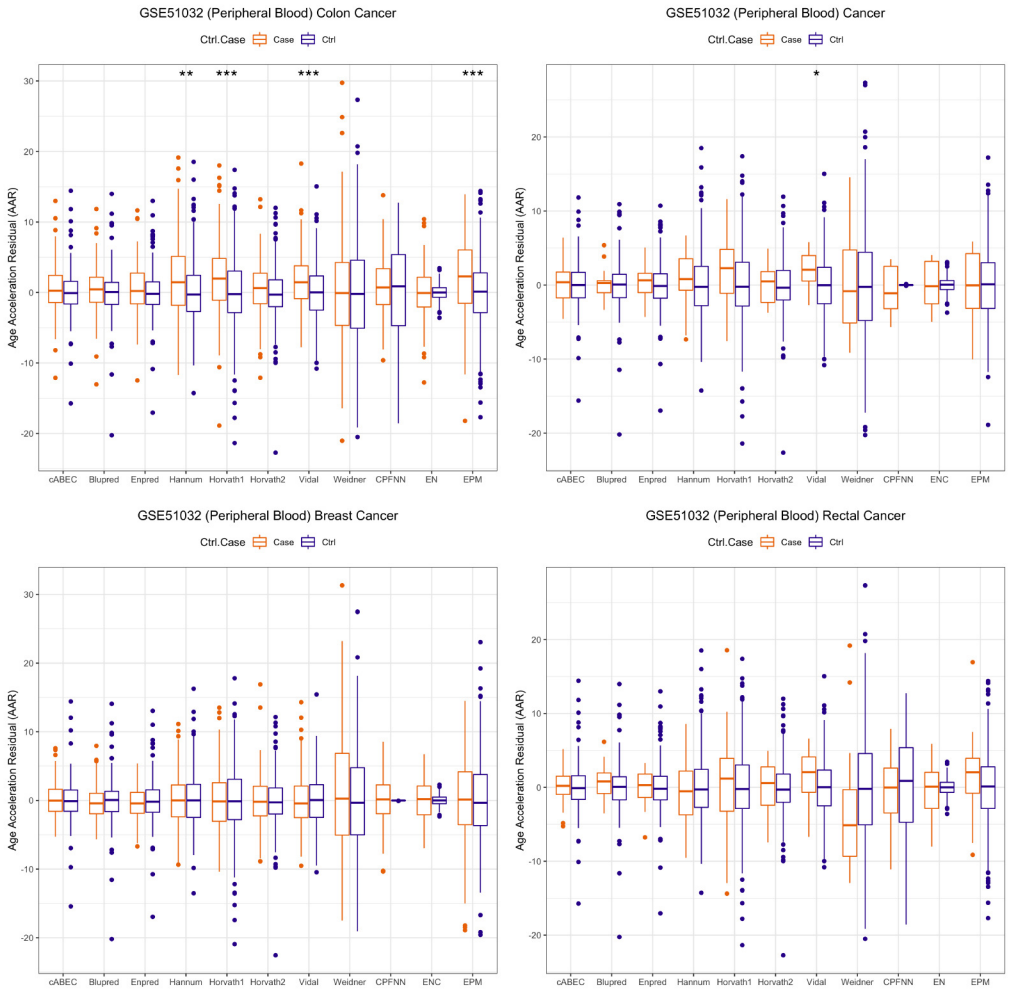

Figure 4: Age Acceleration Residual analysis for early prediction of disease progression: all controls where healthy at the time of blood sampling and developed some form of cancer over a follow-up period covering 15 years. Statistically significantly different distribution of control vs case residuals: $**** \leq 0.0001, *** \leq 0.001, ** \leq 0.01, * \leq 0.05$

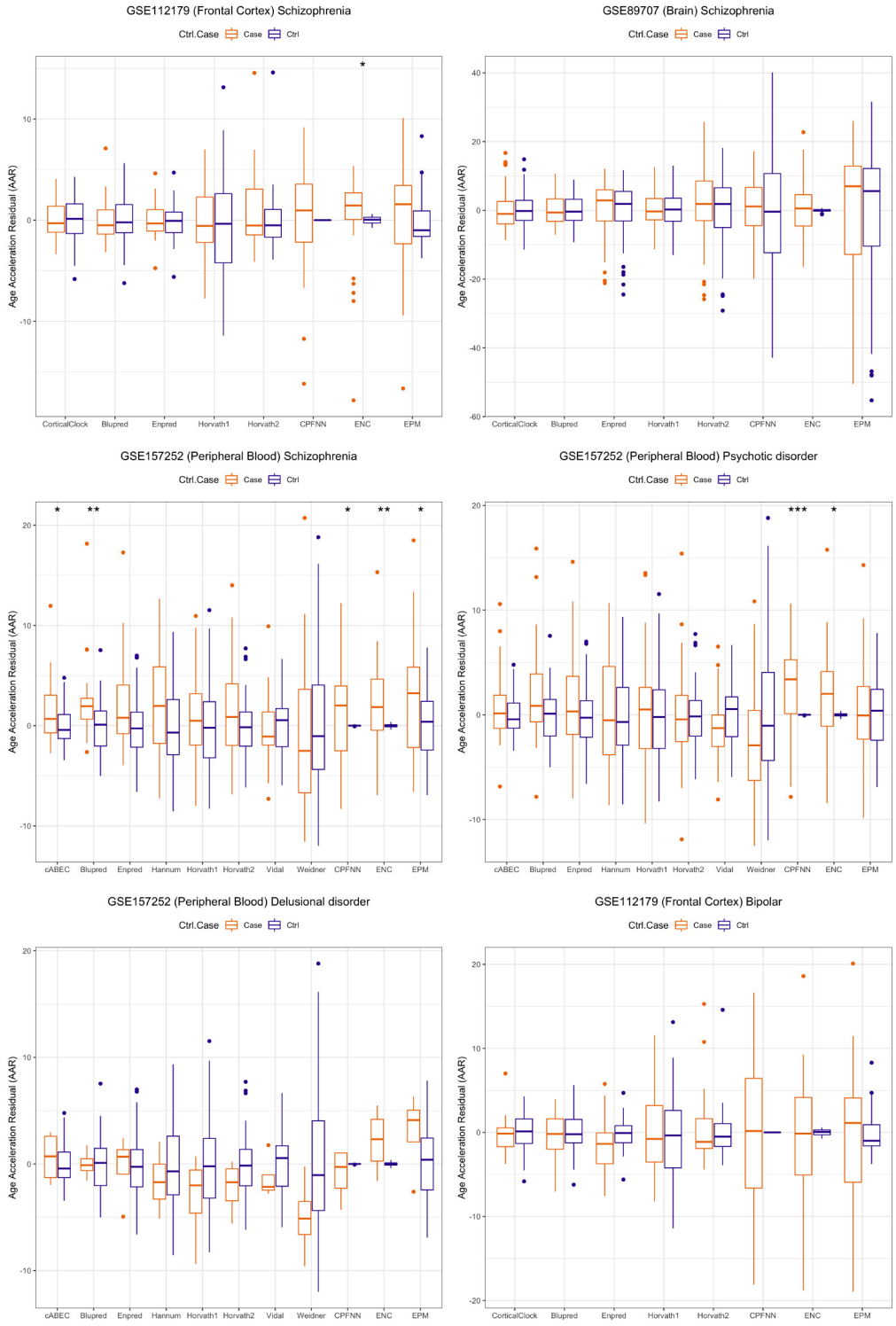

Figure 5: Age Acceleration Residual analysis. Statistically significantly different distribution of control vs case residuals: $**** \leq 0.0001, *** \leq 0.001, ** \leq 0.01, * \leq 0.05$

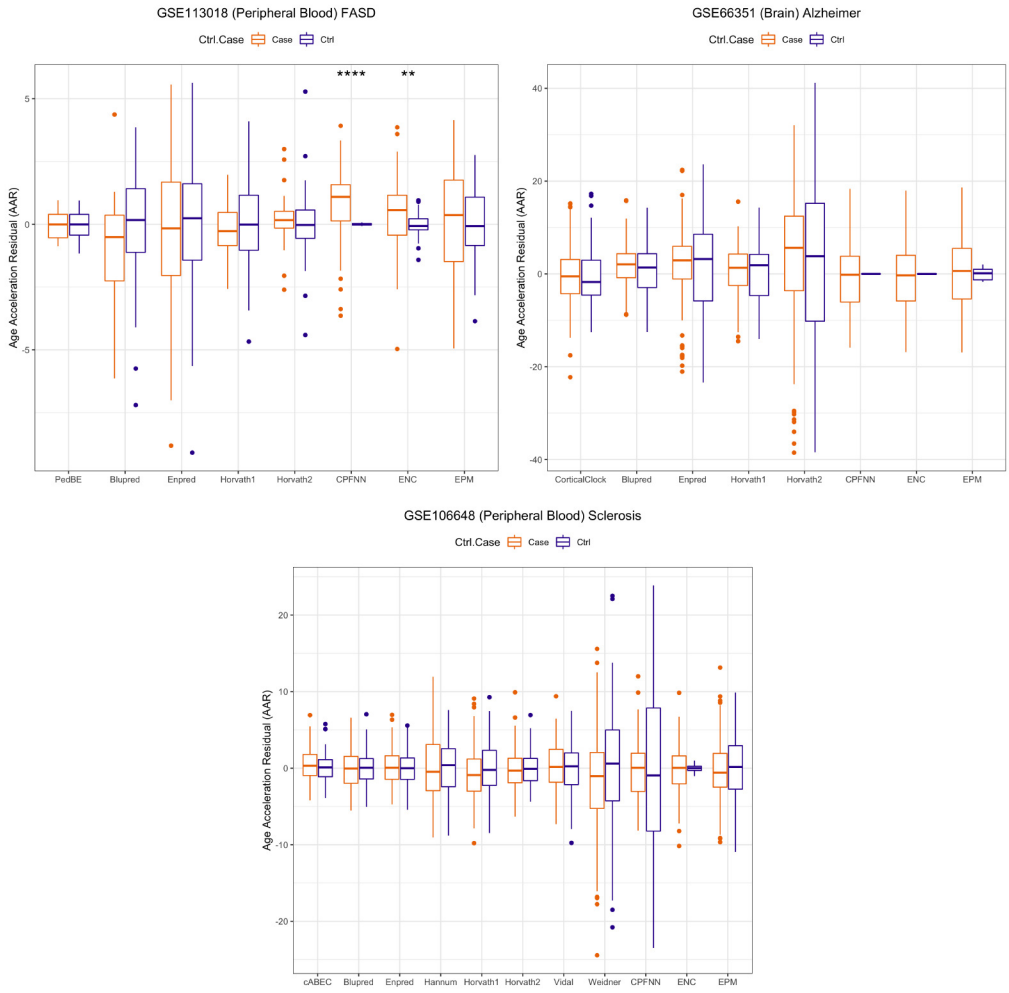

Figure 6: Age Acceleration Residual analysis. Statistically significantly different distribution of control vs case residuals: $**** \leq 0.0001, *** \leq 0.001, ** \leq 0.01, * \leq 0.05$

Conclusion

A number of observations can be drawn from this analysis.

The first is related to the variety of computational methods and techniques used to build DNAm-based clocks. In most cases, we do not find dramatically different performances, which ultimately seem to be more affected by the amount and quality of the training data than by the computational methods. From this perspective, a convenient guiding frame to discuss the results may be the tissue used for training, which defines the analysis at its earlier point in time (design of the case cohort). DNAm clocks can be classified in two broad classes: tissue-specific and multi-tissue. Our analysis brings

us to the conclusion that tissue-specific clocks and multi-tissue predictors have different performances with the former offering overall better results.

It seems relatively assessed that multi-tissue clocks are the most difficult to rely on, being ambiguous, as resulting from an unstandardized selection of tissues. On the contrary, tissue-specific clocks directly take into account the methylation landscape that define cells' specificity and tissues niches. However, not all tissues seem to reveal strong association signals between methylation levels and chronological age. In particular, age estimation accuracy of DNAm clocks appears to be relevantly higher on blood tissues in comparison to other tissues. This may be largely a consequence of the much larger availability of blood methylation data, which allows for larger training datasets and thus higher age estimation accuracy [80], but it may also be related to the biological heterogeneity of methylation patters affecting tissue-specific gene expression [99]. Furthermore, from a practical standpoint, invasive biopsies to assess the methylation status of internal organs, cannot be used as routine screening. At the opposite end lie blood-clocks, whose accessibility, are extremely simple and more cost-effective. It may therefore be interesting to deepen the current analysis with *ad hoc* designed studies to understand which is the trade-off between effective clock design (i.e. needed number of samples) and tissue specificity to grant comparable age accuracy, and finally probes numerosity for final costs assessment.

Finally, some considerations on age acceleration analysis are necessary. Standardization on definitions and implementations of age acceleration would be useful. Although we chose the most commonly used AAR metric, the results obtained are sometimes inconclusive in comparison to what is reported in literature. A second observation is that high age estimation accuracy seems to affect the DNAm clocks capability of detecting significant age acceleration/deceleration, as observed in [80]. In this perspective, our analysis may support the current research trend fostering the design of mortality clocks, where DNAm age plays a relevant but not unique role. This may overall indicate that while DNAm age relevance is important, its relevance in a clinical perspective must be integrated with additional parameters, as indeed the complexity of the empirical observation of the process of ageing intuitively suggests.

Competing interests

There is NO Competing Interest.

Key Points

- DNA methylation (DNAm) is considered one of the most promising biomarker of ageing
- We provide a survey on existing DNAm based epigenetic clocks
- We showed performance evaluation of several DNAm based clocks across different tissues and diseases

References

[1] Teresa Niccoli and Linda Partridge. Ageing as a risk factor for disease. *Current Biology*, 22(17):R741–R752, 2012.

[2] George T. Baker and Richard L. Sprott. Biomarkers of aging. *Experimental Gerontology*, 23(4-5):223–239, 1988.

[3] Alexander Vaiserman and Dmytro Krasnienkov. Telomere Length as a Marker of Biological Age: State-of-the-Art, Open Issues, and Future Perspectives. *Frontiers in Genetics*, 11, 2021.

[4] Xian Xia, Weiyang Chen, Joseph McDermott, and Jing-Dong Jackie Han. Molecular and phenotypic biomarkers of aging. *F1000Research*, 6:860, 2017.

[5] Meaghan J. Jones, Sarah J. Goodman, and Michael S. Kobor. DNA methylation and healthy human aging. *Aging Cell*, 14(6):924–932, 2015.

[6] A Razin and H Cedar. DNA methylation and gene expression. *Microbiological Reviews*, 55(3):451–458, 1991.

[7] En Li, Caroline Beard, and Rudolf Jaenisch. Role for DNA methylation in genomic imprinting. *Nature*, 366(6453):362–365, 1993.

[8] Julie Borgel, Sylvain Guibert, Yufeng Li, Hatsune Chiba, Dirk Schübeler, Hiroyuki Sasaki, Thierry Forné, and Michael Weber. Targets and dynamics of promoter DNA methylation during early mouse development. *Nature Genetics*, 42(12):1093–1100, 2010.

[9] Sven Bocklandt, Wen Lin, Mary E. Sehl, Francisco J. Sánchez, Janet S. Sinsheimer, Steve Horvath, and Eric Vilain. Epigenetic predictor of age. *PLoS ONE*, 6(6):e14821, 2011.

[10] 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 ofELOVL2gene as a new epigenetic marker of age. *Aging Cell*, 11(6):1132–1134, 2012.

[11] Steven Horvath. Dna methylation age of human tissues and cell types. *Genome Biology*, 14(3156):912–926, 2013.

[12] Xiaoqian Gao, Shuai Liu, Haoqiu Song, Xin Feng, Meiyu Duan, Lan Huang, and Fengfeng Zhou. AgeGuess, a methylomic prediction model for human ages. *Frontiers in Bioengneering and Biotechnology*, 8, 2020.

[13] Qian Zhang, Costanza L. Vallerga, Rosie M. Walker, Tian Lin, Anjali K. Henders, Grant W. Montgomery, Ji He, Dongsheng Fan, Javed Fowdar, Martin Kennedy, Toni Pitcher, John Pearson, Glenda Halliday, John B. Kwok, Ian Hickie, Simon Lewis, Tim Anderson, Peter A. Silburn, George D. Mellick, Sarah E. Harris, Paul Redmond, Alison D. Murray, David J. Porteous, Christopher S. Haley, Kathryn L. Evans, Andrew M. McIntosh, Jian Yang, Jacob Gratten, Riccardo E. Marioni, Naomi R. Wray, Ian J. Deary, Allan F. McRae, and Peter M. Visscher. Improved precision of epigenetic clock estimates across tissues and its implication for biological ageing. *Genome Medicine*, 11(1), 2019.

[14] J. Bohlin, S. E. Håberg, P. Magnus, S. E. Reese, H. K. Gjessing, M. C. Magnus, C. L. Parr, C. M. Page, S. J. London, and W. Nystad. Prediction of gestational age based on genome-wide differentially methylated regions. *Genome Biology*, 17(1), 2016.

[15] Mariana Boroni, Alessandra Zonari, Carolina Reis de Oliveira, Kallie Alkatib, Edgar Andres Ochoa Cruz, Lear E. Brace, and Juliana Lott de Carvalho. Highly accurate skin-specific methylome analysis algorithm as a platform to screen and validate therapeutics for healthy aging. *Clinical Epigenetics*, 12(1), 2020.

[16] Gemma L Shireby, Jonathan P Davies, Paul T Francis, Joe Burrage, Emma M Walker, Grant W A Neilson, Aisha Dahir, Alan J Thomas, Seth Love, Rebecca G Smith, Katie Lunnon, Meena Kumari, Leonard C Schalkwyk, Kevin Morgan, Keeley Brookes, Eilis Hannon, and Jonathan Mill. Recalibrating the epigenetic clock: implications for assessing biological age in the human cortex. *Brain*, 143(12):3763–3775, 2020.

[17] Yunsung Lee, Kristine L. Haftorn, William R. P. Denault, Haakon E. Nustad, Christian M. Page, Robert Lyle, Sindre Lee-Ødegård, Gunn-Helen Moen, Rashmi B. Prasad, Leif C. Groop, Line Sletner, Christine Sommer, Maria C. Magnus, Håkon K. Gjessing, Jennifer R. Harris, Per Magnus, Siri E. Håberg, Astanand Jugessur, and Jon Bohlin. Blood-based epigenetic estimators of chronological age in human adults using DNA methylation data from the illumina MethylationEPIC array. *BMC Genomics*, 21(1), 2020.

[18] Yunsung Lee, Sanaa Choufani, Rosanna Weksberg, Samantha L. Wilson, Victor Yuan, Amber Burt, Carmen Marsit, Ake T. Lu, Beate Ritz, Jon Bohlin, Håkon K. Gjessing, Jennifer R. Harris, Per Magnus, Alexandra M. Binder, Wendy P. Robinson, Astanand Jugessur, and Steve Horvath. Placental epigenetic clocks: estimating gestational age using placental DNA methylation levels. *Aging*, 11(12):4238–4253, 2019.

[19] Lechuan Li, Chonghao Zhang, Shiyu Liu, Hannah Guan, and Yu Zhang. Age prediction by DNA methylation in neural networks. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 2021. [published online ahead of print, 2021 May 28].

[20] Sagi Snir, Colin Farrell, and Matteo Pellegrini. Human epigenetic ageing is logarithmic with time across the entire lifespan. *Epigenetics*, 14(9):912–926, 2019.

[21] Gregory Hannum, Justin Guinney, Ling Zhao, Li Zhang, Guy Hughes, SriniVas Sadda, Brandy Klotzle, Marina Bibikova, Jian-Bing Fan, Yuan Gao, Rob Deconde, Menzies Chen, Indika Rajapakse, Stephen Friend, Trey Ideker, and Kang Zhang. Genome-wide methylation profiles reveal quantitative views of human aging rates. *Molecular Cell*, 49(2):359–367, 2013.

[22] 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.

[23] Anna K. Knight, Jeffrey M. Craig, Christiane Theda, Marie Bækvad-Hansen, Jonas Bybjerg-Grauholm, Christine S. Hansen, Mads V. Hollegaard, David M. Hougaard, Preben B. Mortensen, Shantel M. Weinsheimer, Thomas M. Werge, Patricia A. Brennan, Joseph F. Cubells, D. Jeffrey Newport, Zachary N. Stowe, Jeanie L. Y. Cheong, Philippa Dalach, Lex W. Doyle, Yuk J. Loke, Andrea A. Baccarelli, Allan C. Just, Robert O. Wright, Mara M. Téllez-Rojo, Katherine Svensson, Letizia Trevisi, Elizabeth M. Kennedy, Elisabeth B. Binder, Stella Iurato, Darina Czamara, Katri Räikkönen, Jari M. T. Lahti, Anu-Katriina Pesonen, Eero Kajantie, Pia M. Villa, Hannele Laivuori, Esa Hämäläinen, Hea Jin Park, Lynn B. Bailey, Sasha E. Parets, Varun Kilaru, Ramkumar Menon, Steve Horvath, Nicole R. Bush, Kaja Z. LeWinn, Frances A. Tylavsky, Karen N. Conneely, and Alicia K. Smith. An epigenetic clock for gestational age at birth based on blood methylation data. *Genome Biology*, 17(1), 2016.

[24] Jun Li, Xiaoyan Zhu, Kuai Yu, Haijing Jiang, Yizhi Zhang, Biqi Wang, Xuezhen Liu, Siyun Deng, Jie Hu, Qifei Deng, Huizhen Sun, Huan Guo, Xiaomin Zhang, Weihong Chen, Jing Yuan, Meian He, Yansen Bai, Xu Han, Bing Liu, Chuanyao Liu, Yanjun Guo, Bing Zhang, Zhihong Zhang, Frank B. Hu, Wenjing Gao, Liming Li, Mark Lathrop, Catherine Laprise, Liming Liang, and Tangchun Wu. Exposure to polycyclic aromatic hydrocarbons and accelerated DNA methylation aging. *Environmental Health Perspectives*, 126(6):067005, 2018.

[25] Xingyan Li, Weidong Li, and Yan Xu. Human age prediction based on DNA methylation using a gradient boosting regressor. *Genes*, 9(9):424, 2018.

[26] Chunxiao Li, Wenjing Gao, Ying Gao, Canqing Yu, Jun Lv, Ruoran Lv, Jiali Duan, Ying Sun, Xianghui Guo, Weihua Cao, and Liming Li. Age prediction of children and adolescents aged 6-17 years: an epigenome-wide analysis of DNA methylation. *Aging*, 10(5):1015–1026, 2018.

[27] Benjamin T Mayne, Shalem Y Leemaqz, Alicia K Smith, James Breen, Claire T Roberts, and Tina Bianco-Miotto. Accelerated placental aging in early onset preeclampsia pregnancies identified by DNA methylation. *Epigenomics*, 9(3):279–289, 2017.

[28] Sarah Voisin, Nicholas R. Harvey, Larisa M. Haupt, Lyn R. Griffiths, Kevin J. Ashton, Vernon G. Coffey, Thomas M. Doering, Jamie-Lee M. Thompson, Christian Benedict, Jonathan Cedernaes, Malene E. Lindholm, Jeffrey M. Craig, David S. Rowlands, Adam P. Sharples, Steve Horvath, and Nir Eynon. An epigenetic clock for human skeletal muscle. *Journal of Cachexia, Sarcopenia and Muscle*, 11(4):887–898, 2020.

[29] Lisa M. McEwen, Kieran J. O'Donnell, Megan G. McGill, Rachel D. Edgar, Meaghan J. Jones, Julia L. MacIsaac, David Tse Shen Lin, Katia Ramadori, Alexander Morin, Nicole Gladish, Elika Garg, Eva Unternaehrer, Irina Pokhvisneva, Neerja Karnani, Michelle Z. L. Kee, Torsten Klengel, Nancy E. Adler, Ronald G. Barr, Nicole Letourneau, Gerald F. Giesbrecht, James N. Reynolds, Darina Czamara, Jeffrey M. Armstrong, Marilyn J. Essex, Carolina de Weerth, Roseriet Beijers, Marieke S. Tollenaar, Bekh Bradley, Tanja Jovanovic, Kerry J. Ressler, Meir Steiner, Sonja Entringer, Pathik D. Wadhwa, Claudia Buss, Nicole R. Bush, Elisabeth B. Binder, W. Thomas Boyce, Michael J. Meaney, Steve Horvath, and Michael S. Kobor. The pedbe clock accurately estimates dna methylation age in pediatric buccal cells. *Proceedings of the National Academy of Sciences*, 117(38):23329–23335, 2020.

[30] Laura Vidal-Bralo, Yolanda Lopez-Golan, and Antonio Gonzalez. Simplified assay for epigenetic age estimation in whole blood of adults. *Frontiers in Genetics*, 7:126, 2016.

[31] Carola Weidner, Qiong Lin, Carmen Koch, Lewin Eisele, Fabian Beier, Patrick Ziegler, Dirk Bauerschlag, Karl-Heinz Jöckel, Raimund Erbel, Thomas Mühleisen, Martin Zenke, Tim Brümmendorf, and Wolfgang Wagner. Aging of blood can be tracked by DNA methylation changes at just three CpG sites. *Genome Biology*, 15(2):R24, 2014.

[32] Xiaohui Wu, Weidan Chen, Fangqin Lin, Qingsheng Huang, Jiayong Zhong, Huan Gao, Yanyan Song, and Huiying Liang. DNA methylation profile is a quantitative measure of biological aging in children. *Aging*, 11(22):10031–10051, 2019.

[33] Yan Xu, Xingyan Li, Yingxi Yang, Chunhui Li, and Xiaojian Shao. Human age prediction based on DNA methylation of non-blood tissues. *Computer Methods and Programs in Biomedicine*, 171:11–18, 2019.

[34] Pietro Di Lena, Claudia Sala, and Christine Nardini. Estimage: a webserver hub for the computation of methylation age. *Nucleic Acids Research*, 49(W1):W199–W206, 2021.

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

[36] Lara Oblak, Jeroen van der Zaag, Albert T. Higgins-Chen, Morgan E. Levine, and Marco P. Boks. A systematic review of biological, social and environmental factors associated with epigenetic clock acceleration. *Ageing Research Reviews*, 69:101348, 2021.

[37] Daniel J. Simpson and Tamir Chandra. Epigenetic age prediction. *Aging Cell*, 20(9), 2021.

[38] Daniel W Belsky, Avshalom Caspi, Louise Arseneault, Andrea Baccarelli, David L Corcoran, Xu Gao, Eiliss Hannon, Hona Lee Harrington, Line JH Rasmussen, Renate Houts, Kim Huffman, William E Kraus, Dayoon Kwon, Jonathan Mill, Carl F Pieper, Joseph A Prinz, Richie Poulton, Joel Schwartz, Karen Sugden, Pantel Vokonas, Benjamin S Williams, and Terrie E Moffitt. Quantification of the pace of biological aging in humans through a blood test, the DunedinPoAm DNA methylation algorithm. *eLife*, 9, 2020.

[39] 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.

[40] 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.

[41] Yan Zhang, Rory Wilson, Jonathan Heiss, Lutz P. Breitling, Kai-Uwe Saum, Ben Schöttker, Bernd Holleczek, Melanie Waldenberger, Annette Peters, and Hermann Brenner. DNA methylation signatures in peripheral blood strongly predict all-cause mortality. *Nature Communications*, 8(1), 2017.

[42] H. Eka D. Suchiman, Roderick C. Slieker, Dennis Kremer, P. Eline Slagboom, Bastiaan T. Heijmans, and Elmar W. Tobi. Design, measurement and processing of region-specific DNA methylation assays: the mass spectrometry-based method EpiTYPER. *Front Genet*, 6:287, 2015.

[43] 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.

[44] Ieva Rauluseviciute, Finn Drabløs, and Morten Beck Rye. DNA methylation data by sequencing: experimental approaches and recommendations for tools and pipelines for data analysis. *Clinical Epigenetics*, 11(1):193, 2019.

[45] The BLUPRINT consortium. Quantitative comparison of DNA methylation assays for biomarker development and clinical applications. *Nature Biotechnology*, 34(7):726–737, 2016.

[46] Zhenxing Wang, XiaoLiang Wu, and Yadong Wang. A framework for analyzing DNA methylation data from illumina infinium HumanMethylation450 BeadChip. *BMC Bioinformatics*, 19(S5), 2018.

[47] 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.

[48] Marie Forest, Kieran J. ODonnell, Greg Voisin, Helene Gaudreau, Julia L. MacIsaac, Lisa M. McEwen, Patricia P. Silveira, Meir Steiner, Michael S. Kobor, Michael J. Meaney, and Celia M.T. Greenwood. Agreement in DNA methylation levels from the illumina 450k array across batches, tissues, and time. *Epigenetics*, 13(1):19–32, 2018.

[49] Teresia Kling, Anna Wenger, Stephan Beck, and Helena Carén. Validation of the MethylationEPIC BeadChip for fresh-frozen and formalin-fixed paraffin-embedded tumours. *Clinical Epigenetics*, 9(1), 2017.

[50] 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.

[51] Ruth Pidsley, Chloe C Y Wong, Manuela Volta, Katie Lunnon, Jonathan Mill, and Leonard C Schalkwyk. A data-driven approach to preprocessing illumina 450k methylation array data. *BMC Genomics*, 14(1), 2013.

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

[53] Ting Wang, Weihua Guan, Jerome Lin, Nadia Boutaoui, Glorisa Canino, Jianhua Luo, Juan Carlos Celedón, and Wei Chen. A systematic study of normalization methods for infinium 450k methylation data using whole-genome bisulfite sequencing data. *Epigenetics*, 10(7):662–669, 2015.

[54] 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. *British Journal of Cancer*, 109(6):1394–1402, 2013.

[55] Claudia Sala, Pietro Di Lena, Danielle Fernandes Durso, Andrea Prodi, Gastone Castellani, and Christine Nardini. Evaluation of pre-processing on the meta-analysis of DNA methylation data from the illumina HumanMethylation450 BeadChip platform. *PLOS ONE*, 15(3):e0229763, 2020.

[56] W. Evan Johnson, Cheng Li, and Ariel Rabinovic. Adjusting batch effects in microarray expression data using empirical bayes methods. *Biostatistics*, 8(1):118–127, 2006.

[57] 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.

[58] E Magda Price, Allison M Cotton, Lucia L Lam, Pau Farré, Eldon Emberly, Carolyn J Brown, Wendy P Robinson, and Michael S Kobor. Additional annotation enhances potential for biologically-relevant analysis of the illumina infinium HumanMethylation450 BeadChip array. *Epigenetics & Chromatin*, 6(1), 2013.

[59] Shuxia Li, Jesper B. Lund, Kaare Christensen, Jan Baumbach, Jonas Mengel-From, Torben Kruse, Weilong Li, Afsaneh Mohammadnejad, Alison Pattie, Riccardo E. Marioni, Ian J. Deary, and Qihua Tan. Exploratory analysis of age and sex dependent DNA methylation patterns on the x-chromosome in whole blood samples. *Genome Medicine*, 12(1), 2020.

[60] Rajbir Nath Batra, Aviezer Lifshitz, Ana Tufegdzic Vidakovic, Suet-Feung Chin, Ankita Sati-Batra, Stephen-John Sammut, Elena Provenzano, H. Raza Ali, Ali Dariush, Alejandra Bruna, Leigh Murphy, Arnie Purushotham, Ian Ellis, Andrew Green, Francine E. Garrett-Bakelman, Chris Mason, Ari Melnick, Samuel A. J. R. Aparicio, Oscar M. Rueda, Amos Tanay, and Carlos Caldas. DNA methylation landscapes of 1538 breast cancers reveal a replication-linked clock, epigenomic instability and cis-regulation. *Nature Communications*, 12(1), 2021.

[61] Zhifu Sun, High Seng Chai, Yanhong Wu, Wendy M White, Krishna V Donkena, Christopher J Klein, Vesna D Garovic, Terry M Therneau, and Jean-Pierre A Kocher. Batch effect correction for genome-wide methylation data with illumina infinium platform. *BMC Medical Genomics*, 4(1), 2011.

[62] Andrew E. Teschendorff, Francesco Marabita, Matthias Lechner, Thomas Bartlett, Jesper Tegner, David Gomez-Cabrero, and Stephan Beck. A beta-mixture quantile normalization method for correcting probe design bias in illumina infinium 450 k DNA methylation data. *Bioinformatics*, 29(2):189–196, 2012.

[63] David N. Reshef, Yakir A. Reshef, Hilary K. Finucane, Sharon R. Grossman, Gilean McVean, Peter J. Turnbaugh, Eric S. Lander, Michael Mitzenmacher, and Pardis C. Sabeti. Detecting novel associations in large data sets. *Science*, 334(6062):1518–1524, 2011.

[64] Pietro Di Lena, Claudia Sala, Andrea Prodi, and Christine Nardini. Missing value estimation methods for DNA methylation data. *Bioinformatics*, 35(19):3786–3793, 2019.

[65] Victor J. Yohai, Werner A. Stahel, and Ruben H. Zamar. A procedure for robust estimation and inference in linear regression. In *Directions in Robust Statistics and Diagnostics*, pages 365–374. Springer New York, 1991.

[66] Diego Franco Saldana and Yang Feng. SIS: An r package for sure independence screening in ultrahigh-dimensional statistical models. *Journal of Statistical Software*, 83(2), 2018.

[67] Sagi Snir, Colin Farrell, and Matteo Pellegrini. Human epigenetic ageing is logarithmic with time across the entire lifespan. *Epigenetics*, 14(9):912–926, 2019. PMID: 31138013.

[68] Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 67(2):301–320, 2005.

[69] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1):1–22, 2010.

[70] Isabelle Guyon, Jason Weston, Stephen Barnhill, and Vladimir Vapnik. Gene selection for cancer classification using support vector machines. *Machine Learning*, 46(1/3):389–422, 2002.

[71] Xin Feng, Ruochi Zhang, Minge Liu, Quewang Liu, Fei Li, Zhenwei Yan, and Fengfeng Zhou. An accurate regression of developmental stages for breast cancer based on transcriptomic biomarkers. *Biomarkers in Medicine*, 13(1):5–15, 2019.

[72] Jerome H. Friedman. Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, 29(5), 2001.

[73] Jeffrey B. Endelman. Ridge regression and other kernels for genomic selection with r package rrBLUP. *The Plant Genome*, 4(3):250–255, 2011.

[74] Sagi Snir, Yuri I. Wolf, and Eugene V. Koonin. Universal pacemaker of genome evolution. *PLoS Computational Biology*, 8(11):e1002785, 2012.

[75] Bing Cheng and D. M. Titterington. Neural Networks: A Review from a Statistical Perspective. *Statistical Science*, 9(1):2 – 30, 1994.

[76] R. Edgar. Gene expression omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Research*, 30(1):207–210, 2002.

[77] Pietro Di Lena, Claudia Sala, Andrea Prodi, and Christine Nardini. Methylation data imputation performances under different representations and missingness patterns. *BMC Bioinformatics*, 21(1), 2020.

[78] 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, André G. Uitterlinden, Devin Absher, Themistocles Assimes, Morgan E. Levine, Ake T. Lu, Philip S. Tsao, Lifang Hou, JoAnn 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.

[79] Colin Farrell, Sagi Snir, and Matteo Pellegrini. The epigenetic pacemaker: modeling epigenetic states under an evolutionary framework. *Bioinformatics*, 36(17):4662–4663, 2020.

[80] Qian Zhang, Costanza L. Vallerga, Rosie M. Walker, Tian Lin, Anjali K. Henders, Grant W. Montgomery, Ji He, Dongsheng Fan, Javed Fowdar, Martin Kennedy, Toni Pitcher, John Pearson, Glenda Halliday, John B. Kwok, Ian Hickie, Simon Lewis, Tim Anderson, Peter A. Silburn, George D. Mellick, Sarah E. Harris, Paul Redmond, Alison D. Murray, David J. Porteous, Christopher S. Haley, Kathryn L. Evans, Andrew M. McIntosh, Jian Yang, Jacob Gratten, Riccardo E. Marioni, Naomi R. Wray, Ian J. Deary, Allan F. McRae, and Peter M. Visscher. Improved precision of epigenetic clock estimates across tissues and its implication for biological ageing. *Genome Medicine*, 11(1):54, 2019.

[81] Christopher G. Bell, Robert Lowe, Peter D. Adams, Andrea A. Baccarelli, Stephan Beck, Jordana T. Bell, Brock C. Christensen, Vadim N. Gladyshev, Bastiaan T. Heijmans, Steve Horvath, Trey Ideker, Jean-Pierre J. Issa, Karl T. Kelsey, Riccardo E. Marioni, Wolf Reik, Caroline L. Relton, Leonard C. Schalkwyk, Andrew E. Teschendorff, Wolfgang Wagner, Kang Zhang, and Vardhman K. Rakyan. DNA methylation aging clocks: challenges and recommendations. *Genome Biology*, 20(1), 2019.

[82] Tessa Bergsma and Ekaterina Rogaeva. DNA methylation clocks and their predictive capacity for aging phenotypes and healthspan. *Neuroscience Insights*, 15:263310552094222, 2020.

[83] Andrés Esteban-Cantos, Javier Rodríguez-Centeno, Pilar Barruz, Belén Alejos, Gabriel Saiz-Medrano, Julián Nevado, Artur Martin, Francisco Gayá, Rosa De Miguel, Jose I Bernardino, Rocío Montejano, Beatriz Mena-Garay, Julen Cadiñanos, Eric Florence, Fiona Mulcahy, Denes Banhegyi, Andrea Antinori, Anton Pozniak, Cédrick Wallet, François Raffi, Berta Rodés, and Jose R Arribas. Epigenetic age acceleration changes 2 years after antiretroviral therapy initiation in adults with HIV: a substudy of the NEAT001/ANRS143 randomised trial. *The Lancet HIV*, 8(4):e197–e205, 2021.

[84] Steve Horvath and Andrew J. Levine. HIV-1 infection accelerates age according to the epigenetic clock. *Journal of Infectious Diseases*, 212(10):1563–1573, 2015.

[85] Tammy M. Rickabaugh, Ruth M. Baxter, Mary Sehl, Janet S. Sinsheimer, Patricia M. Hultin, Lance E. Hultin, Austin Quach, Otoniel Martínez-Maza, Steve Horvath, Eric Vilain, and Beth D. Jamieson. Acceleration of age-associated methylation patterns in HIV-1-infected adults. *PLOS ONE*, 10(3):e0119201, 2015.

[86] Pierre-Antoine Dugué, Julie K. Bassett, JiHoon E. Joo, Chol-Hee Jung, Ee Ming Wong, Margarita Moreno-Betancur, Daniel Schmidt, Enes Makalic, Shuai Li, Gianluca Severi, Allison M. Hodge, Daniel D. Buchanan, Dallas R. English, John L. Hopper, Melissa C. Southey, Graham G. Giles, and Roger L. Milne. DNA methylation-based biological aging and cancer risk and survival: Pooled analysis of seven prospective studies. *International Journal of Cancer*, 142(8):1611–1619, 2017.

[87] Danielle Fernandes Durso, Maria Giulia Bacalini, Claudia Sala, Chiara Pirazzini, Elena Marasco, Massimiliano Bonafé, Ítalo Faria do Valle, Davide Gentilini, Gastone Castellani, Ana Maria Caetano Faria, Claudio Franceschi, Paolo Garagnani, and Christine Nardini. Acceleration of leukocytes' epigenetic age as an early tumor and sex-specific marker of breast and colorectal cancer. *Oncotarget*, 8(14):23237–23245, 2017.

[88] 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*, 111(10):1051–1058, 2019.

[89] Canhua Xiao, Jonathan J Beitler, Gang Peng, Morgan E Levine, Karen N Conneely, Hongyu Zhao, Jennifer C Felger, Evanthia C Wommack, Cynthia E Chico, Sangchoon Jeon, Kristin A Higgins, Dong M Shin, Nabil F Saba, Barbara A Burtness, Deborah W Bruner, and Andrew H Miller. Epigenetic age acceleration, fatigue, and inflammation in patients undergoing radiation therapy for head and neck cancer: A longitudinal study. *Cancer*, 127(18):3361–3371, 2021.

[90] Pierre Hainaut, Béatrice Vozar, Sabina Rinaldi, Elio Riboli, and Elodie Caboux. The european prospective investigation into cancer and nutrition biobank. In *Methods in Molecular Biology*, pages 179–191. Humana Press, 2010.

[91] B. Kirkpatrick, E. Messias, P. D. Harvey, E. Fernandez-Egea, and C. R. Bowie. Is schizophrenia a syndrome of accelerated aging? *Schizophrenia Bulletin*, 34(6):1024–1032, 2008.

[92] Xiaohui Wu, Junping Ye, Zhongju Wang, and Cunyou Zhao. Epigenetic age acceleration was delayed in schizophrenia. *Schizophrenia Bulletin*, 47(3):803–811, 2020.

[93] Oluwagbenga Dada, Christopher Adanty, Nasia Dai, Richie Jeremian, Sauliha Alli, Philip Gerretsen, Ariel Graff, John Strauss, and Vincenzo De Luca. Biological aging in schizophrenia and psychosis severity: DNA methylation analysis. *Psychiatry Res.*, 296(113646):113646, 2021.

[94] Gabriel R Fries, Isabelle E Bauer, Giselli Scaini, Samira S Valvassori, Consuelo Walss-Bass, Jair C Soares, and Joao Quevedo. Accelerated hippocampal biological aging in bipolar disorder. *Bipolar Disord.*, 22(5):498–507, 2020.

[95] Satoshi Okazaki, Ikuo Otsuka, Yutaka Shinko, Tadasu Horai, Takashi Hirata, Naruhisa Yamaki, Ichiro Sora, and Akitoyo Hishimoto. Epigenetic clock analysis in children with fetal alcohol spectrum disorder. *Alcoholism: Clinical and Experimental Research*, 45(2):329–337, 2021.

[96] Morgan E Levine, Ake T Lu, David A Bennett, and Steve Horvath. Epigenetic age of the pre-frontal cortex is associated with neuritic plaques, amyloid load, and alzheimer's disease related cognitive functioning. *Aging*, 7(12):1198–1211, 2015.

[97] Ake T. Lu, Eilis Hannon, Morgan E. Levine, Eileen M. Crimmins, Katie Lunnon, Jonathan Mill, Daniel H. Geschwind, and Steve Horvath. Genetic architecture of epigenetic and neuronal ageing rates in human brain regions. *Nature Communications*, 8(1), 2017.

[98] Eleftheria Theodoropoulou, Lars Alfredsson, Fredrik Piehl, Francesco Marabita, and Maja Jagodic. Different epigenetic clocks reflect distinct pathophysiological features of multiple sclerosis. *Epigenomics*, 11(12):1429–1439, 2019.

[99] Puja Sinha, Kiran Singh, and Manisha Sachan. Heterogeneous pattern of DNA methylation in developmentally important genes correlates with its chromatin conformation. *BMC Molecular Biology*, 18(1), 2017.