TYPE Original Research
PUBLISHED 04 April 2024
DOI 10.3389/fbinf.2024.1329144

OPEN ACCESS

EDITED BY
Joao Carlos Setubal,
University of São Paulo, Brazil

REVIEWED BY
Manoj Kandpal,
The Rockefeller University, United States
Alejandra Cervera,
National Institute of Genomic Medicine (INMEGEN), Mexico

*CORRESPONDENCE
Lajoyce Mboning,
✉ lajoycemboning@ucla.edu

RECEIVED 28 October 2023
ACCEPTED 01 February 2024
PUBLISHED 04 April 2024

CITATION
Mboning L, Rubbi L, Thompson M, Bouchard L-S and Pellegrini M (2024), BayesAge: A maximum likelihood algorithm to predict epigenetic age.
Front. Bioinform. 4:1329144.
doi: 10.3389/fbinf.2024.1329144

BayesAge: A maximum likelihood algorithm to predict epigenetic age

Lajoyce Mboning[1]*, Liudmilla Rubbi[2], Michael Thompson[2], Louis-S. Bouchard[1] and Matteo Pellegrini[2]

[1]Department of Chemistry and Biochemistry, University of California Los Angeles, Los Angeles, CA, United States, [2]Department of Molecular, Cell and Developmental Biology, University of Los Angeles, Los Angeles, CA, United States

Introduction: DNA methylation, specifically the formation of 5-methylcytosine at the C5 position of cytosine, undergoes reproducible changes as organisms age, establishing it as a significant biomarker in aging studies. Epigenetic clocks, which integrate methylation patterns to predict age, often employ linear models based on penalized regression, yet they encounter challenges in handling missing data, count-based bisulfite sequence data, and interpretation.

Methods: To address these limitations, we introduce BayesAge, an extension of the scAge methodology originally designed for single-cell DNA methylation analysis. BayesAge employs maximum likelihood estimation (MLE) for age inference, models count data using binomial distributions, and incorporates LOWESS smoothing to capture non-linear methylation-age dynamics. This approach is tailored for bulk bisulfite sequencing datasets.

Results: BayesAge demonstrates superior performance compared to scAge. Notably, its age residuals exhibit no age association, offering a less biased representation of epigenetic age variation across populations. Furthermore, BayesAge facilitates the estimation of error bounds on age inference. When applied to down-sampled data, BayesAge achieves a higher coefficient of determination between predicted and actual ages compared to both scAge and penalized regression.

Discussion: BayesAge presents a promising advancement in epigenetic age prediction, addressing key challenges encountered by existing models. By integrating robust statistical techniques and tailored methodologies for count-based data, BayesAge offers improved accuracy and interpretability in predicting age from bulk bisulfite sequencing datasets. Its ability to estimate error bounds enhances the reliability of age inference, thereby contributing to a more comprehensive understanding of epigenetic aging processes.

KEYWORDS

BayesAge, scAge, epigenetic age, maximum likelihood estimation, true age

1 Introduction

While the sequence of a cell's DNA largely remains invariant during its lifespan, its epigenome changes significantly with age. One of the components of the epigenome that shows the most reproducible changes with age is DNA methylation. DNA methylation involves the covalent modification of DNA and is catalyzed by a family of DNA methyltransferases (DNMTs) that transfer a methyl group from S-adenosyl methionine (SAM) to the fifth carbon of a cytosine residue to form 5-methylCytosine (5 mC). Moore et al. (2012) In mammals, most of the methylated cytosines occur in the CpG context, and there are approximately 30M $5'$—C—phosphate—G—$3'$ (CpG) dinucleotides in the human genome.

Several methods have been developed to predict age based on methylation levels. Johnson et al. (2012) These approaches are often referred to as epigenetic clocks. Kabacik et al. (2022)The first epigenetic clocks were constructed using penalized regressionGreenwood et al. (2020), with age as the response and methylation levels as the features. The use of penalties for model coefficients led to sparse models out of the thousands of sites that were measured (typically using DNA methylation microarrays), only a few hundred had non-zero weights in the models. Hannum et al. (2013); Horvath (2013) More recently these approaches have been extended using neural networks, which can generate slightly more accurate predictions than penalized regression models. Galkin et al. (2021)However, these approaches model DNA methylation changes linearly with age and therefore fail to consider non-linear methylation trends with age. This is significant, as it is widely understood that methylation changes are rapid early in life and slow down with age Snir et al. (2019).

To address this limitation, Farrell and colleagues introduced a model named the Epigenetic Pacemaker (EPM) Farrell and Pellegrini (2020). In this model, methylation across a subset of CpGs is described as a non-linear function of an epigenetic state, rather than actual age. Importantly, this epigenetic state can assume non-linear relationships between the epigenetic state and time. For a given set of i methylation sites and j individuals, the methylation level at a single site can be expressed as $\hat{m}_{ij} = m_i^0 + r_i s_j + \epsilon_{ij}$, where $\hat{m}_{ij}$ represents the observed methylation value, m_i^0 denotes the initial methylation level, r_i is the rate of change, s_j signifies the epigenetic state, and ϵ_{ij} is a normally distributed error term.

Given an input matrix $\hat{M} = [\hat{m}_{ij}]$, the objective of the Epigenetic Pacemaker (EPM) is to ascertain the optimal values for m_i^0, r_i, and s_j that minimize discrepancies between predicted and actual methylation values across specific methylation sites. As we've previously demonstrated, in certain datasets, the epigenetic state evolves in correlation with the logarithm of time Snir et al. (2019) This implies rapid methylome alterations early in development, which then decelerate with the organism's aging. Although the EPM adeptly models some nonlinear correlations between methylation and age, it calibrates the epigenetic age in relation to chronological age to optimize alignment across all sites. However, it overlooks the potential nonlinear associations that different sites may exhibit with age.

Here we propose a new approach to overcome the limitations of existing methods. Our method considers count information rather than methylation fractions, estimates non-linear trends of methylation sites with age, and uses maximum likelihood estimation which is robust to missing data. Existing methods for methylome aging modeling often struggle with low coverage or incomplete data. For instance, linear models for epigenetic age, built on weighted sums of methylation values, inadequately handle missing data—typically by assigning it a value of zero—thereby skewing age predictions. Addressing this deficit, maximum likelihood methods have emerged to generate age estimates from sparse data sets. One existing method, "scAge" (single cell age), Trapp and Gladyshev (2021) is designed to analyze methylation in single cells. scAge harnesses a maximum likelihood strategy to ascertain the most likely age of a subject based on low count data. However, this method presumes a linear relationship in methylation changes over time and employs a heuristic for determining methylation value probabilities from the observed data. We therefore expand upon scAge's foundational principles to develop a read count based framework for modeling non-linear methylation changes with age that is resilient against missing data.

2 Materials and methods

2.1 Data acquisition and analysis

To test our approach for estimating age from methylation data we collected targeted bisulfite sequencing data from either buccal swabs or blood of 458 subjects. DNA was extracted from the buccal swabs and blood using standard protocols. Buccal swabs were incubated overnight at 50°C before DNA extraction. We applied targeted bisulphite sequencing (TBS-seq) to characterize the methylomes of the samples. The protocol is described in detail in a methods paper by Morselli et al.Morselli (2021)Briefly, 500 ng of extracted DNA were used for TBS-seq library preparation. Fragmented DNA was subject to end repair, dA-tailing and adapter ligation using the NEBNext Ultra II Library prep kit using custom pre-methylated adapters (IDT). Pools of 16 purified libraries were hybridized to the biotinylated probes according to the manufacturer's protocol. Captured DNA was treated with bisulphite prior to PCR amplification using KAPA HiFi Uracil+(Roche) with the following conditions: 2 min at 98°C; 14 cycles of (98°C for 20 s; 60°C for 30 s; 72°C for 30 s); 72°C for 5 min; hold at 4°C. Library QC was performed using the High-Sensitivity D1000 Assay on a 2,200 Agilent TapeStation. Pools of 96 libraries were sequenced on a NovaSeq6000 (S1 lane) as paired-end 150 bases. The probes used in the capture were designed to capture approximately 3,000 regions that contained CpG sites used in previously published epigenetic clocks. Greenwood et al. (2020); Levine (2018); Lu (2019).

Demultiplexed Fastq files were subject to adapter removal using cutadapt (v2.10) Martin (2011) and aligned to the GRCh38 genome using BSBolt Align (v1.3.0) Colin (2021). PCR duplicates were removed using samtools markdup function (samtools version 1.9) Li (2009). The BSBolt methylation calling function was employed to produce the CGmap files for each subject, utilizing the sorted and indexed bam files. The reference genome used during the methylation calling phase was Genome assembly GRCh38. The BSBolt matrix aggregation function was used to create the methylation matrix dataset, which subsequently informed the training of BayesAge, scAge, LASSO and the EPM model. Methylation values were measured across 46,518 CpG sites.

This study was completed on System76's Lemur Pro laptop. The laptop's specifications include 40 GB RAM and 12-core Intel i7 CPU. CGmap files were converted to Bismark format for scAge testing. Additionally, BayesAge incorporates a function to convert CGmap files into a format compatible with its prediction function.

For BayesAge, scAge, LASSO, and the EPM model we implemented 10-Fold Cross Validation. The dataset of 458 subjects was divided into 10 equal parts and the models were trained on 9 of these folds and tested on the remaining one. So for every 46 subjects left out for testing, we trained the models on the remaining 412 subjects. This process was repeated 10 times to make sure all the subjects used for the age prediction step were not used during the training step. The LASSO model was implemented using the `sklearn` package. The function `linear_model.Lasso` was used to define the lasso model with parameters for the alpha value equal to 0.02 and max iterations of 10,000 for the training. The cross validation function `cross_val_predict` from the `sklearn.model_selection` had a CV parameter equal to 10. The EPM model was implemented using the tutorial from the official website https://epigeneticpacemaker.readthedocs.io. Due to extreme downsampling to 100,000 CpG sites, the methylation calling using BSBolt would return empty matrices. To circumvent this issue, a methylation matrix with very low values of coverage and percent of samples was used. Since the Lasso and EPM model requires finite input values, any NaN entries were imputed as zero methylation. While not ideal, this allowed the model to be trained and make age predictions on the available data.

2.2 The BayesAge framework

The BayesAge framework consists of two phases: training and prediction as seen in Figure 1. In the training step, we use LOWESS (locally weighted scatterplot smoothing) to fit the trend between individual methylation levels and age. We use LOWESS smoothing so that we do not need to make any *a priori* assumptions about the functional form of the association between methylation and age. The τ parameter determines the smoothness of the LOWESS fit. The LOWESS function used was from the `statsmodels.api`. After computing the fit for each site we also calculate the correlation between methylation levels and age using the Spearman rank correlation, which is robust to non-linear trends. We select the top sites to include in the prediction phase using the absolute value of the correlation. At the end of this process, the trained model consists of N sites and their methylation levels across ages, from 1 to 100 in increments of 1 year, based on a predetermined τ parameter of the LOWESS fit.

In the Prediction step, this reference matrix is intersected with the CpG sites measured in a specific sample. A count matrix of these CpG sites is constructed that reports the number of cytosines and thymines. For the chosen age-associated CpG sites, it is posited that the chance of detecting the observed cytosine and thymine counts given the intended methylation level for a specific age based on the trained model, follows a binomial distribution. To compute the probability of observing the counts measured across all sites that are found following the intersection with the training matrix, we compute the product of these probabilities. To prevent underflow errors during computation, a logarithmic sum replaces the product of individual CpG probabilities, which results in a singular probability value for each age.

Utilizing these pre-identified, ranked age-associated CpG sites, the framework calculates the likelihood of observing each age in a single subject, spanning an age spectrum of 0–100 years, at an interval of 1 year. Consequently, for each subject we compute an age-likelihood distribution, with the maximum likelihood age interpreted as the epigenetic age for subject X. Here, Pr_{CpG} represents the methylation probability for a distinct CpG at a specific age, aggregated from 1 CpG to N total CpGs. The associated probability for a unique CpG site state is mathematically detailed as follows:

$$\Pr_{CpG}(x; n) = \binom{n}{x} p^x q^{n-x} \tag{1}$$

where:

n: Reads of all cytosines.

x: Reads of methylated cytosines.

p: The predicted average methylation probability.

$q = 1 - p$: The probability of thymine counts.

2.2.1 Data simulation and average

The data simulation was executed by creating 100 synthetic samples for each real sample for a total 45,800 samples. For each synthetic sample, the counts at each site were simulated using the `scipy.stats.binom` probability distribution function. Futhermore, BayesAge prediction function was used to estimate the age of each of the 100 synthetic samples. The 100 synthetic samples of each real sample was used to calculate the lower interquartile range (IQR) and upper IQR limit bounds. Finally, the average age of the error limit bounds reported were calculated using:

$$\mu = \sum_{i=1}^{458} \frac{X_i - Y_i}{458}, \tag{2}$$

where X and Y represent the upper IQR limit and lower IQR limit of sample i respectively.

3 Results

3.1 BayesAge framework

We set out to develop a framework to estimate the age of an individual from bisulfite sequence data. Bisulfite conversion converts unmethylated cystosines to thymines, while leaving methylated cytosines unconverted. After bisulfite converted DNA is sequenced, the reads are aligned to genome and the methylation state of any cytosine in the genome is measured by counting the number of cytosines and thymines that align to that position. Typically the methylation level is estimated by computing the ratio of cytosines to cytosines plus thymines. However, since sequencing data is inherently count based, it is important to consider not only the methylation level, but also the total coverage, as the confidence of the methylation estimate increases with increased coverage.

To track changes in DNA methylation with age we collected DNA methylation data from over 400 individuals. To identify CpG sites whose methylation changed with age in a tissue independent manner, we collected our sample from both blood, saliva and buccal swabs. These tissues have heterogeneous mixtures of both hematopoietic as well as epithelial cells, and represent typical

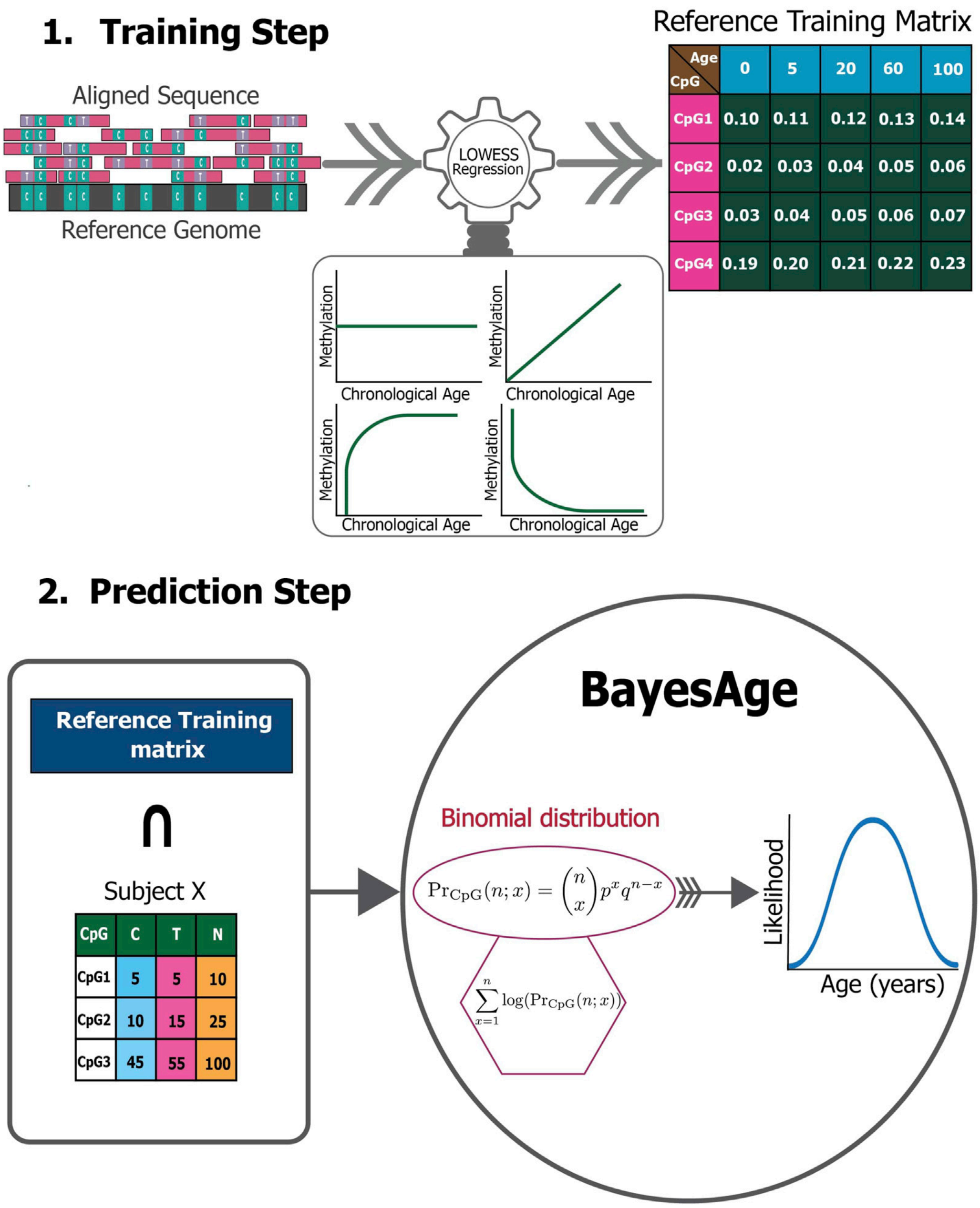
1. Training Step
Aligned Sequence
Reference Genome
LOWESS Regression
Reference Training Matrix
Age
CpG
0
5
20
60
100
CpG1 0.10 0.11 0.12 0.13 0.14
CpG2 0.02 0.03 0.04 0.05 0.06
CpG3 0.03 0.04 0.05 0.06 0.07
CpG4 0.19 0.20 0.21 0.22 0.23
Methylation
Chronological Age
2. Prediction Step
Reference Training matrix
∩
Subject X
CpG C T N
CpG1 5 5 10
CpG2 10 15 25
CpG3 45 55 100
BayesAge
Binomial distribution
$\mathrm{Pr}_{\mathrm{CpG}}(n;x) = \binom{n}{x} p^x q^{n-x}$
$\sum_{x=1}^{n} \log(\mathrm{Pr}_{\mathrm{CpG}}(n;x))$
Likelihood
Age (years)

FIGURE 1
The BayesAge framework.

cells that are found in many tissues. As whole genome sequencing is resource intensive, we used a targeted approach to only sequence a few thousand loci across the genome. This allowed us to obtain about 100x coverage of 46,517 targeted regions. We made sure to include among these loci regions that had been previously shown to have age associated DNA methylation changes. We include samples that covered a broad range of ages, from neonates to 92 years old as seen by the histogram in Figure 2. The data used to plot the histogram is added as supplementary information.

Our BayesAge framework has two steps: in the first we train a model and in the second we predict the age of a sample. In the training phase we first select CpG sites in the genome that have age associated methylation values. Previous work has shown that many CpG sites in the genome have DNA methylation levels that increase with age, but that the association between methylation and age is not necessarily linear. In fact many sites show non-linear changes of methylation with age that are well approximated by exponential functions, with the rate of change decreasing with age. To identify the most significantly age-associated sites, we used the Spearman rank correlation, which is a non-parametric method that does not assume linearity. The 16 sites in the genome with the highest Spearman correlation values are shown in Figure 3.

We find that for many of these sites the association of methylation with age is non-linear, with rates of change that

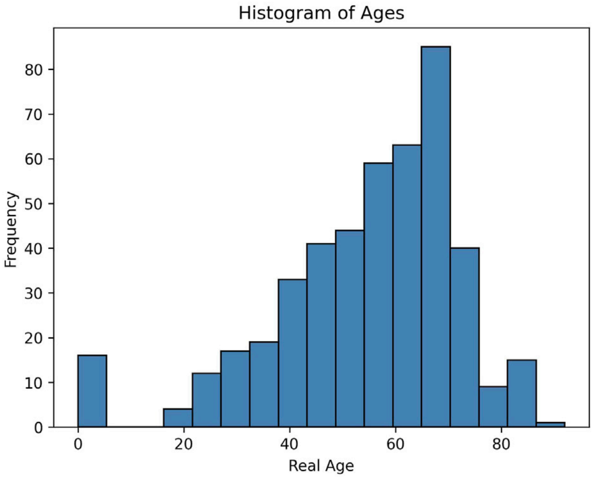

FIGURE 2
Distribution of the ages used in the training and prediction in this study.

decrease with age. BayesAge, recognizing these non-linear trends, uses LOWESS (locally weighted scatterplot smoothing) regression to model the trend lines. This method utilizes locally weighted linear regression to estimate a smoothed line to the data. The extent of this smoothing is governed by the tau parameter, which determines the size of the local neighborhood window for each local linear regression fit. For our dataset, a tau value of 0.7 was chosen, allowing the fit to adaptively capture the nonlinear methylation variations across the age spectrum without succumbing to overfitting, as illustrated in Figure 3. The trend lines of these fits represent our aging model, or the expected methylation with age at each of these sites.

Our prediction step allows us to estimate the age of a sample using the training model. Our approach to estimating age from the methylation data of a single sample builds on the count based nature of bisulfite sequencing data. We propose a Bayesian framework for estimating the most likely age of an individual by computing the probability of the observed counts of cytosines and thymines for any given age, and selecting the age that maximizes this probability. This maximum likelihood approach was first proposed in the scAge method, which was developed to estimate the age of a samples from single cell methylation data. In contrast to scAge, our method is designed for bulk DNA methylation data where the coverage of cytosines in the genome is generally high, and the methylation levels can assume values between zero and one.

To estimate the probability of the observed counts based on the expected methylation levels of a single site at a specific age we use a binomial distribution. The rationale for utilizing a binomial distribution to characterize the sites is anchored in the nature of bisulfite sequencing data. Rather than probabilities, bisulfite sequencing yields count data. Each sequenced read at a CpG site represents a Bernoulli trial with two potential outcomes: the observation of a methylated cytosine, the probability of which equals the intrinsic methylation level at that site, or the observation of an unmethylated cytosine, with the probability being the complement of the methylation level. Consequently, the cumulative counts of methylated and unmethylated reads across all sequenced reads at a particular site adhere to a binomial distribution dictated by the methylation level. This modeling approach for CpG sites echoes the discrete, count-centric nature of sequencing data, diverging from the notion of methylation as continuous probabilities.

Figure 1 and Eq. 1 exemplify the probability calculation phase, which evaluates the likelihood of detecting specific cytosine and thymine counts, given the anticipated methylation level for each CpG across varying ages, as derived from the LOWESS regression. In contrast, scAge's training phase employs a linear regression model, as evident in Eq. 3, to forecast the methylation level for every CpG site across different ages.

$$\mathrm{Meth}_{CpG} = a_{CpG}{}^{*}age + b_{CpG} \tag{3}$$

In the final step, in order to estimate the probability of a specific age of a sample we compute the likelihood of our observed counts across multiple sites for any given age by taking the product of the probabilities of each site. In practice this product is computed by summing the logarithms of the probabilities. In the final stage we compute the age that maximizes this probability.

We applied BayesAge to our samples using a model that contains the top 8 sites and used 10-fold cross validation to train and test the model. We find that the R^2 between the predicted age and the actual age is 0.78, and that the mean absolute error of the age estimate is 7 years, indicating a relatively strong correlation and accuracy in age prediction. The cross validation method is implemented such that the model is trained on 412 random samples and then tested on the remaining 46 samples. This process is repeated until all the samples are used for prediction.

Along with estimating the most likely age of a sample, BayesAge enables the calculation of error bounds on the estimate. To generate the prediction error bounds for bdAge, we employed data simulation. By running 100 simulations for each sample, we derived the interquartile range (IQR) of age predictions as a measure of uncertainty. Across all the simulations, the average IQR was approximately 12 years, providing an estimate of the typical error margins for Bayesian age predictions in our dataset as seen in Figure 4.

3.2 Comparison of BayesAge with scAge and penalized regression models

To evaluate the performance of BayesAge we compared it to scAge and penalized regression models using the same dataset of 458 individuals. As illustrated in Figure 5, 10-Fold Cross Validation coupled with mean absolute error (MAE) was employed to validate the outcomes of all models. The results show that BayesAge age estimations have a slightly higher coefficient of determination (r^2) compared to scAge. Notably, when limited to the top 8 CpG sites for age prediction, BayesAge outperforms scAge by approximately 1%. However, as we increase the number of CpG sites employed in the prediction, this performance disparity becomes more evident. Specifically, utilizing the top 256 CpG sites, BayesAge achieves an R^2 value of

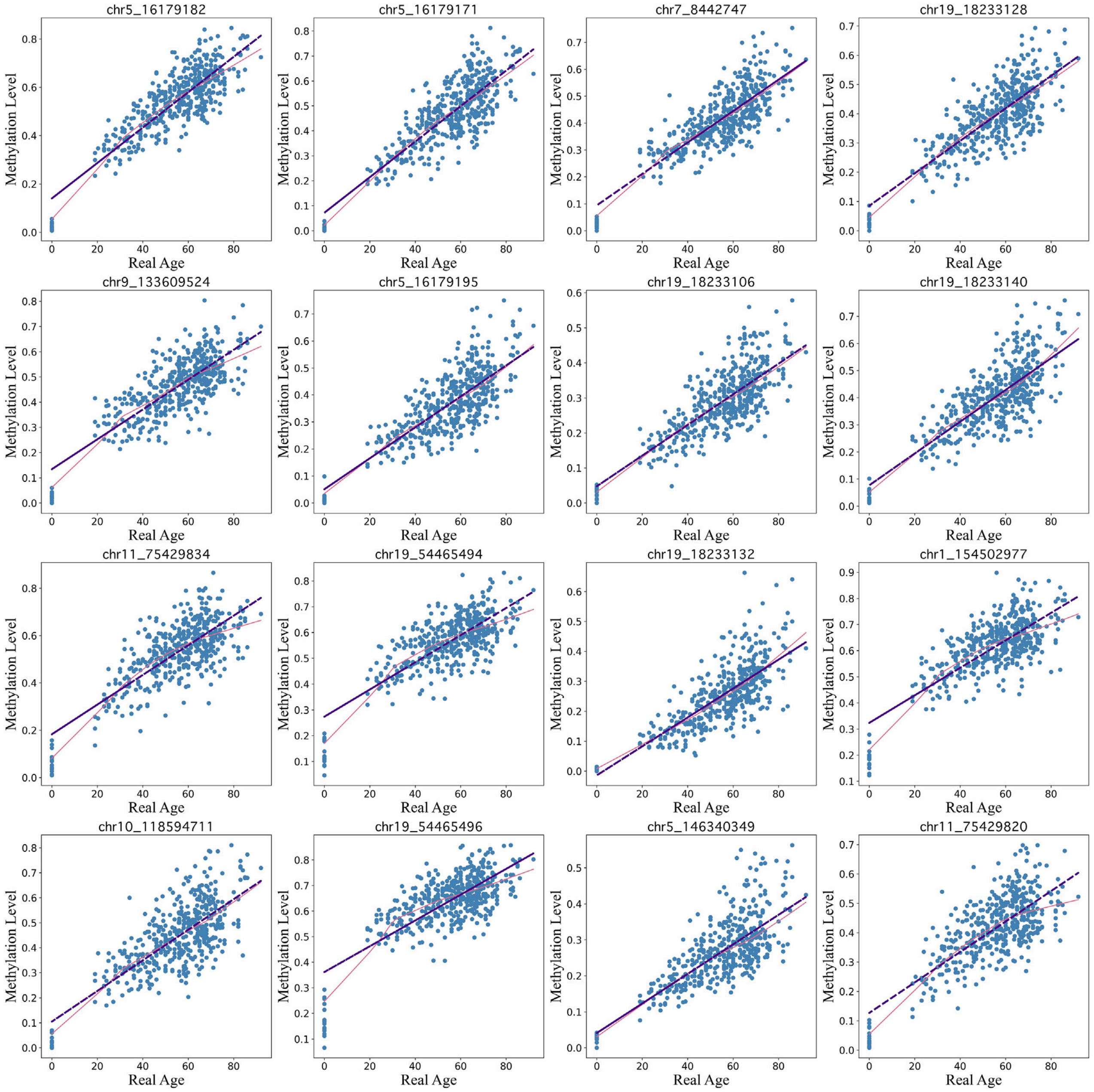

FIGURE 3
Top 16 CpG sites fitted using lowess regression (pink line) with a tau parameter value of 0.7 vs linear regression (purple line)..

73%, while scAge lags slightly behind with 70%. It is worth noting that, across the varying numbers of CpG sites assessed, scAge consistently generated age predictions with slightly lower mean absolute errors relative to BayesAge.

While accurately predicting the age of an individual is an important component of these models, another important metric is whether the errors in the prediction show an age bias. In human studies epigenetic clocks are often used to measure the difference between epigenetic and actual age, and many studies have shown that these differences are associated with disease and longevity. However, in order for these age differences to be interpretable, it is important that they do not demonstrate an age dependence. In other words, it is useful for models to generate residuals that are uncorrelated with age.

For this reason we evaluated the residuals of both scAge and BayesAge to identify any age-related biases. Figure 6 shows that the LOWESS fit of scAge, with a tau setting of 0.9, has a nonlinear residual pattern as age varies. By contrast, the BayesAge model was devoid of such age-associated biases in its residuals. The absence of discernible age associated residual patterns is another advantage of the non-linear BayesAge approach over scAge.

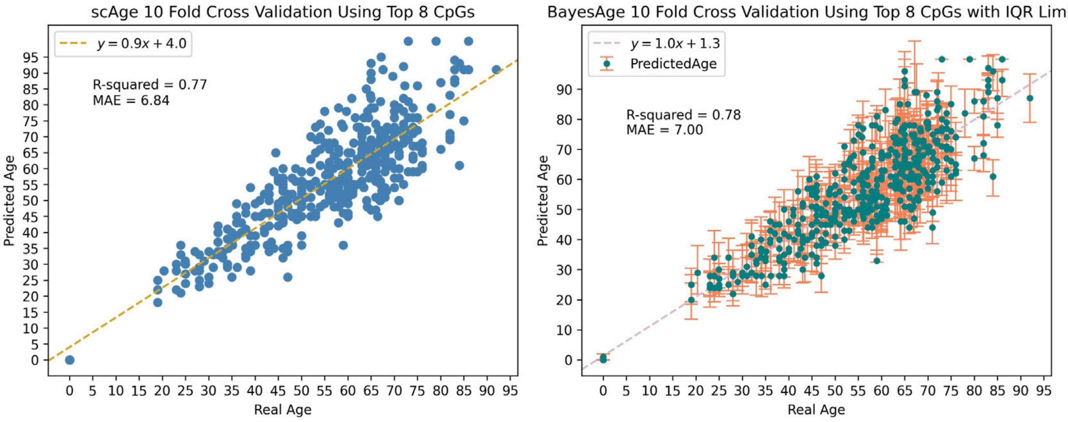

FIGURE 4
Data simulation was used to generate the IQR limits seen in BayesAge's plot with a tau value of 0.5.

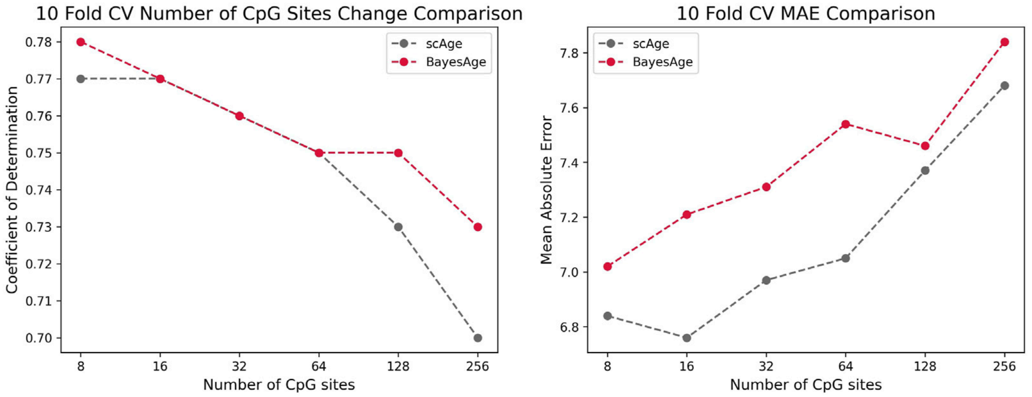

FIGURE 5
Comparison of the r-squared and MAE of scAge vs. BayesAge on a selected number of CpG sites.

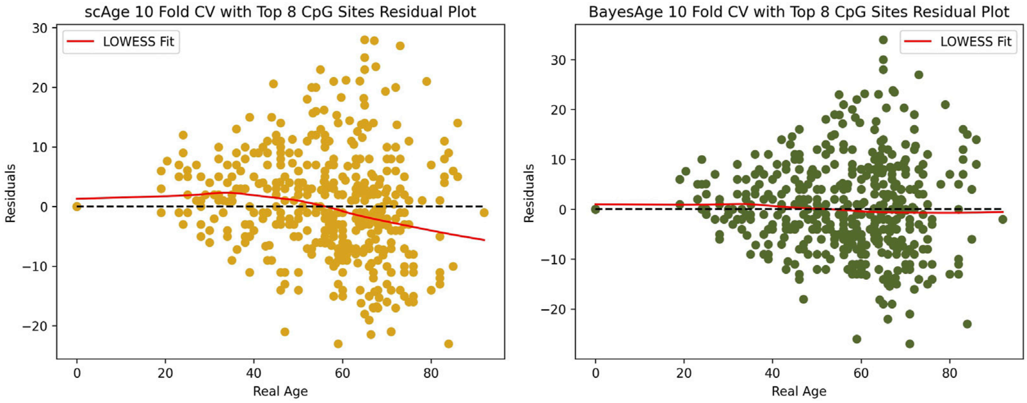

FIGURE 6
Residual plot analysis of scAge compared to BayesAge.

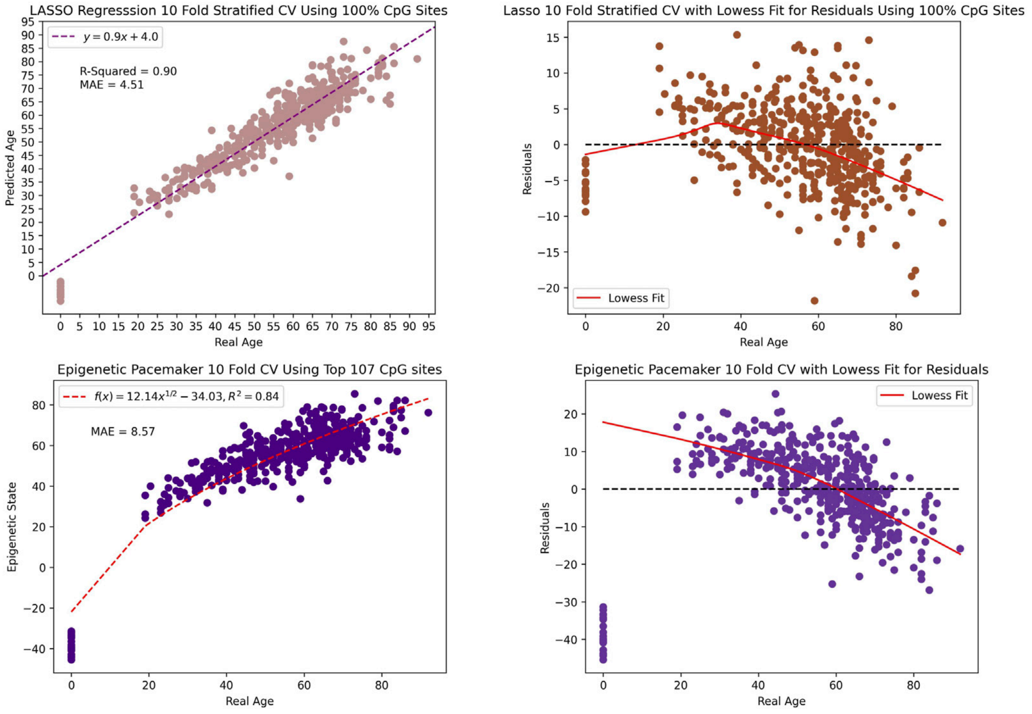

FIGURE 7
The scatter plot and residual plot of the LASSO and EPM model. The tau value of the lowess fit for the residual is 0.9.

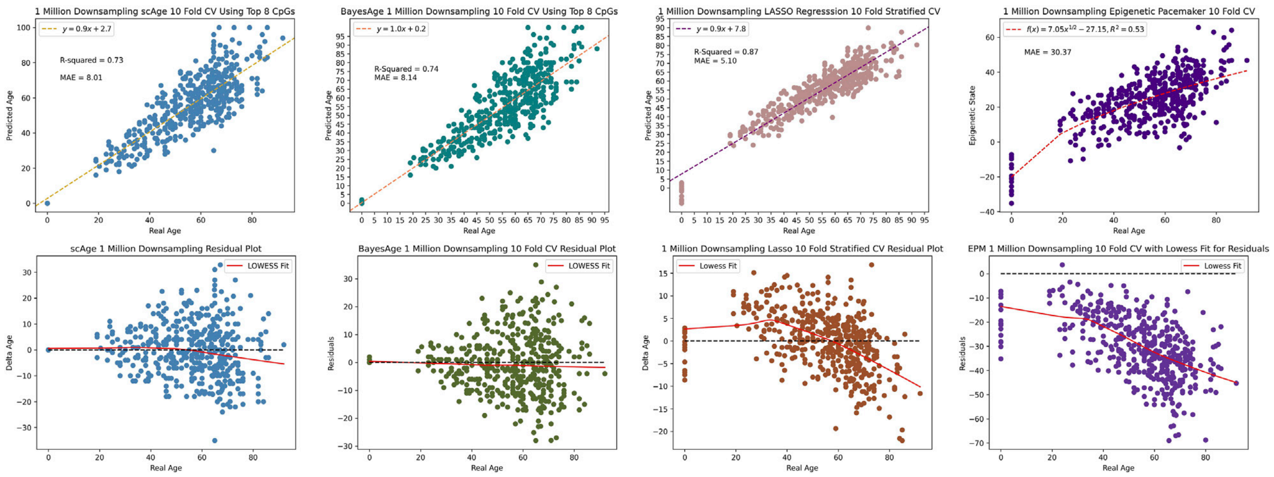

FIGURE 8
1 million downsampling comparison of scAge, BayesAge, LASSO regression and the Epigenetic Pacemaker respectively. The lowess fit has a tau value of 0.9.

Most of the previously published epigenetic clocks have been constructed using regularized regression which constrains the coefficient estimates to zero. Therefore, for comparative purposes, we implemented a LASSO (Least Absolute Shrinkage and Selection Operator) regression and the EPM model and trained on our data. Utilizing repeated 10-fold cross-validation, LASSO attained an r^2 value of 90% and a MAE of 4.51 when including all the 46,519 CpG sites in our dataset. In constract, the EPM model reached an r^2 value of 84% and a MAE of 8.57 when using the top 107 CpG sites. Thus the age estimation of the LASSO and EPM model outperforms both BayesAge and scAge. However, this comes at a cost of creating a significant bias in the age predictions. This bias is evident as the observed residuals display an age-associated trend that was more

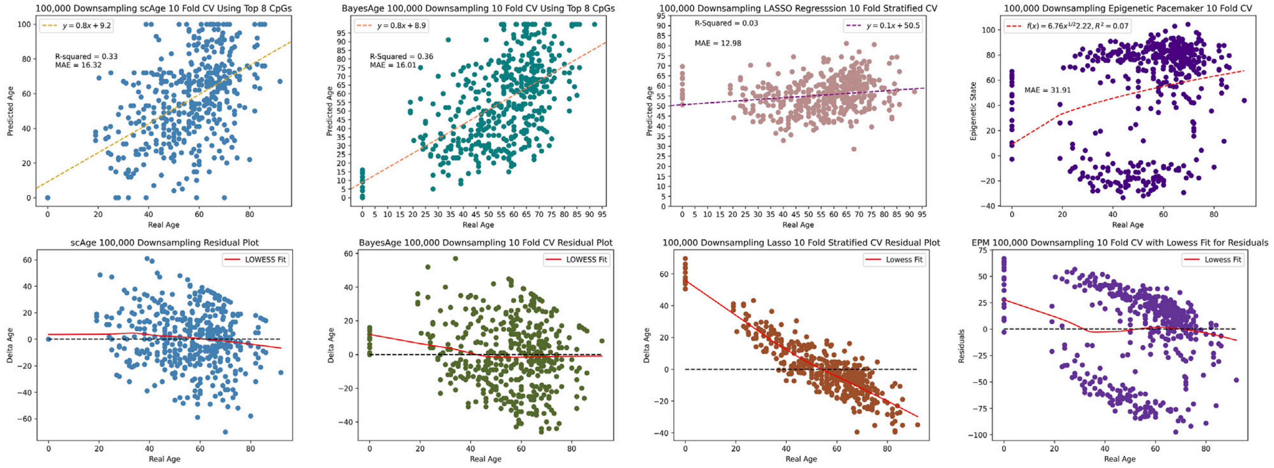

FIGURE 9
100,000 downsampling comparison of scAge, BayesAge, LASSO and the Epigenetic Pacemaker. LASSO shows a higher bias compared to the MLE models.

pronounced than in scAge. Thus while the age predictions of the LASSO and EPM model may be accurate, the residuals, or differences between actual and predicted age, show very strong biases that makes them difficult to interpret as seen in Figure 7.

We used one final metric to evaluate the performance of these four methods, and that involves the measurements of the robustness of the predictions as data is down-sampled. This is important, as bisulfite sequencing data-sets often have varying degrees of coverage. To evaluate model performance at reduced coverage levels, the original bisulfite sequencing data was subjected to random downsampling. At 1 million CpG sites, BayesAge's R^2 was marginally better at 74%, compared to scAge's 73%. Despite this, scAge recorded a slightly superior MAE of 8.01, while BayesAge returned 8.14 as seen in Figure 8. Notably, the LASSO model outperformed both the MLE models and the EPM model in terms of r^2 and MAE. The EPM model had a noticeably higher MAE of 30.37 compared to the other models. Additionally, the age-associated biases in residuals were distinctly evident for LASSO and the EPM model, more so than the other MLE techniques. Among the four methods, BayesAge demonstrated the least age-related biases.

Upon further reduction to 100,000 reads, BayesAge outperformed the other three methods with an R^2 value of 36% and MAE of 16.01, in comparison to scAge's respective metrics of 33% and 16.32. The LASSO and EPM models were the least resilient to this extreme downsampling, recording a considerably diminished R^2 value of 3% and 7% and an elevated MAE of 12.98 and 31.91 respectively as seen by Figure 9. Analyzing the residuals revealed significant age-related biases in the LASSO and EPM model, in stark contrast to the more consistent patterns observed in scAge and BayesAge models. This comparative analysis, even with signficantly reduced methylation data coverage, underscores BayesAge's capability to maintain accuracy and limit systemic biases. This positions BayesAge as a useful tool for epigenetic age prediction, particularly in cases with limited data coverage.

3.3 Computational efficiency

On the computational front, the time it takes to process 46 cgmap files using `load_cgmap_file` to a format compatible with the prediction function of BayesAge is on average 1 minute. The time to construct a reference using BayesAge's `construct_reference` function using 412 subjects takes on average 200 s. The time to predict the age of 46 subjects using BayesAge's `bdAge` function is on average 4.5 s.

scAge, just like BayesAge, is fully functional on a single core. Since BayesAge is a direct extension of scAge, both models experience linear speedup with multiprocessing.

The 10-Fold Cross Validation using the LASSO model takes around 6 min. The EPM package is implemented along with a conditional expectation maximization algorithm to efficiently estimate the parameters of the model. As such, it takes 10 s to implement the 10-Fold Cross Validation to our dataset.

4 Discussion

We introduced BayesAge, a maximum likelihood estimation framework for predicting epigenetic age from DNA methylation data. BayesAge addresses several limitations found in previous epigenetic age estimation methods such as penalized regression and the Epigenetic Pacemaker. It uses a LOWESS smoothing method to model nonlinear trends in methylation patterns with age, avoiding potential biases that can arise from linear assumptions. The BayesAge model is designed for count-based bisulfite sequencing data. By using a binomial distribution, it effectively models methylation probabilities at each CpG site, accounting for variable coverage depths across different sites and individuals. Notably, BayesAge maintains its performance even in the presence of significant data downsampling.

Our application of age prediction methods to a dataset of 458 individuals indicates that BayesAge generates comparably accurate prediction to the scAge MLE method. However, the residuals of

BayesAge show limited age-associated biases, suggesting that our model reflects biological differences in aging that are not correlated with age. This is an important property, as most human epigenetic clock studies are focused on age acceleration rather than age prediction. The fact that many models produce age acceleration estimates that are age associated with age, which confounds their interpretation, and makes it difficult to identify factors that impact the rate of epigenetic aging. By contrast, BayesAge residuals are not age associated and therefore may be more useful for identifying moderators of epigenetic aging.

Another important consideration of epigenetic clocks is that they produce a point estimate of an individual's age, and it is not possible to obtain a confidence interval on that estimate that accounts for the uncertainty of predictions. To overcome this limitation, we have implemented a simulation framework that allows us to model the range of age estimates that can be generated from a single sample. This simulated data suggests that BayesAge's predictions have an uncertainty using interquartile range of around 12 years.

In our evaluation of four methods for epigenetic age prediction, we focused on measuring the robustness of predictions as data was down-sampled, a critical consideration given the varying coverage levels often seen in bisulfite sequencing data. When reducing data to 1 million CpG sites, BayesAge displayed a slightly higher R^2 at 74% compared to scAge's 73%, while scAge had a marginally better MAE at 8.01, *versus* BayesAge's 8.14. The LASSO model outperformed the MLE methods and the EPM model in R^2 and MAE but exhibited noticeable age-associated biases in residuals. In extreme down-sampling to 100,000 sites, BayesAge surpassed scAge with a higher R^2 (36% vs 33%) and lower MAE (16.01 vs 16.32), while the LASSO and EPM models' performance deteriorated significantly (3% R^2, MAE 12.98% and 7%, MAE 31.91) with prominent age-related biases. Even with reduced data, BayesAge consistently maintained high accuracy and minimized biases, positioning BayesAge as a robust tool for epigenetic age prediction, especially in low-coverage scenarios.

The limitations of our binomial distribution age model provide potential avenues for further research in epigenetic aging frameworks. Future studies might consider using beta-binomial distributions to accommodate overdispersion in methylation probabilities. The model can also be expanded to include biological covariates known to impact methylation, such as gender or smoking habits. On the computational front, implementing optimization methods like expectation-maximization algorithms could enhance efficiency for larger epigenome-wide datasets.

In conclusion, BayesAge offers a comprehensive tool for exploring epigenetic aging dynamics. By addressing the challenges of previous models, BayesAge holds promise for enhancing understanding of aging trajectories across populations. This can lead to insights into factors influencing epigenetic aging, with potential applications in various research areas, from forensics to disease studies.

Data availability statement

The data presented in this study are deposited in the Gene Expression Omnibus (GEO) database, accession number GSE261769.

Author contributions

LM: Conceptualization, Data curation, Formal Analysis, Investigation, Methodology, Software, Validation, Visualization, Writing–original draft, Writing–review and editing. LR: Data curation, Methodology, Writing–review and editing. MT: Data curation, Methodology, Writing–review and editing. LB: Supervision, Writing–review and editing. MP: Conceptualization, Methodology, Supervision, Writing–review and editing.

Funding

The author(s) declare that no financial support was received for the research, authorship, and/or publication of this article.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fbinf.2024.1329144/full#supplementary-material

References

Colin, F., Tosevska, A., Oyetunde, A., and Pellegrini, M. (2021). Bisulfite bolt: a bisulfite sequencing analysis platform. *Gigascience* 10, giab033. doi:10.1093/gigascience/giab033

Farrell, C., Snir, S., and Pellegrini, M. (2020). The Epigenetic Pacemaker: modeling epigenetic states under an evolutionary framework. *Bioinformatics* 36, 4662–4663. doi:10.1093/bioinformatics/btaa585

Galkin, F., Mamoshina, P., Kochetov, K., Sidorenko, D., and Zhavoronkov, A. (2021). Deepmage: a methylation aging clock developed with deep learning. *Aging Dis.* 23, 1252. doi:10.14336/ad.2020.1202

Greenwood, C., Youssef, G., Letcher, P., Macdonald, J., Hagg, L., Sanson, A., et al. (2020). A comparison of penalised regression methods for informing the selection of predictive markers. *PLOS ONE* 15, e0242730. doi:10.1371/journal.pone.0242730

Hannum, G., Guinney, J., Zhao, L., Zhang, L., Hughes, G., Sadda, S., et al. (2013). Genome-wide methylation profiles reveal quantitative views of human aging rates. *Mol. Cell* 49, 359–367. doi:10.1016/j.molcel.2012.10.016

Horvath, S. (2013). Dna methylation age of human tissues and cell types. *Genome Biol.* 14, R115. doi:10.1186/gb-2013-14-10-r115

Johnson, A., Akman, K., Calimport, S., Wuttke, D., Stolzing, A., and de Magalhães, J. (2012). The role of dna methylation in aging, rejuvenation, and age-related disease. *Rejuvenation Res.* 15, 483–494. doi:10.1089/rej.2012.1324

Kabacik, S., Lowe, D., Fransen, L., Leonard, M., Ang, S.-L., Whiteman, C., et al. (2022). The relationship between epigenetic age and the hallmarks of aging in human cells. *Nat. Aging* 2, 484–493. doi:10.1038/s43587-022-00220-0

Levine, e. a., Morgan, E., Quach, A., Chen, B. H., Assimes, T. L., Bandinelli, S., et al. (2018). An epigenetic biomarker of aging for lifespan and healthspan. *Aging (Albany NY)* 10, 573–591. doi:10.18632/aging.101414

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., et al. (2009). The sequence alignment/map format and samtools. *Bioinformatics* 25, 2078–2079. doi:10.1093/bioinformatics/btp352

Lu, e. a., Ake, T., Wilson, J. G., Reiner, A. P., Aviv, A., Raj, K., et al. (2019). Dna methylation grimage strongly predicts lifespan and healthspan. *Aging* 11, 303–327. doi:10.18632/aging.101684

Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.J.* 17, 10. doi:10.14806/ej.17.1.200

Moore, L., Le, T., and Fan, G. (2012). Dna methylation and its basic function. *Nat. Neurpsychopharmacology* 38, 23–38. doi:10.1038/npp.2012.112

Morselli, M., Farrell, C., Rubbi, L., Fehling, H. L., Henkhaus, R., and Pellegrini, M. (2021). Targeted bisulfite sequencing for biomarker discovery. *Methods* 187, 13–27. doi:10.1016/j.ymeth.2020.07.006

Snir, S., Farrell, C., and Pellegrini, M. (2019). Human epigenetic ageing is logarithmic with time across the entire lifespan. *Epigenetics* 14, 912–926. doi:10.1080/15592294.2019.1623634

Trapp, K. C., and Gladyshev, V. (2021). Profiling epigenetic age in single cells. *Nat. Aging* 1, 1189–1201. doi:10.1038/s43587-021-00134-3