

Ensaya: An ensemble age model for prediction of chronological age in adolescents and young adults

Kristina Stenløkk^{a,*}, Pål Skage Dahlberg^a, Veslemøy Rolseth^a, Herman Bernitz^b, Mohamed Yasin Gamiieldien^c, Mona Nahas Gholmieh^d, Liam Robinson^c, Antoine Saadé^e, Andre Uys^f, Øyvind Bleka^a

^a Department of Forensic Sciences, Oslo University Hospital, Norway

^b Department of Oral and Maxillofacial Pathology, School of Dentistry, Faculty of Health Sciences, University of Pretoria, Pretoria, South Africa

^c Department of Maxillo-Facial and Oral Surgery, School of Dentistry, Faculty of Health Sciences, University of Pretoria, Pretoria, South Africa

^d Department of Pediatric and Community Dentistry, Faculty of Dental Medicine, Lebanese University, Beirut, Lebanon

^e Department of Orthodontics and Dentofacial Orthopedics, Faculty of Dental Medicine, Lebanese University, Beirut, Lebanon

^f Department of Anatomy, School of Medicine, Faculty of Health Sciences, University of Pretoria, Hatfield, South Africa

ARTICLE INFO

Keywords:

DNA methylation

Epigenetic clock

Age prediction

Blood-based biomarkers

Forensic epigenetics

ABSTRACT

DNA methylation-based models for chronological age estimation have, for over a decade, been among the most important methods for age prediction in humans. While large-scale data have enabled the development of numerous models, their accuracy varies substantially depending on choice of machine learning algorithm, pre-processing pipeline, age range, tissue type, and population characteristics. Accurate age estimation in adolescents and young adults is particularly relevant in legal, investigational, and biomedical contexts. Here, we systematically evaluated blood-based DNA methylation age predictors in individuals aged 12–25 years and developed Ensaya (Ensemble age model for prediction of chronological age in adolescents and young adults), an ensemble model optimized for this age range. We assembled a validation dataset of 687 whole blood or peripheral blood mononuclear cell samples from 12 cohorts, profiled using Illumina 450 K, EPIC, and EPIC v2 arrays. We compared existing clocks and assessed robustness across cohorts, platforms, normalisation procedures, technical replicates, sex, and circadian timing. Among individual models, PAYA, cAge, and Garma showed the strongest performance and were combined into the Ensaya ensemble. In validation, Ensaya achieved improved accuracy, with a median absolute deviation of 0.79 years and a mean absolute error of 1.04 years, along with reduced age-dependent bias. It also demonstrated high technical reproducibility and stable performance across platforms and normalisation methods. These findings show that ensemble-based modelling improves age estimation in adolescents and young adults, supporting Ensaya as a practical blood-based predictor.

1. Introduction

The introduction of DNA methylation-based age estimation models over a decade ago, most notably the epigenetic clocks developed by Hannum et al. [1] and Horvath et al. [2], both in 2013, marked a major advance in understanding how chronological time, biological phenomena, and environmental exposures leave measurable molecular signatures on the genome [3]. The widespread adoption of Illumina's Infinium methylation arrays by the research community enabled cost-effective epigenome-wide profiling and facilitated cross-study dataset integration, thereby creating larger sample sizes [4]. Although early epigenetic

clocks were trained to predict chronological age, newer models use outcomes such as time-to-death, biological traits, and health-related measures [5–7]. In forensic science, chronological age estimation is essential when the age of an individual involved in legal or investigative proceedings is unknown and must be inferred from biological material. This is particularly important around the legal threshold of 18 years, which in many jurisdictions marks the transition from minor to adult status. Crossing this threshold is associated with substantial changes in legal rights, responsibilities, and judicial treatment. Hence, optimizing prediction models for age ranges spanning the late teens is of the essence. Previous studies [8,9] have demonstrated that age-specific DNA

* Correspondence to: Department of Forensic Sciences, Oslo University Hospital, P.O. Box 4950, Oslo 0424, Norway.

E-mail address: krsest@ous-hf.no (K. Stenløkk).

<https://doi.org/10.1016/j.fsigen.2026.103579>

Received 5 May 2026; Received in revised form 10 June 2026; Accepted 29 June 2026

Available online 30 June 2026

1872-4973/© 2026 The Authors. Published by Elsevier B.V. This is an open access article under the CC BY license (<http://creativecommons.org/licenses/by/4.0/>).

methylation clocks can achieve improved prediction accuracy within their intended age ranges compared with more general models trained across broader age spans.

The pursuit of greater model performance in epigenetic models for chronological age has followed several paths over the years. The large scale of DNA methylation array data, with the latest generation, the EPICv2 array, exploring more than 930 000 sites [10], offers opportunities but also poses analytical challenges, such as the imbalance between sample size and dimensionality. Efforts to improve model performance have included increasing sample size and age range [11], exploring a wide range of approaches including deep learning models, principal component models, single cell clocks, preselection of CpG sites, restricting training to specific age intervals and tissues (recently reviewed by Teschendorff and Horvath [12]) and combining multiple prediction models through ensemble approaches [13].

Only a limited number of epigenetic models have been optimized to predict chronological age in adolescents and young adults. PAYA [8] was the first blood-based clock specifically designed for this age range. Although the paediatric clocks PedBE (buccal cells) [9] and Wu (blood) [14] were primarily developed for younger populations, their training age ranges only partially overlap with that of PAYA. More recently, models such as cAge [15] and Garma [16] have adopted a staged approach in which a general age-prediction model is applied first, followed by a refined age-specific prediction model. In addition, both models improve their elastic net predictors by performing a preselection of CpG sites identified through an epigenome-wide association study (EWAS).

In this study, we systematically evaluate the performance of openly available chronological age estimation models for blood samples in adolescents and young adults (12–25 years). Further, we show how an ensemble of three recently developed epigenetic age models improves prediction accuracy. Also, the robustness of the models was investigated with respect to the effects of different normalisation procedures, array versions, sex and circadian sample timing. The test set consisted of 687 individuals from both publicly available data sources and in-house-generated data from Lebanon and South Africa.

2. Materials and methods

2.1. Study participants and data collection

Two cohorts aged 12–25 years were recruited for this study: one in Beirut, Lebanon ($n = 142$), and one in Pretoria, South Africa ($n = 102$). The aim was to recruit a population with normal health and an even age distribution between 12 and 25 years. Exclusion criteria were haemophilia, endocrine disorders affecting growth and severe illnesses such as cancer. Blood samples were collected from the arm vein in EDTA tubes (Lebanon) or using Sarstedt 200 μ l Microvette® tubes, with a small amount collected from a fingertip puncture (South Africa).

Genomic DNA was extracted from blood using chemagic DNA Blood Kit H96 kits (PerkinElmer) on the chemagic Magnetic Separation Module I according to the manufacturer's protocols. DNA was eluted in TE–4 buffer. Extracted DNA was bisulphite converted using EZ–96 DNA Methylation kits (Zymo Research) and analysed by the Infinium MethylationEPIC v2.0 (EPIC2) array at Life and Brain GmbH in Bonn, Germany. 131 frozen blood samples from Lebanon were thawed in separate aliquots at two timepoints, and DNA was independently extracted and analysed on each occasion, generating replicate measurements for assessment of reproducibility.

We also searched the Gene Expression Omnibus (GEO) for publicly available DNA methylation datasets. We restricted our search to whole blood or PBMC (Peripheral Blood Mononuclear Cell) samples from either the Illumina Infinium HumanMethylation450 BeadChip (450k), HumanMethylationEPIC BeadChip (EPIC) or HumanMethylationEPICv2 BeadChip (EPIC2) platforms. Only datasets providing raw intensity data (IDAT files) between 12 and 25 years were included in the performance

analysis. The full list of datasets is available in the [Supplementary Materials \(Table S1\)](#), along with an overview of the age span per cohort ([Figure S1](#)). For studies reporting age in whole years, 0.5 years was added to each value to approximate the midpoint of the reported interval. Methylation datasets were obtained from the Gene Expression Omnibus (GEO) under accession numbers GSE105018, GSE219037, GSE197675, GSE224363, GSE158063, GSE197676, GSE115278, GSE118144, GSE164056, GSE227815, GSE249113, GSE168739, GSE217633, GSE124366, GSE157252, GSE59065, GSE142512, GSE234461, GSE227809, GSE247193, GSE247195, GSE247197, GSE83944, and the dataset from Ämmälä et al. from 10.17605/OSF.IO/QX7W3.

2.1.1. Ethical considerations

All methods in the present study were carried out in accordance with the relevant guidelines and regulations. The study "DNA methylation as a tool for estimation of age in adolescents" performed in collaboration with colleagues in Lebanon and South Africa was approved by the regional committee for medical research ethics south east Norway (REK sør-øst approval 2019/162), the Faculty of Health Sciences Research Ethics Committee, University of Pretoria (Ethics Reference No.: 608/2019) and The Institutional Review Board, Lebanese Hospital Geitaoui-University Medical Center (Code: 2019-IRB–017).

Informed consent was signed by a parent/guardian if a study participant was below the age of 16 years.

2.2. Data analysis

2.2.1. Pre-processing and quality control

Raw IDAT files were imported and processed using the minfi package (v1.52.1) [17]. Background correction and dye-bias normalisation were performed with the preprocessNoob function.

To ensure robust downstream analyses, data were subjected to multi-criteria quality control using the doQC function from the normtools R-package (v0.7.0, <https://github.com/oyvble/normtools>). Quality control was applied independently within each cohort using default thresholds. Samples flagged by at least two distinct quality control criteria were excluded from further analyses. The pipeline integrates multiple complementary quality metrics, including control probe performance (minfi [17]), bisulfite conversion efficiency assessed using control probe- and out-of-band-based approaches (watermelon [18], SeSAME [19]), detection p-values (ewastools [20]), and global beta-value distribution characteristics (watermelon [20]). Samples flagged by at least two distinct quality control criteria were excluded from further analyses (see [Supplementary information 2](#)).

2.2.2. Epigenetic clocks

We aimed to evaluate relevant DNA methylation-based chronological age estimators for blood samples trained on array data. Candidate clocks were identified through the literature and implemented using existing wrapper packages, including MethylClock (v 1.12.0 with normalize=FALSE) [21] and pyaging (v 0.5.3 with the recommended "Illumina Human Methylation Arrays" setting from the authors tutorial) [22], which facilitated standardized application across datasets. In addition, we included PAYA [8], Garma (the Garma Young model) [16], cAge [15] and Gaussian process regression (GP) clocks from Varshavsky et al. [23]. These clocks were imputed by using the mean beta values. A full list of clocks and details about training can be found in [Table S2](#).

2.2.3. Performance metrics

Conventional performance metrics, including mean absolute error (MAE) and median absolute deviation (MAD), root mean square error (RMSE) and R^2 , were used to evaluate the age models.

2.3. External comparisons

2.3.1. Replicate effects

Duplicate processing was performed for 131 of the 142 samples (92.3%) from the Lebanon cohort. We used these samples to assess the absolute difference between age estimates from replicates referred to as replicate distance.

2.3.2. Normalisation effects

To assess the effect of DNA methylation normalisation on the age estimates, multiple pre-processing methods were applied. Raw IDAT files were imported using minfi (v 1.52.1) [17], and beta values were generated following normalisation using NOOB (preprocessNoob), SWAN (preprocessSWAN), quantile normalisation (preprocessQuantile), functional normalisation (preprocessFunnorm), and Illumina normalisation (preprocessIllumina), all with default parameters.

BMIQ normalisation was performed on raw beta values obtained via *preprocessRaw*, using probe-type information to correct type I and type II probe bias with the *BMIQ* function from the *watermelon* package (v2.12.0) [18]. In addition, normalisation was performed using the *SeSAMe* (1.24.0) [19] framework (*openSesame*, *prep* = "QCDBP"). CpGs with multiple probes were collapsed using the *betasCollapseToPfx* function from *SeSAMe*.

2.3.3. Circadian rhythm effects

To assess whether epigenetic age estimates from different DNA methylation clocks exhibit significant rhythmicity across the circadian cycle, we searched for all available blood-based datasets with time of sample collection as metadata. We performed cosinor-based regression analyses as implemented by Koncavičius et al. [24]. A cosinor model incorporating sine and cosine terms with a fixed 24-hour period (H1) was compared to a null model including only the intercept (H0) as suggested by Cornelissen et al. [25]:

$$H1: Y(t) = b_0 + b_1 \sin(2\pi \cdot t/24) + b_2 \cos(2\pi \cdot t/24) + \epsilon$$

$$H0: Y(t) = b_0 + \epsilon$$

Model fits were compared using a likelihood ratio test and the resulting p-value quantified the statistical significance of oscillatory variation for each clock estimate and corrected for false discovery rate with the Benjamini-Hochberg method [26].

2.3.4. Sex differences

For each epigenetic clock, prediction error was defined as the difference between estimated DNA methylation age and chronological age ($\Delta\text{Age} = \text{Estimated_age} - \text{Age}$) to account for age as a confounding effect. Sex effects on systematic prediction error were assessed using linear mixed-effects models with sex as a fixed effect and dataset as a random intercept to account for batch effects. Chronological age was included as a covariate using a natural cubic spline with three degrees of freedom to allow for nonlinear age-dependent bias in clock predictions.

Models were fitted separately for each clock using maximum likelihood estimation. Fixed-effect estimates and 95% confidence intervals were obtained using Wald statistics. Two-sided p-values for sex effects were computed from the Wald test statistics assuming asymptotic normality. To account for multiple testing across clocks, p-values were adjusted using the Benjamini-Hochberg false discovery rate (FDR) procedure. Sex effects were considered statistically significant at FDR-adjusted $p < 0.05$.

3. Results

3.1. Performance of epigenetic clocks in adolescents and young adults

We evaluated the performance of a broad set of publicly available epigenetic clocks for chronological age prediction in an independent test

set comprising DNA methylation measured in blood from 687 samples in the age interval 12–25 years (Figure S1 and Supplementary information 1). The dataset was approximately balanced for sex (44% female) and included samples from 12 cohorts spanning 8 nationalities (Table S1). To assess robustness across technological platforms, the test data included Illumina 450k, EPIC, and EPIC2 arrays.

Among all evaluated clocks, cAge produced age estimates closest to chronological age, as measured by median absolute deviation (MAD = 1.05 years; Table 1), and achieved the lowest mean absolute error (MAE = 1.35 years). Its performance was closely followed by PAYA (MAD = 1.17 years; MAE = 1.39 years) and Garma (MAD = 1.39 years; MAE = 1.69 years). These three models also yielded the lowest root mean squared errors (RMSE), with values of 2.72, 1.81, and 2.39 years for cAge, PAYA, and Garma, respectively. In contrast, the remaining clocks exhibited substantially higher deviations, with MAD values ranging from 2.08 years (Zhang_BLUP) to 20.83 years (Hannum).

Given the consistently strong performance of cAge, PAYA, and Garma, we examined whether ensemble approaches could further improve accuracy. We explored various strategies to combine prediction models and found that a simple combination by using the mean or median value of the predicted ages from the models performed well. Hence, we evaluated all combinations of two to five clocks using both the mean and median of predicted age estimates from the given models (Table S3). Two combinations yielded particularly low deviations from chronological age: the mean of PAYA and Garma (MAD = 0.81 years; MAE = 1.08 years) and the mean of PAYA, Garma, and cAge (MAD = 0.79 years; MAE = 1.04 years) (Table 1). The combination of three clocks also had the lowest MAD and MAE for in-house generated EPIC2 data (Table S4) and will be referred to as Ensaya (ensemble age model for prediction of chronological age in adolescents and young adults).

The number of CpG sites included in PAYA ($n = 267$) is substantially lower than in the Garma model ($n = 3838$ unique sites) and cAge ($n = 2330$). Despite this difference, 49 CpG sites are shared across all three models (Figure S2). Garma and cAge share an additional 882 sites, bringing the total to 931 shared CpG sites between the two models. These shared sites correspond to 40% of the CpG sites in cAge and 23% of those in the Garma model. Garma contains the highest proportion of unique CpG sites (75%), whereas cAge and PAYA show comparable proportions of unique sites (59% and 61%, respectively).

Assessment of model performance stratified by cohort and array platform revealed high robustness for the top-performing clocks (Figs. 1B and 1C). Across cohorts, MAD values remained relatively stable for cAge (range=0.87–2.2 years, IQR=0.73), PAYA (range=0.79–2.99 years, IQR=0.9), and Garma (range=0.75–2.55 years, IQR=0.39). Ensemble predictions generally further reduced error, with MAD ranges

Table 1

Performance metrics for the evaluated epigenetic age prediction models. Reported performance measures include median absolute deviation (MAD), mean absolute error (MAE), root mean squared error (RMSE), and coefficient of determination (R^2). Models are ranked by MAD.

Clock	MAD	MAE	RMSE	R2	Publication
Ensaya	0.79	1.04	1.71	0.77	
PAYA_Garma	0.81	1.08	1.62	0.79	
cAge	1.05	1.35	2.27	0.65	[15]
PAYA	1.17	1.39	1.81	0.72	[8]
Garma	1.39	1.69	2.39	0.75	[16]
Zhang_BLUP	2.08	2.49	3.42	0.60	[11]
Altumage	2.24	3.01	4.41	0.31	[27]
Zhang_EN	2.30	3.78	5.33	0.54	[11]
GP10	2.81	3.32	4.42	0.40	[23]
GP71	2.81	3.32	4.44	0.50	[23]
Horvath	3.41	4.19	5.54	0.49	[2]
GP30	3.78	4.30	5.55	0.46	[23]
PCHorvath	7.78	7.75	8.77	0.48	[28]
Horvath_skin	8.33	8.25	8.58	0.51	[29]
Hannum	20.83	21.11	21.72	0.39	[1]

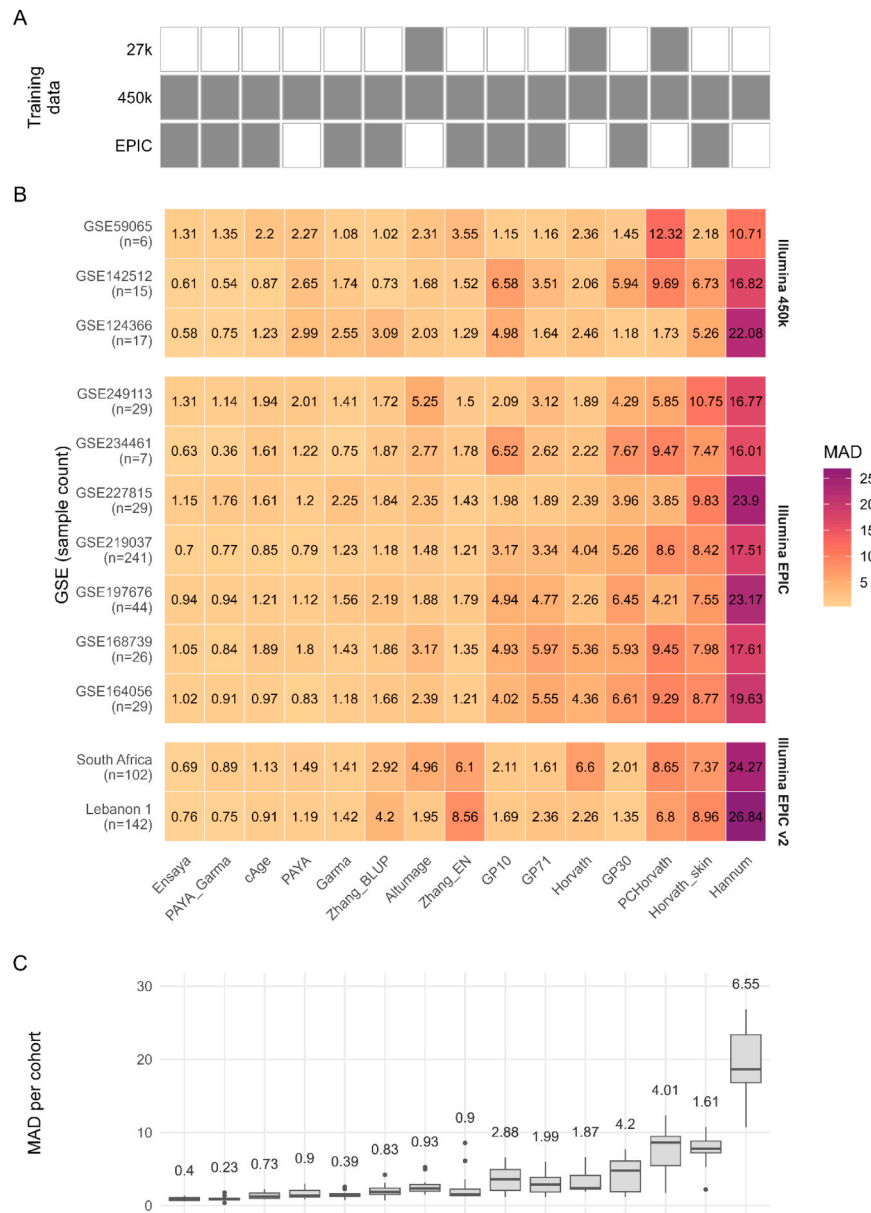


Fig. 1. A) Illumina platform technology used to train the tested models in this study. B) Tile plot of median absolute deviation (MAD) values across models and cohorts are shown on the y-axis, along with the number of included samples. The Illumina platform used in the given cohort is indicated on the right side of the y-axis. C) Box plot of MAD values per cohort. Numbers above the boxplots are the interquartile ranges (IQR, Q3-Q1).

of 0.36–1.76 years for PAYA + Garma (IQR=0.23) and 0.58–1.31 years (IQR=0.4) for Ensaya. No substantial differences in performance were observed between Illumina 450k, EPIC, and EPIC2 platforms for these leading models. This is notable given that none of the top-performing clocks were trained on EPIC2 data (PAYA on 450k; cAge and Garma on 450k and EPIC; Fig. 1A), displaying good transferability to newer array technologies. In contrast, some models appeared more sensitive to platform differences: Zhang_BLUP and Zhang_EN showed a tendency toward increased MAD on EPIC2 data, while the GP-based clocks (GP10, GP30, GP71) performed worse on EPIC arrays despite having been trained on both 450k and EPIC data.

Inspection of age-prediction residuals (Fig. 2) revealed systematic age-dependent biases in individual clocks. PAYA tended to overestimate age at the younger end of the age spectrum and underestimate age among older individuals. Garma exhibited the opposite pattern, with underestimation among younger samples and overestimation among older ones. Notably, these opposing trends largely cancelled out in

ensemble predictions, resulting in minimal systematic bias across age ranges.

To quantify the extent to which different DNA methylation clocks capture overlapping biological signals, concordance was assessed by correlating their absolute prediction errors, which reflect agreement in deviations from the true age. Correlations varied widely (Figure S3), ranging from -0.50 (GP30 vs. Hannum) to 0.82 (GP30 vs. GP71) as strongest association observed between single-clock predictions. Ensemble model predictions were highly correlated with their constituent clocks, and strong correlations were also observed among GP clocks trained on similar data. cAge showed moderate correlations with both PAYA ($r = 0.50$) and Garma ($r = 0.66$), whereas the correlation between PAYA and Garma was substantially weaker ($r = 0.36$).

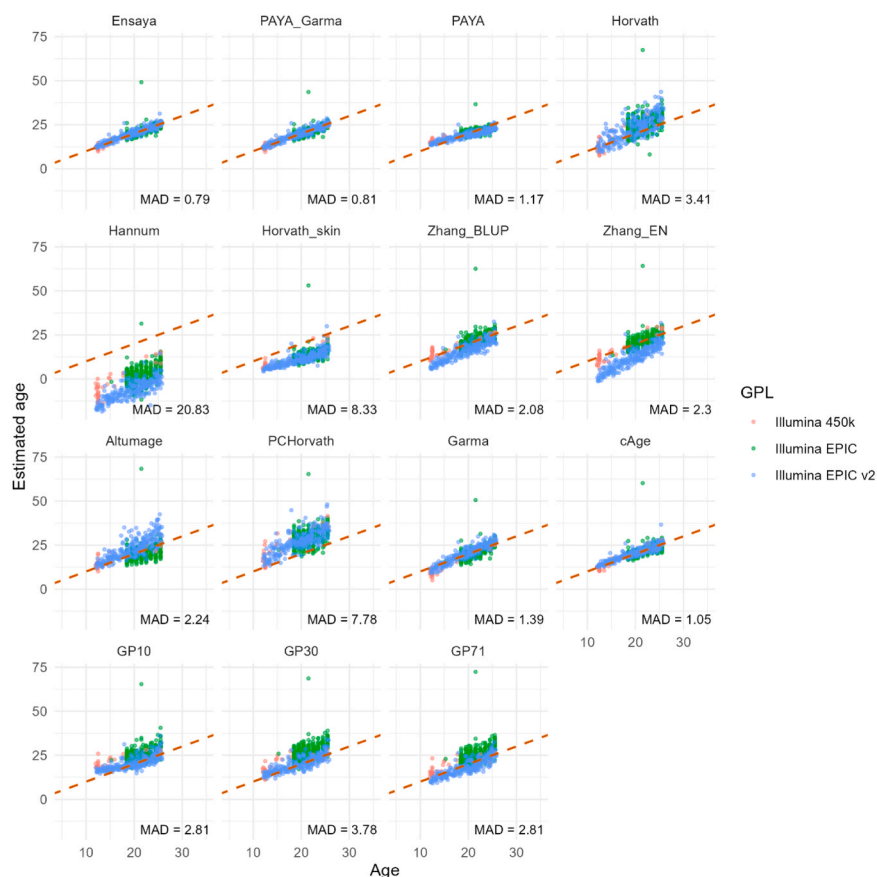


Fig. 2. Scatter plots showing chronological age vs epigenetic age as estimated by tested clocks with median absolute deviation (MAD), tested across all 12 cohorts ($n = 687$).

3.2. External analysis

3.2.1. Replicate effects

To assess absolute difference between age estimates from replicates, we created duplicates from 131 of the samples from Lebanon (termed Lebanon 2). We computed the replicate distance as the absolute difference in age estimates between the two replicates per clock. We found that PAYA had the lowest median replicate distance of 0.36 years (Fig. 3, Table S5) followed by the ensemble models (0.45 years for both

PAYA+Garma and Ensaya). PAYA + Garma has a replicate distance of more than one year for 23 of the samples, and 21 for Ensaya. cAge has a median replicate distance of 0.56 years, but with a considerable standard deviation of 0.73 years, including one sample with a replicate distance of 4.26 years, 10 out of the 131 samples differing by more than two years and 31 by more than one year.

3.2.2. Normalisation effects

Since the evaluated clocks were trained using diverse data

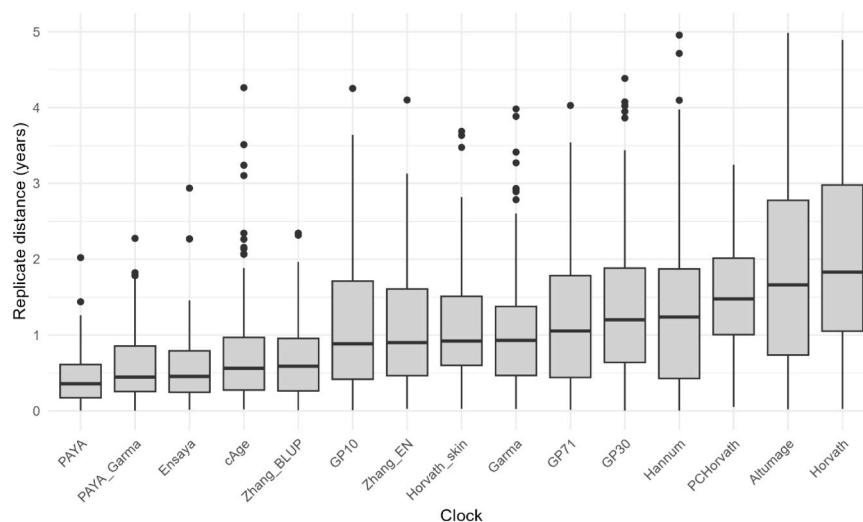


Fig. 3. Boxplot of the absolute difference between age estimates from epigenetic models of 131 replicate samples from Lebanon.

normalisation strategies, we next assessed the impact of normalisation on performance using a subset of the test data (South Africa ($n = 102$) and Lebanon 1 ($n = 142$) processed with several commonly used methods (BMIQ, functional normalisation, Illumina, NOOB, quantile normalisation, SeSAmE, and SWAN). Surprisingly, concordance between training and test normalisation methods did not consistently improve age prediction accuracy (Fig. 4). For example, SkinHorvath performed poorly on NOOB normalised data ($MAD = 8.34$ years), despite being trained on NOOB normalised data, but showed the best performance with quantile normalisation ($MAD = 2.09$ years). For the top-performing clocks and ensembles (cAge, PAYA, Garma, PAYA + Garma, Ensaya), MAD values were generally robust to normalisation choice, with the notable exception of quantile normalisation. Among these models, cAge was most affected by normalisation, with MAD values ranging from 0.93 years (BMIQ) to 4.14 years (quantile normalisation). For ensemble models, differences across normalisation methods were smaller: PAYA + Garma achieved its lowest MAD with NOOB normalisation (0.78 years), while Ensaya performed best with functional normalisation ($MAD = 0.73$ years), closely followed by NOOB ($MAD = 0.75$) and Sesame ($MAD = 0.76$).

3.2.3. Circadian rhythm effects

We tested the hypothesis that epigenetic age oscillates with time of day, as previously suggested by Koncavičius et al. (2024) [24]. We analysed the same datasets included in the original Koncavičius study, along with one additional dataset we identified, consisting of whole-blood cells from 15 individuals (OSF. DOI: 10.17605/OSF.IO/QX7W3), resulting in six datasets with available sample collection times. The analysed datasets also varied in their blood cell fraction compositions, as summarized in Tables S6A and S6B. After correction for multiple testing, a significant oscillatory pattern was observed in only one dataset (GSE247197; Tables S6A and S6B), and only for five epigenetic clocks ($p < 0.05$: Hannum, Horvath, Zhang BLUP, SkinHorvath, and GP71). No consistent effect was detected in the remaining five datasets, suggesting that circadian fluctuations in epigenetic age, if present, are subtle and not readily generalizable across cohorts.

3.2.4. Sex differences

To assess the potential effect of sex on the age prediction, we fitted regression models per clock with sex as fixed effects and dataset as

random intercepts. Four clocks (Horvath, Hannum, PCHorvath and GP30) showed significant (FDR-adjusted $p < 0.05$) sex differences in prediction bias after adjustment for age and dataset effects (Table S7). However, the direction of the effect was not uniform: three clocks predicted higher ages for females, whereas one clock predicted higher ages for males (GP30).

4. Discussion

4.1. Performance of epigenetic clocks in adolescents and young adults

Substantial research efforts have focused on improving chronological age prediction using DNA methylation data, leading to a wide range of methodological approaches. These include the use of different regularised regression [1,2], deep learning frameworks [13,27,30,31], pre-selection of age-associated CpG sites [15,16], dimensionality reduction approaches such as principal component analysis (PCA) [28], training on restricted age ranges (e.g. [8,9,14]), and the development of multi-tissue [1,2,11,27,29,32] and tissue-specific [9] models. In this study, we systematically compared openly available epigenetic age prediction models for blood samples from individuals aged 12–25 years, with particular emphasis on model robustness and the potential benefits of ensemble-based approaches. Our main finding was that an ensemble strategy combining two or three recently published models provided the best overall performance. The combinations of three (cAge, Garma and PAYA) and two (Garma and PAYA) both performed similarly in terms of MAD and MAE for the complete dataset. However, the three clock combination outperforms Garma and PAYA in the in-house dataset, which is more relevant for future age estimations as EPIC v2 is the latest platform. Further, this combination showed lower MAD values for all tested normalisation strategies, except quantile normalisation. This ensemble model, termed Ensaya (Ensemble model for adolescents and young adults), leverages predictors trained on large and heterogeneous datasets that are not readily accessible to the wider research community. When evaluated in independent validation data from adolescents and young adults, Ensaya achieved a median absolute deviation (MAD) of 0.79 years. This is, to our knowledge, the most accurate of the openly accessible blood-based prediction models for chronological age in the range 12–25 years. PAYA [8] reports an even lower MAD of ~0.7 years but is tested on a narrow age span of strictly 18-year-olds. In this study,

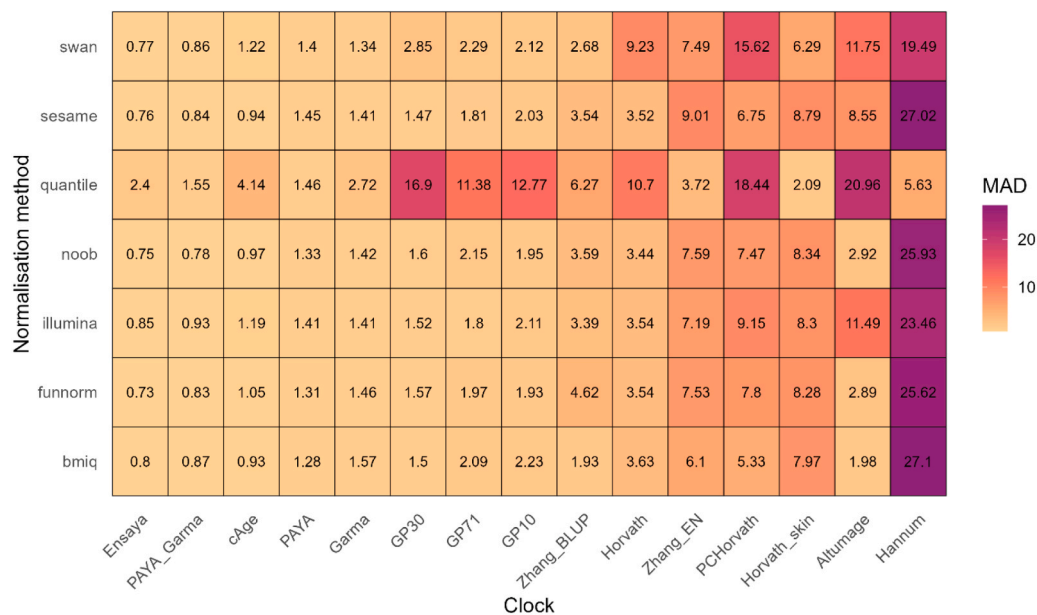


Fig. 4. Performance by median absolute deviation (MAD) of epigenetic clocks using different normalisation methods. This was tested using the in-house datasets South Africa ($n = 102$) and Lebanon 1 ($n = 142$).

we observed a systematic bias for PAYA: younger ages were overestimated, and older ages were underestimated.

The component models included in Ensaya exhibited partly complementary properties. As already highlighted, the PAYA model showed high accuracy around 18 years of age but not for younger or older individuals, displaying an attenuated prediction slope across a broader age range. Surprisingly, the Garma model [16] showed the opposite trend, while cAge [15] did not exhibit a clear systematic bias within the tested interval. Combining these models using a simple mean yielded improved overall performance by mitigating component-specific weaknesses.

DNA methylation changes are more dynamic during childhood [33], which motivates the development of age-specific epigenetic clocks [8,9,13,14]. It has been shown that methylation changes in early life might be better modelled logarithmically [2]. However, more recent studies have demonstrated that linear and logarithmic model formulations yield comparable predictive performance from around 12 years of age [8]. Nonetheless, restricting training data to narrower age intervals reduces variance and influences CpG site selection, with important consequences for model generalizability. Several studies have previously evaluated epigenetic clocks in paediatric and adolescent populations. For example, McEwen et al. (2020) demonstrated that the Paediatric-Buccal Epigenetic (PedBE) clock, based on 94 CpG sites, predicts chronological age with high accuracy in buccal samples from individuals aged 0–20 years (median absolute error ~0.35 years). However, the strong performance may partly reflect the age distribution of the test data, which consisted mainly of children younger than 10 years.

The cAge, PAYA, and Garma models showed the lowest MAD in our test data, and all included optimisation strategies tailored to younger populations. PAYA was trained specifically on the same age interval as our test cohort. cAge applies a two-stage approach, using a general model to classify individuals before applying age-specific models for those younger or older than 20 years. Similarly, Garma was trained across defined age strata and provides a dedicated "young" model for individuals under 46 years. This age-specific model prediction, along with the models' training on large, diverse datasets, are arguably the main reasons for the ensemble approach's effectiveness.

Our analysis of correlations among different epigenetic clocks generally reveals low to moderate concordance in their deviations from chronological age. Exceptions include ensemble clocks composed of overlapping component clocks, as well as individual clocks and the ensemble models that incorporate them (e.g., Garma, PAYA, and cAge with Ensaya), which show higher correlations. Similarly, higher correlations are also observed among clocks trained on the same dataset (e.g., the Zhang clocks (BLUP and EN) and the GP-based clocks (GP10, GP30 and GP71)). In contrast, PAYA and Garma show a relatively low correlation, likely due to their opposing regression biases with respect to chronological age.

Beyond predictive accuracy, our results emphasize the importance of extensive evaluation of different age prediction models across independent datasets. We systematically tested the models on data from 12 independent cohorts representing diverse genetic ancestries, biological conditions, external exposures, study designs, and platform technologies, and observed substantial variability in performance across cohorts for some models. By examining cohort-level variability, it is possible to assess model transportability and robustness, which may be masked in aggregate performance. We found that the ensemble models showed the lowest variation in cohort-specific MAD, as measured by the IQR, indicating that they were the most robust among the evaluated clocks. Analyses of technical replicates further indicated that a proportion of residual variation in age prediction (median replicate distance of ~0.45 years for Ensaya) remains attributable to technical variation. Ensaya combines 6503 CpG sites (PAYA: 267, Garma: 3906 (3838 unique across linear and quadratic sites), cAge: 2330) through three individual age estimates, with the final prediction defined as their mean. While this approach may capture a broad epigenome-wide age signal, further validation in genetically and biologically heterogeneous populations is

essential to assess robustness and interpretability fully. The expansion of such comparative datasets should be a key focus for the field moving forward.

An important consideration in developing epigenetic age models is the quality of the training and test data. Previous work has shown that a proportion of samples in the Gene Expression Omnibus (GEO) are mislabelled or swapped [20], with 133 of 8327 samples (1.6%) showing discordance between reported and methylation-predicted sex. Because sex discordance only detects swaps between individuals of opposite sex, this represents a lower bound of the true mislabelling rate. Assuming random pairwise swaps and a balanced sex ratio, the total mislabelling rate may be closer to ~3%. Such residual mislabeling could manifest as extreme outliers and may attenuate overall model performance. In this context, the use of robust regression methods may help mitigate the impact of such mislabeled samples by downweighting extreme observations during model fitting.

Some of the validation datasets were selections of individuals with disease (cancer survivors, diabetes, and COVID–19). These populations do not represent healthy populations, and prior studies have shown that disease and trauma can influence DNA methylation patterns, potentially reducing the performance of epigenetic age models. However, these datasets did not exhibit any significant outliers in our MAD analysis, compared to studies with individuals labelled as normal health, and they do not constitute a disproportionate share of the total test data. Also, they represent conditions of interest that may be prevalent in real-world populations. We therefore included them in our analyses.

4.2. External analysis

4.2.1. Effect of normalisation pre-processing

We assessed the impact of different normalisation strategies on model performance. For most models, normalisation had only a moderate effect. However, quantile normalisation consistently led to a pronounced reduction in predictive accuracy for the top performing clocks. This is consistent with a prior benchmarking study that reported poorer performance for quantile normalisation compared to other normalisation methods [34]. Interestingly, testing the models with the same normalisation strategy as they were built on did not necessarily yield the highest performance. These findings suggest that, for most models, normalisation strategies might not affect age estimation accuracy as much as previously expected, but should be evaluated further.

4.2.2. Circadian rhythm effects are not consistently detectable

Circadian effects on epigenetic age estimation have been reported for several established epigenetic clocks and may represent a substantial source of biological variation in age estimates. Koncinski et al. [24] proposed that these effects partially reflect time-dependent changes in cell-type composition, although similar patterns were also observed in purified neutrophil fractions. One limitation of that study is the small sample size of the individual analyses used to fit circadian patterns with different acrophases. We therefore analysed the same datasets as in the original study, along with one additional whole-blood dataset from 15 individuals. An oscillatory pattern was detected in five epigenetic clocks in only one dataset, GSE247197, which consisted of whole blood depleted of neutrophils. This is the same dataset on which Koncinski et al. based their main findings. No circadian oscillatory pattern was observed in any of the other datasets or in any other clocks, such as the more recent models PAYA, cAge, and Garma nor in the ensemble clocks. Together, these findings suggest that certain epigenetic clocks may be more sensitive than others to variation in blood cell composition, and that the effect observed in GSE247197 may, in part, reflect stochasticity associated with the small sample size.

4.2.3. No significant sex differences found for the Ensaya model

All epigenetic clocks included in our study were trained and validated in mixed-sex populations. Previous studies have reported sex-

associated biases in certain epigenetic clocks. In our analyses, we observed small but statistically significant sex differences for four epigenetic clocks (Horvath, Hannum, PCHorvath and GP30). However, no such differences were detected for Ensaya or for any of the clocks included in this ensemble model.

4.2.4. Limitation of the study

Our model is largely built and exclusively tested on a population spanning 12–25 years. As such, it is not suitable as a first line estimator for age in e.g. trace samples from a crime scene where the assumed age may lie outside this range. Its primary application is to estimate age in living humans and other situations (e.g. autopsies) where the expected age is within the interval of the model. The amount and quality of DNA needed to run the Infinium array would further complicate the model's applicability to forensic trace evidence, where samples are often limited and degraded.

As DNA methylation-based models for chronological age estimation move toward practical application in real-world populations, rigorous assessment of their generalizability is essential. An important limitation of the present study is that the Ensaya evaluation did not include a systematic analysis of the sources of performance variation across populations. Future studies should examine the extent to which biological states, including disease, external exposures such as smoking and medication use, and regional genetic heterogeneity, influence model performance, including both accuracy and calibration. Such analyses will require large, well-characterised cohorts with broad, evenly distributed age ranges and sufficiently detailed metadata to enable robust assessment of population-specific effects.

5. Conclusion

In conclusion, this study evaluates openly available DNA methylation-based chronological age prediction models in adolescents and young adults and demonstrates that combining complementary clocks can substantially improve predictive performance. While several recently developed models performed well individually, their integration into the ensemble model Ensaya yielded the most accurate and consistent predictions across independent datasets. The robustness of the ensemble across cohorts, array platforms, and normalisation strategies further supports its potential applicability in real-world settings. Continued validation in larger and more diverse populations, as well as refinement of uncertainty estimation, will be important steps toward the reliable application of epigenetic age prediction in forensic and biomedical contexts.

CRedit authorship contribution statement

Stenlökk Kristina Severine Rudskjær: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Øyvind Bleka:** Writing – review & editing, Writing – original draft, Validation, Software, Methodology, Investigation, Formal analysis, Conceptualization. **Andre Uys:** Writing – review & editing, Resources. **Mohamed Yasin Gamieldeen:** Writing – review & editing, Resources. **Herman Bernitz:** Writing – review & editing, Resources. **Veslemøy Rolseth:** Writing – review & editing, Writing – original draft, Project administration, Investigation, Data curation, Conceptualization. **Pål Skage Dahlberg:** Writing – review & editing, Writing – original draft, Resources, Project administration, Methodology, Formal analysis, Data curation, Conceptualization. **Antoine Saadé:** Writing – review & editing, Resources. **Liam Robinson:** Writing – review & editing, Resources. **Mona Nahas Gholmieh:** Writing – review & editing, Resources.

Declaration of Generative AI and AI-assisted technologies in the writing process

During the preparation of this work the authors used Chat-GPT in order to improve readability and language. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the content of the published article.

Code availability

A script for running Ensaya is available at <https://github.com/kristinastenlökk/Ensaya>.

Funding

The Norwegian Directorate of Immigration has financed the study “DNA methylation as a tool for estimation of age in adolescents” and the academic work concerning biological age assessment carried out by the Department of Forensic Sciences, Oslo University Hospital, where KS, ØB, PSD and VR are currently employed.

Declaration of Competing Interest

The authors declare no conflicts of interest.

Acknowledgements

We want to thank all the researchers who generously made their data public available. We would also like to thank Kristina Totland Carm and Debbie Hege Svensrud for technical assistance, and Haavard Aanes for sharing scripts on the PAYA model.

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at [doi:10.1016/j.fsigen.2026.103579](https://doi.org/10.1016/j.fsigen.2026.103579).

References

- [1] G. Hannum, J. Guinney, L. Zhao, L. Zhang, G. Hughes, S. Sada, B. Klotzle, M. Bibikova, J.-B. Fan, Y. Gao, R. Deconde, M. Chen, I. Rajapakse, S. Friend, T. Ideker, K. Zhang, Genome-wide methylation profiles reveal quantitative views of human aging rates, *Mol. Cell.* 49 (2013) 359–367, <https://doi.org/10.1016/j.molcel.2012.10.016>.
- [2] S. Horvath, DNA methylation age of human tissues and cell types, *Genome Biol.* 14 (2013) R115, <https://doi.org/10.1186/gb-2013-14-10-r115>.
- [3] S. Jiang, Y. Guo, Epigenetic clock: DNA methylation in aging, *Stem Cells Int.* 2020 (2020) 1047896, <https://doi.org/10.1155/2020/1047896>.
- [4] T.J. Peters, B. Meyer, L. Ryan, J. Achinger-Kawecka, J. Song, E.M. Campbell, W. Qu, S. Nair, P. Loi-Luu, P. Stricker, E. Lim, C. Stirzaker, S.J. Clark, R. Pidsley, Characterisation and reproducibility of the HumanMethylationEPIC v2.0 BeadChip for DNA methylation profiling, *BMC Genom.* 25 (2024) 251, <https://doi.org/10.1186/s12864-024-10027-5>.
- [5] D.W. Belsky, A. Caspi, D.L. Corcoran, K. Sugden, R. Poulton, L. Arseneault, A. Baccarelli, K. Chamarti, X. Gao, E. Hannon, H.L. Harrington, R. Houts, M. Kothari, D. Kwon, J. Mill, J. Schwartz, P. Vokonas, C. Wang, B.S. Williams, T. E. Moffitt, DunedinPACE, a DNA methylation biomarker of the pace of aging, *eLife* 11 (2022) e73420, <https://doi.org/10.18632/eLife.73420>.
- [6] M.E. Levine, A.T. Lu, A. Quach, B.H. Chen, T.L. Assimes, S. Bandinelli, L. Hou, A. A. Baccarelli, J.D. Stewart, Y. Li, E.A. Whitsel, J.G. Wilson, A.P. Reiner, A. Aviv, K. Lohman, Y. Liu, L. Ferrucci, S. Horvath, An epigenetic biomarker of aging for lifespan and healthspan, *Aging (Albany NY)* 10 (2018) 573–591, <https://doi.org/10.18632/aging.101414>.
- [7] A.T. Lu, A. Quach, J.G. Wilson, A.P. Reiner, A. Aviv, K. Raj, L. Hou, A.A. Baccarelli, Y. Li, J.D. Stewart, E.A. Whitsel, T.L. Assimes, L. Ferrucci, S. Horvath, DNA methylation GrimAge strongly predicts lifespan and healthspan, *Aging (Albany NY)* 11 (2019) 303–327, <https://doi.org/10.18632/aging.101684>.
- [8] H. Aanes, Ø. Bleka, P.S. Dahlberg, K.T. Carm, T. Lehtimäki, O. Raitakari, M. Kähönen, M. Hurme, V. Rolseth, A new blood based epigenetic age predictor for adolescents and young adults, *Sci. Rep.* 13 (2023) 2303, <https://doi.org/10.1038/s41598-023-29381-7>.
- [9] L.M. McEwen, K.J. O'Donnell, M.G. McGill, R.D. Edgar, M.J. Jones, J.L. MacIsaac, D.T.S. Lin, K. Ramadori, A. Morin, N. Gladish, E. Garg, E. Unternaehrer, I. Pokhvisneva, N. Karnani, M.Z.L. Kee, T. Klengel, N.E. Adler, R.G. Barr,

- N. Letourneau, G.F. Giesbrecht, J.N. Reynolds, D. Czamara, J.M. Armstrong, M. J. Essex, C. de Weerth, R. Beijers, M.S. Tollenaar, B. Bradley, T. Jovanovic, K. J. Ressler, M. Steiner, S. Entringer, P.D. Wadhwa, C. Buss, N.R. Bush, E.B. Binder, W.T. Boyce, M.J. Meaney, S. Horvath, M.S. Kobor, The PedBE clock accurately estimates DNA methylation age in pediatric buccal cells, *Proc. Natl. Acad. Sci. USA* 117 (2020) 23329–23335, <https://doi.org/10.1073/pnas.1820843116>.
- [10] Infinium™ Methylation Screening Array, (2024). (<https://www.illumina.com/content/dam/illumina/gcs/assembled-assets/marketing-literature/infinium-methylation-screening-array-data-sheet-m-gl-01893/infinium-methylation-screening-array-data-sheet-m-gl-01893.pdf>).
- [11] Q. Zhang, C.L. Vallerger, R.M. Walker, T. Lin, A.K. Henders, G.W. Montgomery, J. He, D. Fan, J. Fowdar, M. Kennedy, T. Pitcher, J. Pearson, G. Halliday, J. B. Kwok, I. Hickie, S. Lewis, T. Anderson, P.A. Silburn, G.D. Mellick, S.E. Harris, P. Redmond, A.D. Murray, D.J. Porteous, C.S. Haley, K.L. Evans, A.M. McIntosh, J. Yang, J. Gratten, R.E. Marioni, N.R. Wray, I.J. Deary, A.F. McRae, P.M. Visscher, Improved precision of epigenetic clock estimates across tissues and its implication for biological ageing, *Genome Med.* 11 (2019) 54, <https://doi.org/10.1186/s13073-019-0667-1>.
- [12] A.E. Teschendorff, S. Horvath, Epigenetic ageing clocks: statistical methods and emerging computational challenges, *Nat. Rev. Genet.* 26 (2025) 350–368, <https://doi.org/10.1038/s41576-024-00807-w>.
- [13] D. Martínez-Enguita, T. Hillerton, J. Åkesson, D. Kling, M. Lerm, M. Gustafsson, Precise and interpretable neural networks reveal epigenetic signatures of aging across youth in health and disease, *Front. Aging* 5 (2025) 1526146, <https://doi.org/10.3389/fragi.2024.1526146>.
- [14] X. Wu, W. Chen, F. Lin, Q. Huang, J. Zhong, H. Gao, Y. Song, H. Liang, DNA methylation profile is a quantitative measure of biological aging in children, *Aging (Albany NY)* 11 (2019) 10031–10051, <https://doi.org/10.18632/aging.102399>.
- [15] E. Bernabeu, D.L. McCartney, D.A. Gadd, R.F. Hillary, A.T. Lu, L. Murphy, N. Wrobel, A. Campbell, S.E. Harris, D. Liewald, C. Hayward, C. Sudlow, S.R. Cox, K.L. Evans, S. Horvath, A.M. McIntosh, M.R. Robinson, C.A. Vallejos, R.E. Marioni, Refining epigenetic prediction of chronological and biological age, *Genome Med.* 15 (2023) 12, <https://doi.org/10.1186/s13073-023-01161-y>.
- [16] L.D. Garma, M. Quintela-Fandino, Applicability of epigenetic age models to next-generation methylation arrays, *Genome Med.* 16 (2024) 116, <https://doi.org/10.1186/s13073-024-01387-4>.
- [17] M.J. Aryee, A.E. Jaffe, H. Corrada-Bravo, C. Ladd-Acosta, A.P. Feinberg, K. D. Hansen, R.A. Irizarry, Minfi: a flexible and comprehensive Bioconductor package for the analysis of Infinium DNA methylation microarrays, *Bioinformatics* 30 (2014) 1363–1369, <https://doi.org/10.1093/bioinformatics/btu049>.
- [18] R. Pidsley, C.C. Y Wong, M. Volta, K. Lunnon, J. Mill, L.C. Schalkwyk, A data-driven approach to preprocessing Illumina 450K methylation array data, *BMC Genom.* 14 (2013) 293, <https://doi.org/10.1186/1471-2164-14-293>.
- [19] W. Zhou, T.J. Triche, P.W. Laird, H. Shen, SeSAMe: reducing artifactual detection of DNA methylation by Infinium BeadChips in genomic deletions, *Nucleic Acids Res.* 46 (2018) e123, <https://doi.org/10.1093/nar/gky691>.
- [20] J.A. Heiss, A.C. Just, Identifying mislabeled and contaminated DNA methylation microarray data: an extended quality control toolset with examples from GEO, *Clin. Epigenetics* 10 (2018) 73, <https://doi.org/10.1186/s13148-018-0504-1>.
- [21] D. Pelegi-Sisó, P. de Prado, J. Ronkainen, M. Bustamante, J.R. González, methylclock: a Bioconductor package to estimate DNA methylation age, *Bioinformatics* 37 (2021) 1759–1760, <https://doi.org/10.1093/bioinformatics/btaa825>.
- [22] L.P. de Lima Camillo, pyaging: a Python-based compendium of GPU-optimized aging clocks, *Bioinformatics* 40 (2024) btae200, <https://doi.org/10.1093/bioinformatics/btae200>.
- [23] M. Varshavsky, G. Harari, B. Glaser, Y. Dor, R. Shemer, T. Kaplan, Accurate age prediction from blood using a small set of DNA methylation sites and a cohort-based machine learning algorithm, *Cell. Rep. Methods* 3 (2023) 100567, <https://doi.org/10.1016/j.crmeth.2023.100567>.
- [24] K. Koncevičius, A. Nair, A. Šveikauskaitė, A. Šestokaitė, A. Kazlauskaitė, A. Dulskas, A. Petronis, Epigenetic age oscillates during the day, *Aging Cell.* 23 (2024) e14170, <https://doi.org/10.1111/ace14170>.
- [25] G. Cornelissen, Cosinor-based rhythmometry, *Theor. Biol. Med. Model.* 11 (2014) 16, <https://doi.org/10.1186/1742-4682-11-16>.
- [26] Y. Benjamini, Y. Hochberg, Controlling the false discovery rate: a practical and powerful approach to multiple testing, *J. R. Stat. Soc. Ser. B (Methodol.)* 57 (1995) 289–300.
- [27] L.P. de Lima Camillo, L.R. Lapierre, R. Singh, A pan-tissue DNA-methylation epigenetic clock based on deep learning, *NPJ Aging* 8 (2022) 4, <https://doi.org/10.1038/s41514-022-00085-y>.
- [28] A.T. Higgins-Chen, K.L. Thrush, Y. Wang, C.J. Minter, P.-L. Kuo, M. Wang, P. Niimi, G. Sturm, J. Lin, A.Z. Moore, S. Bandinelli, C.H. Vinkers, E. Vermetten, B. P.F. Rutten, E. Geuze, C. Okhuisen-Pfeifer, M.Z. van der Horst, S. Schreiter, S. Gutwinski, J.J. Luykx, M. Picard, L. Ferrucci, E.M. Crimmins, M.P. Boks, S. Hägg, T.T. Hu-Seliger, M.E. Levine, A computational solution for bolstering reliability of epigenetic clocks: Implications for clinical trials and longitudinal tracking, *Nat. Aging* 2 (2022) 644–661, <https://doi.org/10.1038/s43587-022-00248-2>.
- [29] S. Horvath, J. Oshima, G.M. Martin, A.T. Lu, A. Quach, H. Cohen, S. Felton, M. Matsuyama, D. Lowe, S. Kabacik, J.G. Wilson, A.P. Reiner, A. Maierhofer, J. Flunkert, A. Aviv, L. Hou, A.A. Baccarelli, Y. Li, J.D. Stewart, E.A. Whitsel, L. Ferrucci, S. Matsuyama, K. Raj, Epigenetic clock for skin and blood cells applied to Hutchinson Gilford Progeria Syndrome and ex vivo studies, *Aging (Albany NY)* 10 (2018) 1758–1775, <https://doi.org/10.18632/aging.101508>.
- [30] S.A. Dip, D. Ma, L. Zhang, DeepAge: harnessing deep neural network for epigenetic age estimation from DNA methylation data of human blood samples, *Proc. AAAI Symp. Ser.* 4 (2024) 267–274, <https://doi.org/10.1609/aaais.v4i1.31801>.
- [31] F. Galkin, P. Mamoshina, K. Kochetov, D. Sidorenko, A. Zhavoronkov, DeepMAge: a methylation aging clock developed with deep learning, *Aging Dis.* 12 (2021) 1252–1262, <https://doi.org/10.14336/AD.2020.1202>.
- [32] K.A. Vijayakumar, G. Cho, Pan-tissue methylation aging clock: recalibrated and a method to analyze and interpret the selected features, *Mech. Ageing Dev.* 204 (2022) 111676, <https://doi.org/10.1016/j.mad.2022.111676>.
- [33] C.G. Bell, R. Lowe, P.D. Adams, A.A. Baccarelli, S. Beck, J.T. Bell, B.C. Christensen, V.N. Gladyshev, B.T. Heijmans, S. Horvath, T. Ideker, J.-P.J. Issa, K.T. Kelsey, R. E. Marioni, W. Reik, C.L. Relton, L.C. Schalkwyk, A.E. Teschendorff, W. Wagner, K. Zhang, V.K. Rakan, DNA methylation aging clocks: challenges and recommendations, *Genome Biol.* 20 (2019) 249, <https://doi.org/10.1186/s13059-019-1824-y>.
- [34] H. Welsh, C.M.P.F. Batalha, W. Li, K.L. Mpye, N.C. Souza-Pinto, M.S. Naslavsky, E. J. Parra, A systematic evaluation of normalization methods and probe replicability using Infinium EPIC methylation data, *Clin. Epigenet* 15 (2023) 41, <https://doi.org/10.1186/s13148-023-01459-z>.