Journal of Zhejiang University-SCIENCE B (Biomedicine & Biotechnology) in press
www.jzus.zju.edu.cn; www.springer.com/journal/11585
E-mail: jzus_b@zju.edu.cn

Research Article

https://doi.org/10.1631/jzus.B2500398

Development of epigenetic clocks for age estimation in human sperm and semen: Multi-Platform discovery and forensic validation

Ming ZHAO[1,2], Fanzhang LEI[1], Meiming CAI[1], Qinglin LIANG[1], Xi YUAN[1], Qiong LAN[1]✉, Yating FANG[3]✉, Bofeng ZHU[1]✉

[1]*Guangzhou Key Laboratory of Forensic Multi-Omics for Precision Identification, School of Forensic Medicine, Southern Medical University, Guangzhou 510515, China*
[2]*School of Forensic Medicine, Kunming Medical University, Kunming 650500, China*
[3]*School of Basic Medical Sciences, Anhui Medical University, Hefei 230031, China*

Abstract: Accurate age estimation from semen evidence is crucial for forensic investigations in sexual assault cases. While DNA methylation is a promising biomarker for predicting the donor's chronological age in forensic cases, most existing DNA methylation-based age estimation models primarily focus on somatic cells, with limited exploration of sperm-specific methylation signatures. Given that tissue-specific differences in CpG methylation may reduce the accuracy of existing epigenetic clocks for semen samples, there is a need to develop age-prediction models for this tissue in particular. For this study, we employed publicly available sperm methylation microarray datasets (GSE185920, n = 1471, aged 20-60 years) from the Gene Expression Omnibus (GEO) to identify age-related CpG sites (AR-CpGs). To identify AR-CpGs, we subsequently implemented a multi-algorithm feature selection strategy (maximum mutual information, L1 regularization, and sequential feature selection). We developed an optimized sperm-epigenetic clock by evaluating 69 machine learning regression model frameworks, achieving a mean absolute error (MAE) of 1.63 years in the training cohort. Validation on independent sperm datasets (GSE185445, n = 379, GSE149318, n = 90) yielded MAEs of 2.93 and 2.58 years, respectively, demonstrating robust generalization. To identify additional markers, we screened for sperm-specific AR-CpGs using whole-genome bisulfite sequencing (WGBS) data from the publicly available GEO dataset GSE222340. Subsequently, based on the pyrosequencing data of nine selected AR-CpG markers analyzed in 95 semen samples (ages 20–42 years), we developed a robust forensic model for human semen age estimation and determined the optimal algorithm by systematically evaluating 23 regression methods. The best-performing model, support vector machine (radial basis function kernel), exhibited an MAE of 2.21 years and a root mean square error (RMSE) of 3.15 years on the test set. This work provides a valuable set of AR-CpGs, develops an optimized sperm-chronological epigenetic clock, and delivers a practical model for estimating age from semen.

Key words: Age-related CpG; Sperm chronological epigenetic clock; Age estimation; Semen age prediction model

1 Introduction

Age estimation from biological evidence plays a crucial role in forensic investigations, offering valuable information for solving crimes and guiding judicial decision-making. When traditional DNA profiling methods fail due to DNA database mismatches, obtaining age information about the sample donor using forensic technology can significantly narrow down suspect lists, identify severely decomposed remains in the case of mass disasters, and supplement skeletal age assessments in cases involving juvenile delinquency. Traditional

✉ Qiong LAN, joan_lan1205@126.com
Qiong, LAN,https://orcid.org/0000-0001-9634-3603
✉ Yating FANG, fighting9216@163.com
Yating FANG, https://orcid.org/0000-0001-9108-8738
✉ Bofeng ZHU, zhubofeng7372@126.com
Bofeng ZHU, https://orcid.org/0000-0002-9038-2342

Received July 10, 2025; Revision accepted Nov. 26, 2025;
Crosschecked xxx. xx, 20xx; Published online xxx. xx, 20xx

forensic age estimation methods primarily rely on anthropological theories that analyze morphological changes in skeletal and dental structures (Meissner et al., 1999; Niño-Sandoval et al., 2017; Thodberg et al., 2017; Lee et al., 2020). However, these methods are constrained by subjectivity and biological variability, particularly when applied to biological trace evidence, which necessitates the development of robust molecular alternatives.

Recent technological advances have addressed these constraints by facilitating the discovery of age-related biomarkers, including transcriptomic markers (Zubakov et al., 2016; Fleischer et al., 2018; Huan et al., 2018; Wang et al., 2022), mitochondrial DNA deletions (Meissner, et al., 1999; Tengan et al., 2002; Ro et al., 2003; Zapico and Ubelaker, 2016), telomere length alterations (Friedrich et al., 2001; Mather et al., 2011; Opstad et al., 2011; Sanders and Newman, 2013; Breitling et al., 2016; Marioni et al., 2016; Ruiz et al., 2017; Vasu et al., 2017; Wang et al., 2018), aspartic acid racemization (Griffin et al., 2008; Griffin et al., 2009; Arany and Ohtani, 2011; Rajkumari et al., 2013; Elfawal et al., 2015; Hassan et al., 2017), and DNA methylation (DNAm) (Christensen et al., 2009; Bell et al., 2012; Horvath, 2013). Among these biomarkers, DNAm stands out for its unparalleled forensic applicability, largely attributable to its inherent chemical stability and the resulting longer detection window in compromised samples (Jylhävä et al., 2017). Extensive studies have demonstrated strong correlations between DNAm patterns and chronological age. Furthermore, DNAm-based age estimation demonstrates superior accuracy, high inter-laboratory reproducibility, and broad applicability to diverse biological sample types. Collectively, these attributes establish DNAm as one of the most reliable and versatile molecular tools for estimating the age of unknown individuals in forensic casework (Teschendorff et al., 2010; Hannum et al., 2013).

Advances in epigenetic technologies have enabled the development of methylation-based age prediction models, known as chronological epigenetic clocks, that use age-related CpG sites (AR-CpGs). Examples of such clocks include the Horvath pan-tissue clock (Horvath, 2013) and the Hannum blood-specific clock (Hannum, et al., 2013). These models not only enhance our understanding of aging but also provide quantifiable tools for age estimation in forensic contexts. The discovery of age-related changes in methylation that are conserved across tissues has enabled the creation of pan-tissue clocks. In parallel, the tissue and body fluid specificity of DNA methylation patterns has driven the development of specialized epigenetic clocks for distinct biological sources (Theda et al., 2018; Jung et al., 2019; Naue, 2023). However, clocks developed primarily based on somatic cells may be unsuitable for application to sperm cells. This incompatibility arises because sperm cells undergo unique epigenetic reprogramming during meiosis, resulting in methylation patterns distinct from those of somatic cells (Eckhardt et al., 2006; Oakes et al., 2007; Cui et al., 2016). Additionally, the development of DNAm-based age estimation models using sperm cells or semen samples remains relatively limited compared with those using samples such as blood and saliva. There is therefore a clear need for epigenetic clocks optimized for male germ cells to improve age estimation from semen evidence in forensic investigations.

Current age estimation models primarily rely on AR-CpGs identified from methylation microarray data, such as those generated by Illumina Infinium HumanMethylation platforms (Alsaleh and Haddrill, 2019; Xiao et al., 2021). Although high-density microarrays (e.g., the EPIC v1.0 chip covering over 850,000 CpG sites) have been instrumental in epigenomic research (Moran et al., 2016; Pidsley et al., 2016), they only involve approximately 4% of the CpGs in the human methylome (Vidaki and Kayser, 2018). This limited genomic coverage leaves vast regulatory regions unexplored. Additionally, many microarray-derived AR-CpGs suffer from limited generalizability in forensic contexts due to issues such as tissue specificity, biological variability, and a paucity of methylation data from forensically relevant tissues. These limitations highlight the need to identify novel CpG markers for forensic age estimation.

Whole-genome bisulfite sequencing (WGBS) is a high-throughput sequencing method that enables genome-wide DNA methylation profiling at single-base resolution (Feng and Lou, 2019; Beck et al., 2022). Unlike methylation microarrays, which are restricted to predefined CpGs, WGBS can detect rare methylation variants and characterize epigenetic patterns in traditionally understudied regions, including intergenic regions and transposable elements (Li et al., 2018; Ortega-Recalde et al., 2021). This comprehensive capability makes

WGBS a powerful tool for epigenetic research, particularly for identifying novel AR-CpGs that cannot be detected by conventional microarray platforms (Wang et al., 2013; Suzuki et al., 2018).

Given the above considerations, the present study was designed with two aims. First, we analyzed public methylation microarray data (GSE185920) to construct a high-precision sperm epigenetic clock. Second, we used WGBS data (GSE222340) to explore the entire genome to identify novel AR-CpGs. We then selected the most promising age-related markers identified from these two analytical streams and validated them via pyrosequencing in an independent semen cohort. The ultimate goal was to integrate these findings to establish a robust, forensically applicable age prediction model for semen stains.

2 Materials and methods

2.1 Samples and datasets

2.1.1 Sample collection

This study was conducted in accordance with the ethical principles outlined in the Declaration of Helsinki and received formal approval from the Ethics Committee of Southern Medical University (Approval NO. 2023-KY-097-02). We enrolled 95 healthy male volunteers (aged 20-42 years) from the Chinese Han population who had no documented systemic diseases or a history of long-term medication use. In line with ethical guidelines, all participants received thorough explanations of study procedures and voluntarily provided signed informed consent before participation. Prior to semen sample collection, donors adhered to a minimum 72-hour abstinence period before self-collecting ejaculate specimens through masturbation, which were promptly preserved in pre-sterilized cryotubes at -80°C to preserve molecular integrity.

2.1.2 Dataset collection

DNA methylation datasets were sourced from the Gene Expression Omnibus (GEO) database of the National Center for Biotechnology Information (NCBI). The analysis included both Illumina Infinium HumanMethylationEPIC BeadChip microarray data (accessions: GSE185920, GSE185445, and GSE149318) and WGBS data (accession: GSE222340). These datasets were filtered to include only semen or sperm samples from healthy donors that had confirmed availability of chronological age metadata.

2.2 Development and validation of a sperm-specific DNA methylation clock for chronological age estimation

2.2.1 Processing and analysis of sperm microarray datasets

We selected the GSE185920 dataset for AR-CpG screening and epigenetic clock construction (Jenkins et al., 2022). Raw methylation data were preprocessed using RnBeads version 2.0.1 in R version 4.0.3. This involved four sequential quality control steps: probe filtering (detection p-values > 0.01), the exclusion of samples with > 10% failed probes, the application of a bead count threshold (< 5 beads in at least 5% of samples), and the removal of probes that were SNP-overlapping or sex chromosomal. Missing values were subsequently imputed using the k-nearest neighbors (k-NN) algorithm (Sahoo and Sundararajan, 2024), followed by Beta mixture quantile dilation (BMIQ) normalization to adjust for technical differences between Infinium I and II probes (Teschendorff et al., 2013). Potential batch effects were corrected using the ComBat method in the sva package (Leek et al., 2012).

Age-related CpGs (AR-CpGs) were identified through a two-step procedure: initial screening was conducted via linear regression ($\beta \sim$ chronological age) using the *dmrfinder()* function in the minfi package, applying a significance threshold of FDR-adjusted $p < 0.05$ (Aryee et al., 2014). Subsequently, candidate AR-CpGs were prioritized based on the absolute value of Spearman's rank correlation coefficient ($|\rho|$), where higher values indicated more robust monotonic relationships with increasing age.

2.2.2 Functional enrichment analysis and body fluid specificity validation of age-related CpGs

Functional enrichment analysis was conducted on the top 5000 sperm AR-CpGs ranked by absolute Spearman's ρ using the ChAMP package (version 2.21.1). We performed Gene Ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway analyses to preliminarily explore the potential biological mechanisms underlying these methylation changes. The body fluid specificity of AR-CpGs was validated using methylation microarray data (peripheral blood, semen, saliva, menstrual blood, and vaginal secretion samples) previously generated by our team (Fang et al., 2023).

2.2.3 Feature selection algorithms for sperm-specific AR-CpGs

To construct a sperm chronological epigenetic clock, we employed three distinct feature selection algorithms, including L1 regularization (Lasso) (Schmidt et al., 2007), maximal information coefficient (MIC) (Kinney and Atwal, 2014b), and sequential feature selection (SFS) (Aha and Bankert, 1995) using MATLAB version 2023b to identify an optimal subset of AR-CpGs. The input was a methylation beta-value matrix (1471 sperm samples × top 5000 AR-CpGs) derived from the GSE185920 dataset. This matrix was randomly divided into training (80%), validation (10%), and test (10%) sets. We adopted multiple linear regression as the unified base model across all algorithms to ensure a consistent, comparable framework for feature selection. An upper limit of 1000 CpG features was set for the iterative testing to mitigate overfitting and identify the point of performance saturation. This range is sufficient to cover the number of CpG sites used by most high-performance epigenetic clocks in the existing literature. Model performance was quantified for both training and test sets using the coefficient of determination (R^2), mean absolute error (MAE), and root mean square error (RMSE). The CpG subset from each method that achieved the optimal balance of high R^2 and the low MAE/RMSE was designated as the respective "best_set" (i.e., SFS_best_set, MIC_best_set, and L1_best_set). To assess the novelty of our identified AR-CpGs, we systematically compared our optimal CpG set against AR-CpGs reported in key published studies on sperm and semen. We used an online Venn diagram tool to visualize the overlap and clarify the distinct and shared CpGs.

2.2.4 Multiple machine learning regression algorithms for constructing a sperm chronological epigenetic clock

(1) Optimal AR-CpG set for a sperm chronological epigenetic clock

To ensure our evaluation was comprehensive and methodologically robust, we selected 23 regression algorithms from the Scikit-learn library based on their representativeness across major machine learning paradigms. This selection encompasses linear models (e.g., Ridge, Lasso) for their interpretability and efficiency with high-dimensional data; tree-based ensembles (e.g., Random Forest, Gradient Boosting) for their capacity to capture complex non-linear interactions; and support vector machines for their effectiveness in high-dimensional feature spaces. We then integrated three optimal AR-CpG sets with this diverse algorithmic portfolio to construct and compare candidate sperm chronological epigenetic clocks. The input methylation beta-value matrix (the optimal AR-CpG set × 1471 sperm samples) was split into training (80%) and test (20%) sets using a stratified sampling method based on Python version 3.10.4. Hyperparameters were optimized with Optuna version 1.4.0 over 100 trials per model, employing 5×10-fold cross-validation to minimize MAE on the validation folds and mitigate random bias. The final model was selected by synthesizing the performance on the test set (RMSE, MAE, R^2) with a Taylor diagram analysis of prediction deviations. This comprehensive evaluation prioritized balancing the number of requisite CpG loci with predictive accuracy.

(2) Evaluation of the sperm-specific chronological epigenetic clock

To evaluate the generalizability of our sperm-specific chronological epigenetic clock, we conducted external validation on two independent datasets (GSE185445: 379 sperm samples; GSE149318: 90 sperm samples). We benchmarked the predictive performance of our model against established epigenetic clocks, including the Horvath pan-tissue clock, the skin-blood clock, and sperm-specific clocks, using these same

datasets to ensure a direct and fair comparison.

2.3 Semen age prediction model construction using WGBS data

2.3.1 Processing and analysis of the WGBS dataset

Sperm WGBS data from the GEO dataset GSE222340 were analyzed using the Dnmtools v1.4.2 package in R version 4.1.3 on a Linux platform. We employed WGBS technology to systematically analyze the global variation characteristics of the human sperm methylome across a temporal scale. Our objective was to deeply explore the combined influence of individual specificity and the aging process on the epigenetic landscape of sperm. The dataset included 20 sperm samples obtained from 10 healthy, fertile male donors (aged 23–56 years at initial sampling). A second sample was collected from each donor 10–18 years later. Raw methylation data (.meth files) were processed into methylation level matrices, with each entry representing the percentage of cytosine methylation and corresponding read coverage. We grouped samples into two batches (Batch A and Batch B) based on filename identifiers (a and b), which corresponded to distinct experimental time points. Quality control involved removing CpGs with missing methylation values or coverage below 5. To ensure robustness across technical replicates, we computed Spearman's rank correlation between methylation levels for each batch separately and defined AR-CpGs as those exhibiting a significant correlation ($p < 0.01$) in both batches.

The inherent limitations of our WGBS discovery approach should be noted. WGBS is known for its variable sequencing depth, which can result in incomplete CpG coverage and less reliable methylation quantification. Furthermore, the epigenome-wide association study (EWAS) was conducted with a very small sample (n=10), severely limiting the statistical power and robustness of the identified AR-CpGs. Therefore, the AR-CpGs identified in this phase should be considered exploratory candidates requiring rigorous validation, which we subsequently performed using pyrosequencing.

2.3.2 Selection of a minimal marker panel for forensic application

To develop a robust detection system for forensic practice, we screened AR-CpGs from two complementary sources: (i) those exhibiting the highest absolute Spearman's correlation coefficients ($|\rho|$) in the microarray dataset (GSE185920), and (ii) those demonstrating significant age correlations in an independent WGBS dataset (GSE222340). Subsequently, primers were designed for candidate CpGs, and a panel of nine markers was selected based on a combined evaluation of their statistical association with age and the practical feasibility of primer design. This minimal panel was then carried forward for pyrosequencing-based validation.

2.3.3 DNA extraction, quantification, and bisulfite conversion

Genomic DNA was extracted from 50 μL of each semen sample using the MicroElute Genomic DNA Kit (D3096–02, Omega Bio-tek, Inc., Norcross, Georgia, USA) according to the manufacturer's instructions. DNA quantification was performed using a QubitTM 4 fluorometer (Thermo Fisher Scientific, USA) with the Qubit dsDNA HS Assay Kit (Thermo Fisher Scientific, USA) according to the manufacturer's instructions. The extracted genomic DNA (100 ng) was bisulfited using the EZ DNA Methlyation-DirectTM Kit (Zymo Research, Irvine, CA), also following the provided instructions. The converted DNA was then eluted with 30 μL of preheated M-Elution Buffer, yielding the final bisulfite-converted DNA for subsequent analysis.

2.3.4 Pyrosequencing

Primers targeting AR-CpGs were designed using PyroMark Assay Design 2.0 software and were synthesized by Sangon Biotech (Shanghai, China). Each 25 μL polymerase chain reaction (PCR) mixture contained 4 μL of PyroMark PCR Master Mix (2×), 2.5 μL of CoralLoad Concentrate (10×), 2 μL of each PCR primer (2.5 μM), and 2 μL of bisulfite-converted DNA, brought to the final volume with sterile, deionized water. PCR amplification was performed using the PyroMark PCR Kit (Qiagen) on a GeneAmp PCR System 9700 (Applied Biosystems) under standardized conditions: initial denaturation at 95°C for 15 min; 45 cycles of denaturation (94°C, 30 s), annealing (56°C, 30 s), and extension (72°C, 30 s); and final extension at 72°C for 10 min. Pyrosequencing was conducted on the PyroMark Q24 system (Qiagen) with PyroMark Gold CpG Reagents, strictly adhering to the manufacturer's instructions. Methylation levels were quantified and visualized using the Hiplot (https://hiplot.cn/).

2.3.4 Construction and evaluation of the semen age estimation model

We developed and evaluated 23 regression models from the Scikit-learn library in Python (v3.10.4) using the pyrosequencing data of nine AR-CpGs from 95 sperm samples. The data were stratified by chronological age and randomly split into training (80%) and test (20%) sets. The semen age prediction model was constructed using the sperm epigenetic clock framework. Based on a comprehensive assessment of multiple performance metrics, we have made the optimal model available on GitHub as an open-source, user-friendly software (https://github.com/zuigaoming/DNAmethylation.git).

3 Results

3.1 Development and validation of a sperm-specific epigenetic clock

3.1.1 Analysis of sperm methylation microarray data

The analysis commenced with three methylation microarray datasets identified from the NCBI GEO database using the search terms “EPIC”, “Sperm”, and “Semen” and predefined criteria. These datasets included GSE185920 (1471 sperm samples from males aged 20–60 years), which was derived from a randomized clinical trial investigating the effects of folic acid and zinc supplementation (FAZST) on sperm DNA methylation (Jenkins, et al., 2022); GSE185445 (379 sperm samples from males aged 19–40 years), derived from a study aiming to develop a sperm epigenetic clock for predicting pregnancy outcomes (Pilsner et al., 2022); and GSE149318 (90 paired blood–sperm samples from males aged 22–51 years), which was collected for research comparing the methylation profiles of obese and non-obese individuals (Åsenius et al., 2020). As summarized in Fig. 1, the mean ages of donors for each dataset were 32.8 years (GSE185920), 29.9 years (GSE185445), and 35.2 years (GSE149318).

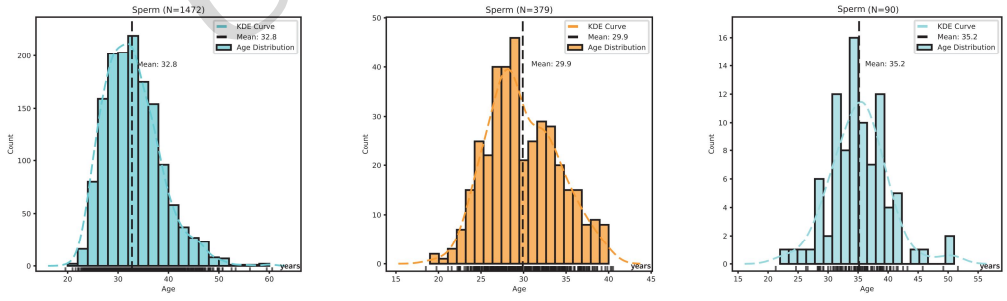

Fig. 1 Age distribution histograms and kernel density curves of samples from three DNA methylation microarray datasets. From left to right, the histograms are GSE185920 (sperm sample, n=1471), GSE185445 (sperm sample, n=379), and GSE149318 (sperm sample, n=90).

The GSE185920 dataset was employed for AR-CpG screening. Following data preprocessing and quality control (Fig. S1), 721,063 probes and all 1471 samples were retained according to the inclusion criteria. We performed an age-correlation analysis using the *dmpFinder()* function (minfi package) via univariate linear regression, identifying 248,097 significant AR-CpGs (adjusted $p < 0.01$). We selected the top 5000 markers ($0.37 < |\rho| < 0.58$) for downstream analysis (Table S1) based on their absolute Spearman's correlation coefficient from a subsequent analysis, which identified 195,543 significant CpGs (adjusted $p < 0.01$; $1.54E\text{-}05 < |\rho| < 0.58$).

The Circos plot illustrated the distribution of these 5000 AR-CpGs across 22 autosomes (Fig. S2). Track A

employed a heatmap to visualize a genome-wide methylation pattern associated with aging, while Tracks B–E illustrate directional methylation evolution trends within specific age groups. We further visualized the top 20 AR-CpGs ranked by absolute Spearman coefficients ($|\rho| > 0.50$, $p < 0.01$) using methylation-age scatterplots (Fig. 2). Among these, cg18037145 exhibited the strongest correlation with age ($\rho = 0.58$, $p = 2.2e\text{-}10$), indicating high predictive potential.

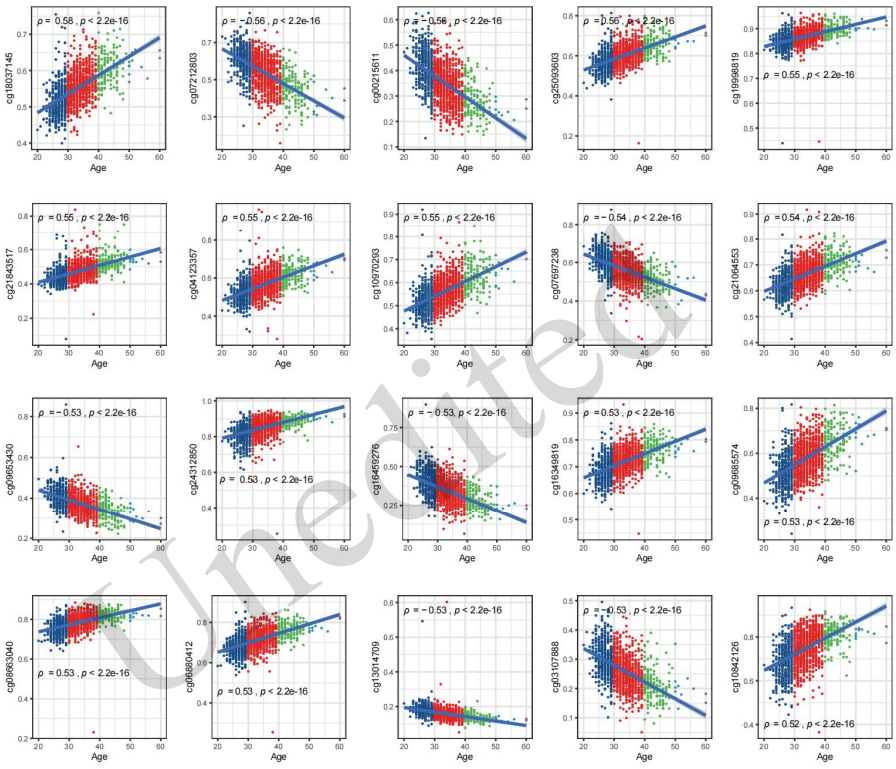

Fig. 2 Scatter plots showing the relationship between methylation levels (β-values) and chronological age for the top 20 AR-CpGs. The blue line represents the linear regression fit. ρ represents Spearman's rank correlation coefficient.

3.1.2 Gene functional annotation of sperm AR-CpGs and enrichment analysis of KEGG and GO functions

Functional and positional annotation of the top 5000 AR-CpGs revealed distinct DNA methylation patterns (Fig. S3, Table S1). Positionally, these sites were predominantly found in intergenic regions (37.91%) and gene bodies (40.62%) (Fig. S3A). The majority of AR-CpGs were located in open sea regions (67.07%) (Fig. S3B). KEGG pathway enrichment analysis identified significant associations with immune-metabolic regulation, cellular signaling, and structural organization. In parallel, GO term enrichment analysis exhibited associations with biological processes including cellular morphogenesis, membrane dynamics, and ion transport (Fig. S3C, S3D).

3.1.3 Evaluation of body fluid specificity of sperm AR-CpGs

The body fluid specificity of the top 100 AR-CpGs based on their absolute Spearman's correlation coefficient was analyzed using 850K methylation microarray data from five forensically relevant body fluids. Semen samples exhibited a distinct methylation signature at these AR-CpGs, distinguishing them from those of other body fluids (Fig. S4A). Analysis of the 1471 sperm samples revealed age-dependent methylation changes

(Fig. S4B). Furthermore, the methylation profiles of semen samples aligned more closely with those of sperm than with those of other body fluid samples.

3.1.4 Feature selection of sperm AR-CpGs

To develop a sperm-specific epigenetic clock, we performed feature selection on the top 5000 AR-CpGs using three machine learning algorithms (MIC, Lasso, SFS). Candidate CpG subsets were evaluated using multivariate linear regression, and the optimal combination from each method was designated as its respective "best_set" based on performance metrics. We identified three optimal feature sets based on these metrics. The MIC method demonstrated stable model performance on 470 AR-CpGs, achieving test-set metrics of R2 = 0.71, MAE = 2.21 years, and RMSE = 2.90 years (Fig. S5A). In comparison, the Lasso algorithm achieved optimal predictive performance on 464 CpGs, with test-set values of R^2 = 0.88, MAE = 1.71 years, and RMSE = 2.31 years (Fig. S6A). The SFS method exhibited peak efficiency using 261 CpGs, yielding test-set results of R^2 = 0.8331, MAE = 1.71 years, and RMSE = 2.31 years (Fig. 3A). Comparative analysis of test-set performance revealed that Lasso_best_set demonstrated superior predictive accuracy, while SFS_best_set contained the fewest AR-CpGs, and MIC_best_set exhibited the lowest test-set performance despite its size.

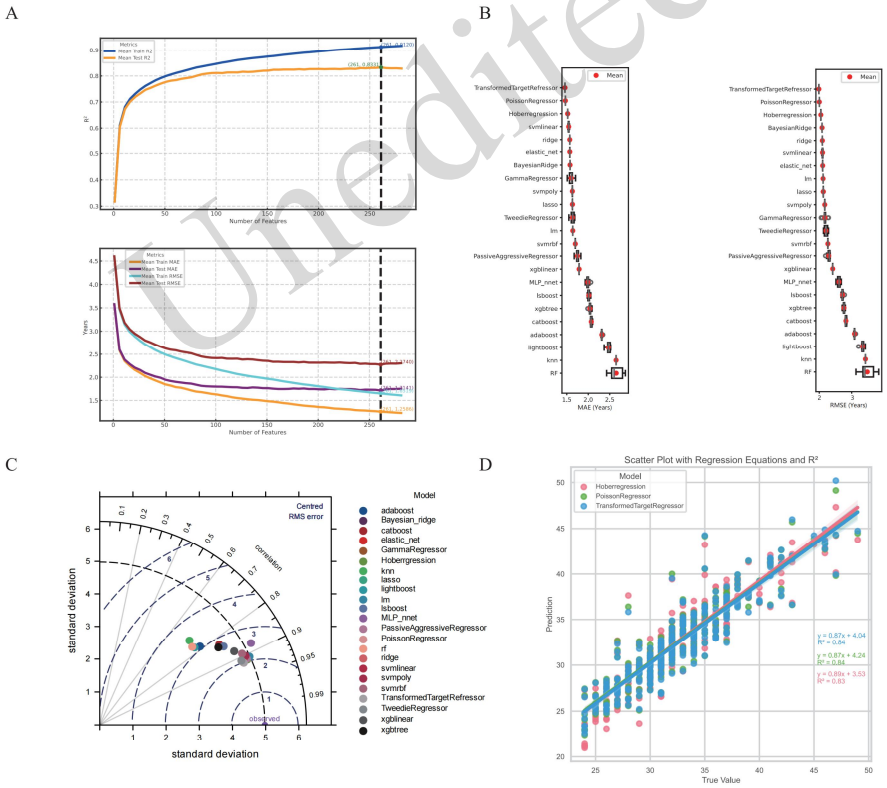

Fig. 3 Construction and evaluation of the sperm epigenetic clock based on a sequential feature selection algorithm. A sequential feature selection algorithm was used to select the top 5000 sperm AR-CpGs ranked by the absolute value of Spearman correlation coefficient ρ. When the specific combination of CpGs was 261, (A) the performance of the training set and test set of the multiple linear regression model showed no significant improvement; (B) Box plots of the mean absolute deviation and root mean square error of the test sets of 23 machine learning regression models; (C) Taylor diagrams of the test sets of 23 machine learning regression models based on 261 CpG markers; (D) Scatter plots of the test sets of the top three optimal models and the true values.

We evaluated 23 machine learning regression models across three feature sets (MIC_best_set, L1_best_set, SFS_best_set). For MIC_best_set, the models yielded MAE values ranging from 1.94 to 2.70 years and RMSE values ranging from 2.61 to 3.41 years, with SVMPoly demonstrating optimal performance (MAE = 1.94, RMSE = 2.61 years) (Fig. S5B). For L1_best_set, models yielded MAE values ranging from 1.37 to 2.62 years and RMSE values ranging from 1.90 to 3.40 years, with PoissonRegressor demonstrating optimal performance (MAE = 1.37, RMSE = 1.90 years) (Fig. S6B). For SFS_best_set, models yielded MAE values ranging from 1.46 to 2.86 years and RMSE values ranging from 1.99 to 3.83 years, with Transformed Target Regressor demonstrating optimal performance (MAE = 1.46, RMSE = 1.99 years) (Fig. 3B).

Taylor diagrams were employed to evaluate regression model performance across the three optimal feature sets (Fig. 3C, S5C, S6C). Each diagram plots models as points relative to an "Observed" reference point on the x-axis, which represents the standard deviation of the true age distribution in the test set. The diagrams assess three performance metrics: (1) correlation coefficient, indicated by angular proximity to the reference point; (2) standard deviation agreement, shown by radial alignment with the reference circle; and (3) centered RMSE, represented by Euclidean distance from the reference point. Optimal models cluster in the lower-right region near the reference point, exhibiting high correlation, matched standard deviation, and minimal centered RMSE, whereas inferior models are displaced toward the upper-left region. For MIC_best_set, three models (XGBTree, Transformed Target Regressor, SVMPoly) showed balanced performance with superior composite metrics (Fig. S5C), yielding test-set R^2 values ranging from 0.71 to 0.72 (Fig. S5D). For L1_best_set, three models (PoissonRegressor, Transformed Target Regressor, Tweedie Regressor) outperformed all others (Fig. S6C), each achieving identical test set R^2 values of 0.85 (Fig. S6D). For SFS_best_set, the three top-performing models (Transformed Target Regressor, PoissonRegressor, HuberRegressor) achieved superior metrics (Fig. 3C), yielding test-set R^2 values of 0.83–0.84 (Fig. 3D).

We evaluated regression models across optimal CpG sets derived from three feature selection methods (MIC, SFS, Lasso), comparing the performance of the top three models from each method (Table 1). For L1_best_set (464 AR-CpGs), the PoissonRegressor demonstrated superior performance, achieving R2 = 0.85, MAE = 1.39 years, and RMSE=1.90 years. For MIC_best_set, the top-performing models exhibited reduced accuracy, yielding R2 < 0.75, MAE > 1.90 years, and RMSE > 2.60 years. For SFS_best_set (261 AR-CpGs, 43.8% fewer than L1_best_set), Transformed Target Regressor achieved comparable precision (R^2 = 0.84, MAE = 1.46 years, RMSE = 1.99 years). The SFS-derived Transformed Target Regressor was selected as the final sperm epigenetic clock based on its optimal balance of marker efficiency and predictive accuracy (Table S2). Table S3 and Fig. S7 present hyperparameter configurations for all 23 SFS-based models.

We compared these 261 AR-CpG sites with previously reported semen/sperm-specific AR-CpG markers and found that the vast majority did not overlap (n = 239), indicating that the markers selected using the multiple feature selection strategy are highly novel (Fig. S8).

Table 1 A performance comparison of the top three regression models under maximum mutual information, sequential feature selection, and L1 regularization feature selection.

Method	Number of CpGs	Models	R^2	MAE(Years)	RMSE(Years)*
Maximum mutual information	470	xgbtree	0.72	1.9462	2.6709
		Transformed Target Regressor	0.71	1.9661	2.6885
		svmpoly	0.72	1.9405	2.6082
Lasso	464	Poisson Regressor	0.85	1.3737	1.8996
		Transformed Target Regressor	0.85	1.3855	1.9182
		TweedieRegressor	0.84	1.4084	1.9636
Sequential feature selection	261	Transformed Target Regressor	0.84	1.4583	1.9888
		Poisson Regressor	0.84	1.4677	2.0024
		Hobber Regression	0.83	1.5213	2.0488

*MAE, mean absolute error; RMSE, root mean square error

3.1.5 Evaluation of sperm chronological epigenetic clock

The sperm epigenetic clock was externally validated using two independent methylation microarray datasets— GSE185445 (379 healthy male sperm samples) and GSE149318 (90 sperm samples)—to assess its robustness in external datasets. Performance was benchmarked against nine published epigenetic clocks including tissue-specific models—Hannum's blood clock (Hannum, et al., 2013), Horvath's pan-tissue clock (Horvath2013)(Horvath, 2013), Horvath's skin–blood clock (Horvath2018)(Horvath et al., 2018), McEwen's oral clock (McEwen2019)(Mcewen et al., 2020), Jenkins' sperm clock (Jenkins2017) (Jenkins et al., 2018), and Zhang's blood clock (ZhangQ2019)(Zhang et al., 2019)—and optimization-based models—Higgins–Chen's principal component-enhanced clocks (PCHannumG2013, PCHorvathS2013, PCHorvathS2018)(Higgins-Chen et al., 2022). All 10 clocks, including our newly developed sperm clock, were evaluated using standardized performance metrics (R^2, MAE, RMSE) across both cohorts. Complete model citations are provided in Table S4.

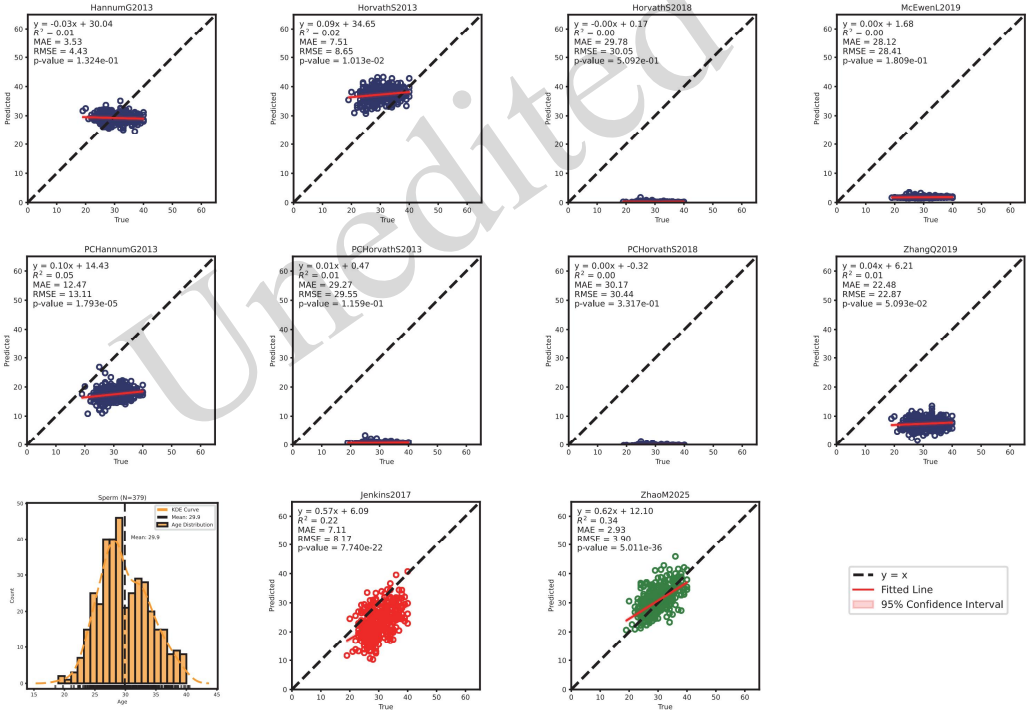

Fig. 4 Age prediction performances of 379 sperm samples (GSE185445) using 10 chronological epigenetic clocks. HannumG2013, HorvathS2013, HorvathS2018, Jenkins2017, McEwenL2019, ZhangQ2019, PCHannumG2013, PCHorvathS2013, PCHorvathS2018, and ZhaoM2025 represent the Hannum clock (blood), Horvath's pan-tissue clock, Horvath's skin-oral clock, Jenkins'sperm clock, McEwen's oral clock, Zhang's whole blood clock, the principal component analysis-based optimized clocks PCHannumG2013, PCHorvathS2013, and PCHorvathS2018, and our sperm epigenetic clock, respectively.

The performance of all 10 epigenetic clocks was evaluated using a screening cohort of 1471 sperm samples (GSE185920; Fig. S9). The newly developed sperm epigenetic clock demonstrated superior predictive accuracy

(R^2 = 0.85, MAE = 1.63 years, RMSE = 2.20 years), outperforming all nine published comparison clocks. The Jenkins2017 clock achieved the second-best performance (R^2 = 0.54, MAE = 2.97 years, RMSE = 4.04 years). The remaining eight clocks, developed for non-sperm tissues, exhibited low accuracy (R^2 < 0.1), with MAE ranging from 4.72 to 32.49 years and RMSE from 6.22 to 32.99 years. Among these clocks, HannumG2013 (MAE = 4.72 years, RMSE = 6.22 years) and HorvathS2013 (MAE = 5.28 years, RMSE = 6.55 years) demonstrated limited generalizability to sperm samples. The optimization-enhanced clocks exhibited the poorest accuracy (MAE > 29 years, RMSE > 30 years).

In the independent validation cohort (GSE185445, n = 379 sperm samples), all 10 epigenetic clocks exhibited reduced predictive performance compared to the larger benchmark cohort (GSE185920, n = 1471 sperm samples). Nevertheless, the newly developed sperm clock maintained superior accuracy (R2 = 0.34, MAE = 2.93 years, RMSE = 3.90 years; Fig. 4), with the Jenkins2017 clock ranking second (R^2 = 0.22, MAE = 7.11 years, RMSE = 8.17 years). The four tissue-specific clocks developed for non-sperm tissues (HannumG2013, HorvathS2018, McEwenL2019, ZhangQ2019) demonstrated no significant correlations between predicted and chronological ages (p > 0.05). Among the three optimization-enhanced models, only PCHannumG2013 showed statistical significance (p < 0.05), while PCHorvathS2013 and PCHorvathS2018 did not (p > 0.05). Despite reduced performance for the external validation cohort, the newly developed clock maintained substantially higher accuracy than all nine published models, demonstrating superior generalizability to independent datasets.

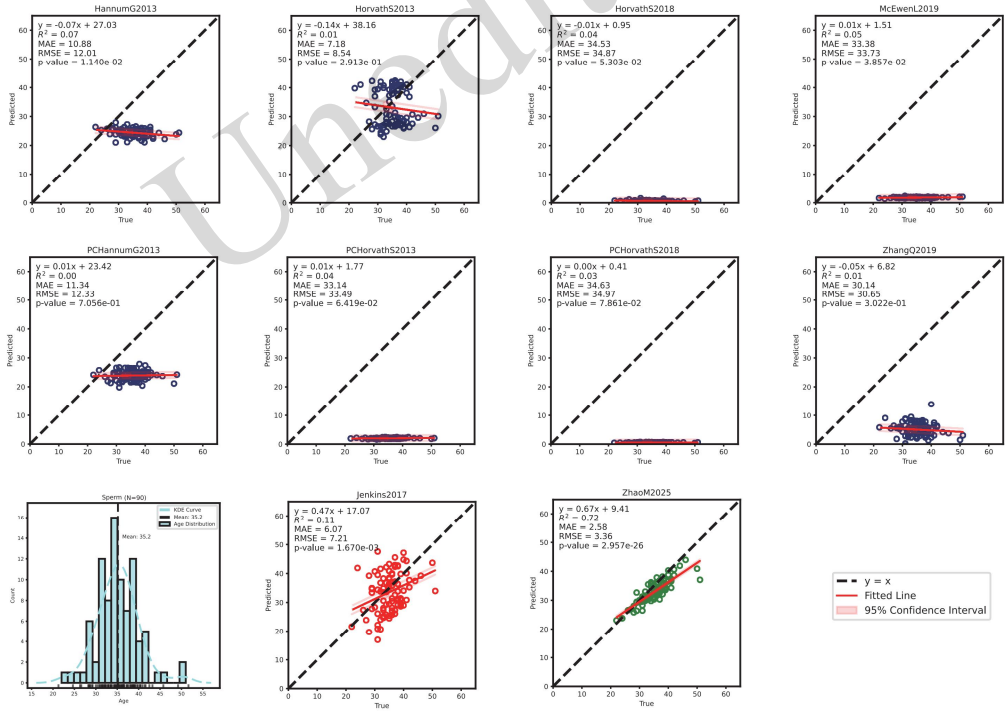

Fig. 5 Age prediction performances of 90 sperm samples (GSE149318) using 10 chronological epigenetic clocks. HannumG2013, HorvathS2013, HorvathS2018, Jenkins2017, McEwenL2019, ZhangQ2019, PCHannumG2013, PCHorvathS2013, PCHorvathS2018, and ZhaoM2025 represent the Hannum clock (blood), Horvath's pan-tissue clock, Horvath's skin-oral clock, Jenkins' sperm clock, McEwen's oral clock, Zhang's whole blood clock, the principal component analysis-based optimized clocks PCHannumG2013, PCHorvathS2013, and PCHorvathS2018, and our sperm epigenetic clock, respectively.

Evaluation of the 90 sperm samples (GSE149318) revealed distinct performance patterns across the 10 epigenetic clocks (Fig. 5). Our sperm epigenetic clock demonstrated superior predictive accuracy ($R^2 = 0.72$, MAE = 2.58 years, RMSE = 3.36 years), significantly outperforming comparative models. The Jenkins2017 clock exhibited reduced performance ($R^2 = 0.11$, MAE = 6.07 years, RMSE = 7.21 years), and three tissue-specific clocks (HorvathS2013, HorvathS2018, ZhangQ2019) showed no significant age correlations ($p > 0.05$). Similarly, optimization-enhanced clocks (PCHannumG2013, PCHorvathS2013, PCHorvathS2018) yielded non-significant correlations ($p > 0.05$). Notably, the HannumG2013 clock displayed a significant negative correlation between predicted and chronological ages ($p < 0.05$).

To assess our clock's performance relative to existing sperm-specific models, we provide comparative metrics with SEACpG (Pilsner et al., 2022), the most recently published sperm clock. Since the original publication did not disclose model parameters and feature specifications for SEACpG, direct replication or head-to-head testing on identical datasets is not feasible; therefore, we compared reported performance metrics across respective training and validation cohorts. SEACpG reported a training performance of $r = 0.91$ ($R^2 \approx 0.83$) with MAE = 1.6 years using a subset of GSE185445 (n = 379), whereas our clock achieved $R^2 = 0.85$ with MAE = 1.63 years using GSE185920 (n = 1471). That the clocks exhibit similar performance metrics despite a nearly four-fold difference in training set size suggests that both feature selection strategies effectively captured age-related methylation patterns, although dataset-specific characteristics prevent definitive model comparison. For external validation, Pilsner et al. (2022) reported $r = 0.79$ ($R^2 \approx 0.62$) without specifying the composition or size of the validation dataset, whereas our clock achieved $R^2 = 0.72$, MAE = 2.58 years, and RMSE = 3.36 years in an independent sperm cohort (GSE149318, n=90). Both models demonstrate generalization capability with external validation $R^2 > 0.62$, though the different test cohorts limit direct comparability.

3.2 Construction of the semen age estimation model

3.2.1 Analysis and AR-CpG screening of WGBS dataset

We analyzed the WGBS dataset GSE222340 to identify novel AR-CpGs. This longitudinal study comprised paired sperm samples from 10 donors collected at two time points (De Sena Brandine et al., 2023): baseline samples (Timepoint A) and follow-up samples obtained 10–18 years later (Timepoint B). Raw methylation data files (.meth files) containing CpG methylation levels and read coverage depths were processed using the Dnmtools package (Ubuntu version 20.0.4) to generate site-by-sample methylation matrices. Age-correlation analysis identified 270,019 AR-CpGs in Timepoint A samples and 243,027 AR-CpGs in Timepoint B samples (both at adjusted $p < 0.01$). Intersection analysis revealed that 243,027 AR-CpGs were shared between time points (Fig. S10). The top 100 AR-CpGs ranked by absolute Spearman's correlation coefficient are listed in Table S5, with all age correlation coefficients reaching at least 0.95.

3.2.2 Pyrosequencing results

To develop a robust detection system suitable for forensic casework, we selected eight candidate AR-CpGs from two discovery sources. Four AR-CpGs were selected from the microarray-based sperm clock (Section 3.1.4), prioritizing top-ranked markers with strong age correlation ($|\rho| > 0.5$). An additional four CpG sites were selected from the WGBS analysis (Section 3.2.1), prioritizing top-ranked markers with very strong age associations ($|\rho| > 0.9$) in the GSE222340 longitudinal cohort. The design characteristics of site-specific primers were also taken into account. Pyrosequencing validation was performed on 95 independent semen samples (age range: 20–42 years, median: 29 years) for all eight selected AR-CpGs (Table 2, Fig. S11). Linear regression analysis revealed that seven sites demonstrated significant age correlations ($p < 0.05$), while cg15932627 exhibited a positive trend that did not reach statistical significance ($p = 0.11$) (Fig. 6). Cross-platform comparison revealed that pyrosequencing-derived age correlations were generally weaker than those observed in the discovery datasets. Additionally, absolute methylation levels differed systematically between platforms. Three CpG sites, cg18857873 (microarray-derived), chrY:785986 (WGBS-derived), and chr14:30407630 (WGBS-derived), exhibited higher methylation beta values in pyrosequencing compared to their respective discovery platforms ($p < 0.05$), likely reflecting technical differences between platforms. Notably, chr10:71661611 (CpG9) displayed opposing correlation trends ($\rho = 0.18$ in pyrosequencing vs. $\rho = -0.97$ in WGBS). Despite this discordance, the marker was retained based on its pyrosequencing performance. Adjacent age-related sites were identified in CpG4 (chr14:30407626, CpG5) and CpG7 (cg18857873+1, CpG8)

amplification regions. These flanking sites provided supplementary age information within the existing amplicons, which may enhance model robustness.

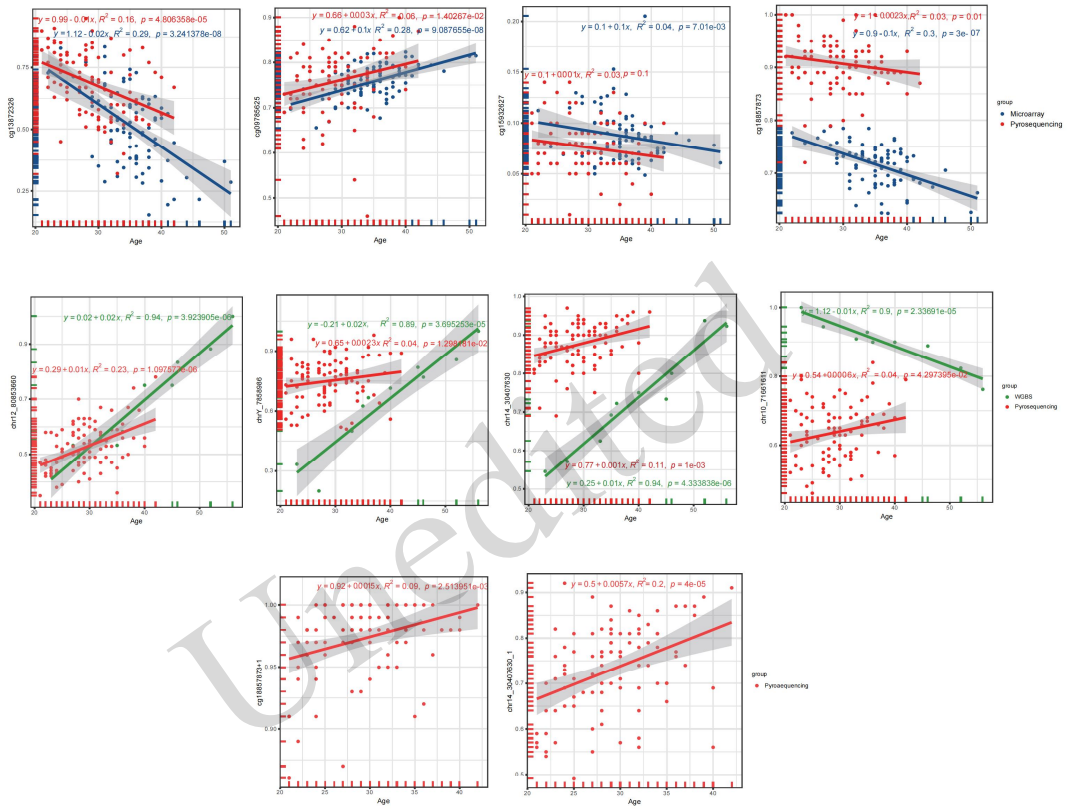

Fig. 6 Linear regression plots of sperm AR-CpG site expression levels versus individual age on WGBS, microarray, and pyrosequencing platforms. WGBS data were based on the GSE222340 dataset containing 10 sperm samples, microarray data were based on the GSE149318 dataset containing 90 sperm samples, and pyrosequencing was the validation result of 95 semen samples.

Table 2 Information of amplification primers and pyrosequencing sequencing primers for sperm age-related CpGs.

No. CpG	ID	Location	Primer Sequence (5'→3')	Primer length (bp)	Type	Amplicon length (bp)
CpG1	-	chr12:80853660[a]	tgaggtaggaggattatgaggttag	25	PCR-Forward	134
			Biotin-ctccctcaacctctaaaactac	22	PCR-Reverse	
			attaaaaaatagaaaaaattagt	23	Sequencing	
CpG2	cg13872326	chr7:27901067[b]	Biotin-gggaagttgatttttatgttggttttattg	30	PCR-Forward	158
			ccccacccctttaattttaacttt	23	PCR-Reverse	
			ccccccccctacctacctattaaatatataaac	32	Sequencing	
CpG3	cg09785625	chr16:80840984[b]	atgattaataggatgtgtgtgtattg	26	PCR-Forward	91
			Biotin-cactcaatttaaaaccttaatcatctcta	29	PCR-Reverse	
			tgaggatttggggag	15	Sequencing	
CpG4*	-	chr14:30407630[a]	tgtttattgggtattaaatagtaagtgtt	29	PCR-Forward	210
			Biotin-tttctctctaaatatacctttaaaaatcc	29	PCR-Reverse	
			aaattattatgagaggttgaataaa	25	Sequencing	
CpG5*	-	chr14:30407634[a]	tgtttattgggtattaaatagtaagtgtt	29	PCR-Forward	210
			Biotin-tttctctctaaatatacctttaaaaatcc	29	PCR-Reverse	
			aaattattatgagaggttgaataaa	25	Sequencing	
CpG6	-	chrY:7858986[a]	ggagtggtttaggtgtgattaatt	24	PCR-Forward	188
			Biotin-tttaaaccccaaacctatactcaaat	26	PCR-Reverse	
			ttgagttttaattgtttgagttaat	25	Sequencing	
CpG7*	cg18857873	chr17:71164157[b]	ttttagtaaggaagtaatataggagttagt	30	PCR-Forward	101
			Biotin-tcaaaacatcccctctactcctca	24	PCR-Reverse	
			ttgttttaaattaaagtttagttat	25	Sequencing	
CpG8*	-	chr17:71164161[b]	ttttagtaaggaagtaatataggagttagt	30	PCR-Forward	101
			Biotin-tcaaaacatcccctctactcctca	24	PCR-Reverse	
			ttgttttaaattaaagtttagttat	25	Sequencing	
CpG9	-	chr10:71661611[a]	gggagtttaagggttatttggattat	26	PCR-Forward	136
			Biotin-attctatacatctccaaaataccatacc	28	PCR-Reverse	
			agggttatttggattatattag	22	Sequencing	
CpG10	cg15932627	chr3:139048065[b]	attagatagggttttgtagtaaagttattt	30	PCR-Forward	175
			Biotin-ttctccttttcatacaacctcatact	26	PCR-Reverse	
			agtaaagttatttttttttttagaa	25	Sequencing	

*CpG 4 and CpG 5 are co-amplified within a single amplicon, while CpG 7 and CpG 8 are co-amplified within another amplicon; [a] Telomere-to-Telomere assembly of the CHM13 cell line, with chr Y from NA24385 (T2T CHM13v2.0); [b] Genome reference consortium human genome build 37

3.2.3 Evaluation of semen age prediction models

We developed a semen age prediction model based on pyrosequencing validation results, using pyrosequencing data from nine AR-CpG sites: seven sites with significant age correlations ($p < 0.05$, excluding non-significant cg15932627) and two additional AR-CpGs identified within amplicon regions (CpG 5 and CpG 8, Section 3.2.2). Using the modeling approach described in Section 3.1, 23 regression algorithms were compared via Taylor diagram analysis (Fig. 7) to identify the optimal predictor. The radial basis function support vector machine (SVM-RBF) exhibited optimal performance, achieving a test-set correlation coefficient (r) of 0.79 with MAE and RMSE values of 2.21 and 3.15 years, respectively (Fig. 7). Detailed hyperparameter configurations and comparative performance metrics for all 23 models are provided in Table S6, and Fig. S12 visualizes the model performance distributions. The final SVM-RBF model was packaged as the Age-Semen-Predictor software for operational forensic use and is openly accessible at https://github.com/zuigaoming/DNAmethylation, with source code and documentation.

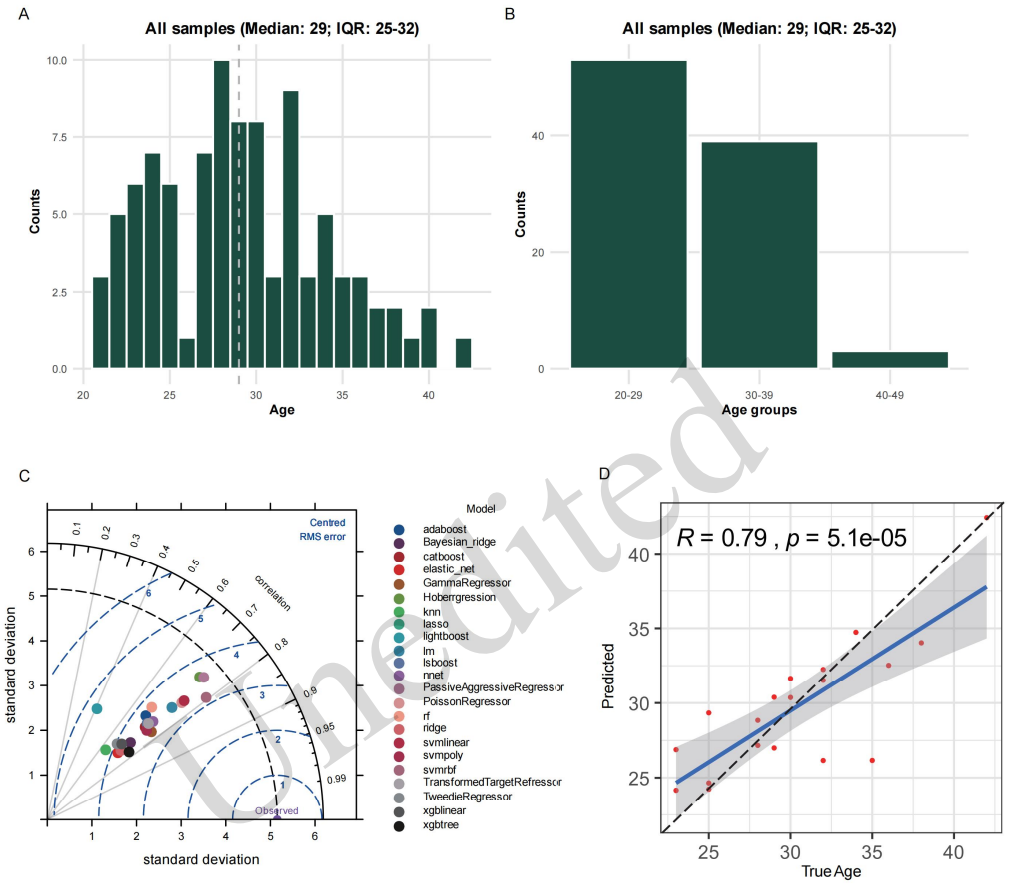

Fig. 7 Age distribution of semen samples for pyrosequencing and the construction and evaluation of semen age estimation models. (A, B) Age distribution histograms of semen samples; (C) Taylor diagram of regression model performance based on nine sperm AR-CpG sites; (D) Scatter plot of the true age and predicted age in the test set under the support vector machine regression model with radial basis function kernel.

4. Discussion

Epigenetic clocks have been categorized into two principal types based on their prediction targets: chronological and biological epigenetic clocks (Se et al., 2017; Onofri et al., 2023). Chronological epigenetic clocks focus on accurately predicting chronological age, whereas biological epigenetic clocks integrate epigenetic drift markers to assess biological aging processes (Levine et al., 2018; Lu et al., 2019; Mcgreevy et al., 2023). Forensic applications primarily employ chronological clocks to estimate age from biological evidence in investigative contexts. Substantial progress has been made in developing such clocks, including pan-tissue models applicable across multiple sample types (Hannum, et al., 2013; Horvath, 2013), and tissue-specific models optimized for particular biological materials, including blood, saliva, and buccal cells (Zhang, et al., 2019; Mcewen, et al., 2020). The pan-tissue clocks appear to be unsuitable for semen samples, given their poor predictive accuracy. Additionally, chronological clocks for semen samples remain relatively undeveloped compared to somatic tissue clocks, primarily constrained by the scarcity of publicly available sperm methylation microarray datasets. Previous semen age estimation studies employing small CpG panels

have achieved prediction errors of 4–5 years, suggesting that markers identified from microarray coverage may not fully capture age-related methylation dynamics in semen evidence.

This study attempted to address these issues using a feasible strategy. First, we developed a sperm epigenetic clock utilizing the largest available sperm methylation microarray dataset (GSE185920, n=1471). This larger cohort enabled more robust marker selection and reduced the risk of overfitting compared to models trained on fewer than 100 samples. AR-CpGs were identified through bioinformatics analysis, and these markers revealed methylation signatures in sperm samples distinct from those in somatic body fluid types. Functional enrichment analyses (KEGG/GO) further implicated these sites in biological processes critical to spermatogenesis, sperm maturation, and motility regulation, providing mechanistic insights into their epigenetic regulatory roles.

The optimal AR-CpG sets for the sperm epigenetic clock were selected using multiple AI-driven feature selection algorithms (MIC, Lasso, and SFS). The MIC algorithm can detect nonlinear relationships between age and methylation levels using nonparametric estimation, effectively capturing complex biological associations (Kinney and Atwal, 2014a). L1 regularization selects minimally redundant features by imposing sparsity constraints, reducing dimensionality while retaining critical CpG sites from correlated clusters (Schmidt, et al., 2007). The SFS algorithm iteratively constructs optimal CpG combinations by evaluating incremental predictive improvements, ensuring synergistic marker interactions (Aha and Bankert, 1995). Diverse feature selection strategies are integrated through this multi-method approach to reduce bias risks associated with single-algorithm methodologies. Model performance was comparatively evaluated across different optimal CpG sets, enabling a comprehensive evaluation of AR-CpG markers through multidimensional feature analysis.

We evaluated 23 machine learning regression models across the three optimal feature sets to identify the most accurate chronological epigenetic clock. The model using sequential feature selection with 261 AR-CpGs achieved the best training performance (R^2= 0.85, MAE = 1.63 years, RMSE = 2.20 years). The sperm-specific epigenetic clock was evaluated using two independent sperm methylation microarray datasets, with its performance benchmarked against that of nine published epigenetic clocks on the same validation sets. In 379 sperm samples (GSE185445), the developed clock demonstrated modest performance (R^2 = 0.34, MAE = 2.93 years) and superior accuracy (R^2 = 0.72, MAE = 2.58 years) in 90 sperm samples (GSE149318). The performance reduction compared to training metrics reflects multiple factors, including smaller validation sample sizes, narrower age distributions (20–42 years in validation versus 21–56 years in training), and potential inter-cohort batch effects. Such attenuation is commonly observed in epigenetic biomarker validation studies and does not indicate poor marker transferability. We noticed that the datasets used for training (GSE185920) and validation (GSE185445 and GSE149318) may have been generated using different sperm DNA extraction protocols, including gradient centrifugation and somatic cell lysis. This technical heterogeneity could introduce systematic biases, potentially compromising the model's generalizability in independent cohorts and contributing to the observed discrepancies in external validation results.

Our clock substantially outperformed the Jenkins sperm clock (which achieved R^2 ranging from 0.11 to 0.54 and MAE from 2.97 to 7.11 years) across our validation cohorts, likely due to Jenkins' smaller training samples and the use of the earlier feature selection approaches available at that time. More recently, Pilsner et al. developed SEACpG, the performance of which is comparable to that of our clock (reported R^2 approximately 0.83, MAE 1.6 years using n=379 from GSE185445)(Pilsner, et al., 2022). Direct performance comparison is complicated by methodological differences: SEACpG was trained on GSE185445 (which we used for external validation), and the characteristics of its validation cohort remain undisclosed. Nevertheless, our clock achieved numerically higher external validation R^2 (0.72 versus their reported 0.62), although different test datasets preclude definitive superiority claims. Our study's primary contributions extend beyond gains in incremental accuracy to include comprehensive multi-cohort validation (n=1940), demonstrated robustness in sperm samples, and complete methodological transparency with open-source implementation. In addition, pan-tissue clocks and tissue-specific and optimization-enhanced clocks exhibited significantly higher prediction errors in sperm samples, underscoring the need to develop sperm-specific epigenetic clocks.

While microarray platforms enable economical large-cohort screening, their coverage of approximately 850,000 CpG sites represents only 3–4 percent of the roughly 28 million CpG sites in the human genome (Vidaki and Kayser, 2018). To identify novel age-related markers beyond microarray coverage, we performed a comprehensive analysis of WGBS data from a longitudinal aging cohort (GSE222340, n=10 individuals with paired samples spanning 10–18 years). Based on sperm WGBS data, we identified novel sperm-specific

AR-CpGs, among which more than 240,000 were significant (adjusted $p < 0.01$). These markers were highly correlated with donor age ($|\rho| > 0.9$), far outperforming the microarray-derived markers (cg18037145, $\rho = 0.58$, $p < 0.01$). Highly age-correlated CpGs mean that better age estimation predictive accuracy can be achieved with fewer markers and simpler models.

However, several limitations affect the reliability of these WGBS findings. The small WGBS cohort (n=10 paired samples) limits the reliability of individual marker estimates. This concern was confirmed when we validated these markers using pyrosequencing: some showed opposite correlation directions to those observed in the WGBS discovery phase. Additionally, the exceptionally high correlations (exceeding 0.9) we observed in this small longitudinal dataset likely overestimate what can be achieved in larger, more diverse populations. While WGBS clearly identifies promising candidate markers, each must be validated independently in larger cohorts using the intended implementation platform before forensic application.

Prior forensic studies have developed targeted approaches to predicting age from semen, making substantive progress in recent years. Lee et al. (2015) employed three array-derived CpG sites, achieving an MAE of 4.8 years (Lee et al., 2015), with follow-up validation across diverse specimen conditions reporting an MAE of 3.9–5.2 years (Lee et al., 2018; Li et al., 2020). The VISAGE Consortium's massively parallel sequencing assay, which used six selected markers, achieved an MAE of 5.1 years (Pisarek et al., 2021). Xiao et al. identified 21 AR-CpGs from microarray data and developed two detection panels. They integrated markers from panel I (n=11), panel II (n=10), or both to construct the optimal model, achieving MAEs of 2.526–4.746 years, 3.890–5.715 years, and greater than 9.800 years on the test sets of sperm, semen, and whole blood, respectively (Xiao et al., 2023). The current research revealed that age estimation from semen samples exhibited upgradable predictive accuracy, and novel AR-CpGs need to be enrolled in model construction.

To develop a robust detection system for forensic practice, this study validated the forensic applicability of the sperm-specific AR-CpGs identified from sperm methylation microarray and WGBS data using pyrosequencing in semen samples. Four AR-CpGs were selected from the microarray-based sperm clock ($\rho > 0.5$) and four from WGBS analysis ($\rho > 0.9$) for the validation. When we tested these markers in 95 semen samples (ages 20–42 years), the age correlations were weaker than in discovery. Seven markers showed significant correlations ($p < 0.05$), whereas one (cg15932627) did not ($p = 0.11$). Additionally, one marker (chr10:71661611) showed a strong negative correlation in WGBS ($\rho = -0.97$) but a weak positive correlation in pyrosequencing ($\rho = 0.2$).

It is worth noting that these AR-CpGs did not exhibit the anticipated age-related characteristics ($|\rho|$ ranging from 0.20 to 0.56, Fig. 6). There may be several explanations for this weaker performance. First, technical differences between methylation detection platforms could affect data comparability (Kacmarczyk et al., 2018). Compared with methylation microarrays, the measurement error caused by insufficient sequencing depth (< 30×) in WGBS overestimated the age-related performance of its CpG sites to some extent. In addition, the sample size of the WGBS dataset in this study was relatively small (n = 10), which also increased the statistical risk. Second, differences in age distributions among the pyrosequencing validation cohort (20–42 years), sperm microarray dataset (21–52 years), and WGBS dataset (23–56 years) might increase biological heterogeneity in methylation profiles, underscoring the need to expand age-range validations. Additionally, Y-chromosomal amplification bias caused by repetitive sequences may explain the inverted age correlation observed at CpG6 (Reed et al., 2010; Hong et al., 2019; Freire-Aradas et al., 2020; Schwender et al., 2021). Furthermore, semen cellular heterogeneity (5–15% somatic cell contamination) may dilute sperm-specific methylation signals, compromising age correlations (Siebert-Kuss et al., 2024). Research also demonstrates that the methylation profile of semen is predominantly shaped by sperm-derived DNA and influenced by somatic cell-free DNA or non-cellular elements in seminal plasma (Schütte et al., 2013; Barney et al., 2022). Therefore, it is necessary to further optimize the semen DNA extraction method to reduce the interference of biological background noise.

Although the performance of these AR-CpGs in the pyrosequencing platform did not meet expectations, we still explored the overall performance of the age prediction model constructed using these markers in semen. Notably, the support vector machine with a radial basis function for age estimation, evaluated using Taylor diagram analysis, exhibited ideal performance. The finalized model, named "Age-Semen-Predictor", and associated resources (source code, test datasets) have been made publicly available on GitHub to facilitate cross-laboratory replication and refinement.

Although WGBS datasets provide a broader genomic coverage context for screening methylation markers, given their technical limitations (such as measurement errors due to insufficient sequencing depth and

systematic differences across detection platforms), results derived from them should be interpreted with caution. Still, several limitations need to be considered. Validation using pyrosequencing platforms was constrained by a relatively narrow age range and a limited sample size of seminal specimens. Crucially, the potential linkage between these newly identified CpGs and DNA polymorphisms remains unexplored in this preliminary investigation. Furthermore, given that DNA methylation patterns are influenced by multiple confounding factors, including ethnicity, gender, and environmental exposures, the evidentiary value of these epigenetic markers requires more comprehensive validation using systematic population studies and inter-laboratory reproducibility assessments.

5. Conclusions

This study established and validated a sperm-specific epigenetic clock through systematic analysis of microarray data from a large cohort, employing three distinct feature selection methods. WGBS analysis of sperm samples revealed novel AR-CpGs, which hold exploratory research value for age estimation. The finalized semen age estimation model, demonstrating robust predictive capacity, has been implemented as an open-source tool accessible through a dedicated GitHub repository. This work provides new strategies and insights for forensic epigenetics, particularly for sexual assault investigations requiring suspect age estimation. Future investigations will expand the sample size and age range for validation to enhance forensic applicability.

Acknowledgments

This work is supported by the National Natural Science Foundation of China (No. 82293652). We thank the volunteers for their contributions to this study.

Data availability statement

The data and materials used to support the findings of this study are available from the corresponding authors upon request.

Author contributions

Ming ZHAO: Writing-original draft, Visualization, Validation, Software, Methodology, Formal analysis, Data curation. Fanzhang LEI: Visualization, Software, Methodology, Formal analysis, Data curation. Meiming CAI: Methodology, Formal analysis, Data curation. Qinglin LIANG：Methodology, Formal analysis, Data curation, Writing-original draft. Xi YUAN: Formal analysis, Data curation. Qiong LAN: Writing-review & editing, Resources, Project administration, Conceptualization. Yating FANG: Writing-review & editing, Resources, Project administration, Conceptualization. Bofeng ZHU: Writing-review & editing, Resources, Project administration, Conceptualization.

Compliance with ethics guidelines

Ming ZHAO, Fanzhang LEI, Meiming CAI, Qinglin LIANG, Xi YUAN, Qiong LAN, Yating Fang, and Bofeng ZHU declare that they have no conflict of interest.

All procedures followed were in accordance with the ethical standards of the Ethics Committee of Southern Medical University (Approval NO. 2023-KY-097-02) and with the Helsinki Declaration of 1975, as revised in 2013. Informed consent was obtained from all patients for being included in the study. Additional informed consent was obtained from all patients for whom identifying information is included in this article.

Declaration on the use of generative AI tools

No generative AI tools were used in the preparation of this manuscript.

References

Alsaleh H, Haddrill PR, 2019. Identifying blood-specific age-related DNA methylation markers on the illumina methylationepic® beadchip. *Forensic Science International*, 303:109944. https://doi.org/10.1016/j.forsciint.2019.109944

Arany S, Ohtani S, 2011. Age estimation of bloodstains: A preliminary report based on aspartic acid racemization rate. *Forensic Sci Int*, 212(1-3):e36-39. https://doi.org/10.1016/j.forsciint.2011.05.015

Aryee MJ, Jaffe AE, Corrada-Bravo H, et al., 2014. Minfi: A flexible and comprehensive bioconductor package for

the analysis of infinium DNA methylation microarrays. *Bioinformatics*, 30(10):1363-1369. https://doi.org/10.1093/bioinformatics/btu049

Åsenius F, Gorrie-Stone TJ, Brew A, et al., 2020. The DNA methylome of human sperm is distinct from blood with little evidence for tissue-consistent obesity associations. *PLOS Genetics*, 16(10):e1009035. https://doi.org/10.1371/journal.pgen.1009035

Barney R, Stalker K, Lutes A, et al., 2022. Assessment of seminal cell-free DNA as a potential contaminate in studies of human sperm DNA methylation. *Andrology*, 10(4):702-709. https://doi.org/10.1111/andr.13163

Beck D, BM, Millissia, And Skinner MK, 2022. Genome-wide cpg density and DNA methylation analysis method (medip, rrbs, and wgbs) comparisons. *Epigenetics*, 17(5):518-530. https://doi.org/10.1080/15592294.2021.1924970

Bell JT, Tsai P-C, Yang T-P, et al., 2012. Epigenome-wide scans identify differentially methylated regions for age and age-related phenotypes in a healthy ageing population. *PLoS genetics*, 8(4):e1002629. https://doi.org/10.1371/journal.pgen.1002629

Breitling LP, Saum K-U, Perna L, et al., 2016. Frailty is associated with the epigenetic clock but not with telomere length in a german cohort. *Clinical Epigenetics*, 8:21. https://doi.org/10.1186/s13148-016-0186-5

Christensen BC, Houseman EA, Marsit CJ, et al., 2009. Aging and environmental exposures alter tissue-specific DNA methylation dependent upon cpg island context. *PLoS genetics*, 5(8):e1000602. https://doi.org/10.1371/journal.pgen.1000602

Cui X, Jing X, Wu X, et al., 2016. DNA methylation in spermatogenesis and male infertility (review). *Experimental and Therapeutic Medicine*, 12(4):1973-1979. https://doi.org/10.3892/etm.2016.3569

De Sena Brandine G, Aston KI, Jenkins TG, et al., 2023. Global effects of identity and aging on the human sperm methylome. *Clinical Epigenetics*, 15(1):127. https://doi.org/10.1186/s13148-023-01541-6

Eckhardt F, Lewin J, Cortese R, et al., 2006. DNA methylation profiling of human chromosomes 6, 20 and 22. *Nat Genet*, 38(12):1378-1385. https://doi.org/10.1038/ng1909

Elfawal MA, Alqattan SI, Ghallab NA, 2015. Racemization of aspartic acid in root dentin as a tool for age estimation in a kuwaiti population. *Medicine, Science, and the Law*, 55(1):22-29. https://doi.org/10.1177/0025802414524383

Fang Y, Chen M, Cai M, et al., 2023. Selection and validation of a novel set of specific differential methylation markers and construction of a random forest prediction model for the accurate tissue origin identifications of body fluids involving young and middle-aged group of chinese han population. 137(5):1395-1405.

Feng L, Lou J, 2019. DNA methylation analysis. *Methods in Molecular Biology (Clifton, NJ)*, 1894:181-227. https://doi.org/10.1007/978-1-4939-8916-4_12

Fleischer JG, Schulte R, Tsai HH, et al., 2018. Predicting age from the transcriptome of human dermal fibroblasts. *Genome Biology*, 19(1):221. https://doi.org/10.1186/s13059-018-1599-6

Freire-Aradas A, Pośpiech E, Aliferi A, et al., 2020. A comparison of forensic age prediction models using data from four DNA methylation technologies. *Front Genet*, 11:932. https://doi.org/10.3389/fgene.2020.00932

Friedrich U, Schwab M, Griese EU, et al., 2001. Telomeres in neonates: New insights in fetal hematopoiesis. *Pediatric Research*, 49(2):252-256. https://doi.org/10.1203/00006450-200102000-00020

Griffin RC, Moody H, Penkman KEH, et al., 2008. The application of amino acid racemization in the acid soluble fraction of enamel to the estimation of the age of human teeth. *Forensic Science International*, 175(1):11-16. https://doi.org/10.1016/j.forsciint.2007.04.226

Griffin RC, Chamberlain AT, Hotz G, et al., 2009. Age estimation of archaeological remains using amino acid racemization in dental enamel: A comparison of morphological, biochemical, and known ages-at-death. *American Journal of Physical Anthropology*, 140(2):244-252. https://doi.org/10.1002/ajpa.21058

Hannum G, Guinney J, Zhao L, et al., 2013. Genome-wide methylation profiles reveal quantitative views of human aging rates. *Molecular Cell*, 49(2):359-367. https://doi.org/10.1016/j.molcel.2012.10.016

Hassan Q, Rakha A, Bashir MZJJCPSP, 2017. Aspartic acid racemization with correlation to age: Aforensic perspective. 27(5):283-287.

Higgins-Chen AT, Thrush KL, Wang Y, et al., 2022. A computational solution for bolstering reliability of epigenetic clocks: Implications for clinical trials and longitudinal tracking. *Nature Aging*, 2(7):644-661. https://doi.org/10.1038/s43587-022-00248-2

Hong SR, Shin K-J, Jung S-E, et al., 2019. Platform-independent models for age prediction using DNA methylation data. *Forensic Science International Genetics*, 38:39-47. https://doi.org/10.1016/j.fsigen.2018.10.005

Horvath S, 2013. DNA methylation age of human tissues and cell types. *Genome biology*, 14:1-20.

Horvath S, Oshima J, Martin GM, et al., 2018. Epigenetic clock for skin and blood cells applied to hutchinson gilford progeria syndrome and ex vivo studies. *Aging*, 10(7):1758-1775. https://doi.org/10.18632/aging.101508

Huan T, Chen G, Liu C, et al., 2018. Age-associated microrna expression in human peripheral blood is associated with all-cause mortality and age-related traits. *Aging Cell*, 17(1):e12687. https://doi.org/10.1111/acel.12687

Jenkins T, Aston K, Carrell D, et al., 2022. The impact of zinc and folic acid supplementation on sperm DNA methylation: Results from the folic acid and zinc supplementation randomized clinical trial (fazst). *Fertility and Sterility*, 117(1):75-85. https://doi.org/10.1016/j.fertnstert.2021.09.009

Jenkins TG, Aston KI, Cairns B, et al., 2018. Paternal germ line aging: DNA methylation age prediction from human sperm. *BMC Genomics*, 19(1):1-10. https://doi.org/10.1186/s12864-018-5153-4

Jung S-E, Lim SM, Hong SR, et al., 2019. DNA methylation of the *elovl2, fhl2, klf14, c1orf132/mir29b2c,* and *trim59* genes for age prediction from blood, saliva, and buccal swab samples. *Forensic Science International: Genetics*, 38:1-8. https://doi.org/10.1016/j.fsigen.2018.09.010

Jylhävä J, Pedersen NL, Hägg S, 2017. Biological age predictors. *EBioMedicine*, 21:29-36. https://doi.org/10.1016/j.ebiom.2017.03.046

Kacmarczyk TJ, Fall MP, Zhang X, et al., 2018. "Same difference": Comprehensive evaluation of four DNA methylation measurement platforms. *Epigenetics & Chromatin*, 11(1):21. https://doi.org/10.1186/s13072-018-0190-4

Kinney JB, Atwal GS, 2014a. Equitability, mutual information, and the maximal information coefficient. *Proceedings of the National Academy of Sciences*, 111(9):3354-3359. https://doi.org/10.1073/pnas.1309933111

Kinney JB, Atwal GSJPOTNaOS, 2014b. Equitability, mutual information, and the maximal information coefficient. 111(9):3354-3359.

Lee HY, Jung S-E, Oh YN, et al., 2015. Epigenetic age signatures in the forensically relevant body fluid of semen: A preliminary study. *Forensic Science International: Genetics*, 19:28-34. https://doi.org/https://doi.org/10.1016/j.fsigen.2015.05.014

Lee HY, Hong SR, Lee JE, et al., 2020. Epigenetic age signatures in bones. *Forensic Science International Genetics*, 46:102261. https://doi.org/10.1016/j.fsigen.2020.102261

Lee JW, Choung CM, Jung JY, et al., 2018. A validation study of DNA methylation-based age prediction using semen in forensic casework samples. *Legal Medicine (Tokyo, Japan)*, 31:74-77. https://doi.org/10.1016/j.legalmed.2018.01.005

Leek JT, Johnson WE, Parker HS, et al., 2012. The sva package for removing batch effects and other unwanted variation in high-throughput experiments. *Bioinformatics*, 28(6):882-883.

Levine ME, Lu AT, Quach A, et al., 2018. An epigenetic biomarker of aging for lifespan and healthspan. *Aging*, 10(4):573-591. https://doi.org/10.18632/aging.101414

Li LY, Song F, Lang M, et al., 2020. Methylation-based age prediction using pyrosequencing platform from seminal stains in han chinese males. *Journal of Forensic Sciences*, 65(2):610-619. https://doi.org/10.1111/1556-4029.14186

Li Q, Hermanson PJ, Springer NM, 2018. Detection of DNA methylation by whole-genome bisulfite sequencing. In: Lagrimini, L.M. (Ed. Maize: Methods and protocols. Springer, New York, NY, p.185-196.

Lu AT, Quach A, Wilson JG, et al., 2019. DNA methylation grimage strongly predicts lifespan and healthspan. *Aging*, 11(2):303-327. https://doi.org/10.18632/aging.101684

Marioni RE, Harris SE, Shah S, et al., 2016. The epigenetic clock and telomere length are independently associated with chronological age and mortality. 45(2):424-432.

Mather KA, Jorm AF, Parslow RA, et al., 2011. Is telomere length a biomarker of aging? A review. *The Journals of Gerontology Series A, Biological Sciences and Medical Sciences*, 66(2):202-213. https://doi.org/10.1093/gerona/glq180

Mcewen LM, O'donnell KJ, Mcgill MG, et al., 2020. The pedbe clock accurately estimates DNA methylation age in pediatric buccal cells. *Proceedings of the National Academy of Sciences of the United States of America*, 117(38):23329-23335. https://doi.org/10.1073/pnas.1820843116

Mcgreevy KM, Radak Z, Torma F, et al., 2023. Dnamfitage: Biological age indicator incorporating physical fitness. *Aging*, 15(10):3904-3938. https://doi.org/10.18632/aging.204538

Meissner C, Von Wurmb N, Schimansky B, et al., 1999. Estimation of age at death based on quantitation of the 4977-bp deletion of human mitochondrial DNA in skeletal muscle. *Forensic Science International*, 105(2):115-124. https://doi.org/10.1016/s0379-0738(99)00126-7

Moran S, Arribas C, Esteller M, 2016. Validation of a DNA methylation microarray for 850,000 cpg sites of the human genome enriched in enhancer sequences. *Epigenomics*, 8(3):389-399.

Naue J, 2023. Getting the chronological age out of DNA: Using insights of age-dependent DNA methylation for forensic DNA applications. *Genes & Genomics*, 45(10):1239-1261. https://doi.org/10.1007/s13258-023-01392-8

Niño-Sandoval TC, Guevara Pérez SV, González FA, et al., 2017. Use of automated learning techniques for predicting mandibular morphology in skeletal class i, ii and iii. *Forensic Science International*, 281:187.e181-187.e187. https://doi.org/10.1016/j.forsciint.2017.10.004

Oakes CC, La Salle S, Smiraglia DJ, et al., 2007. Developmental acquisition of genome-wide DNA methylation occurs prior to meiosis in male germ cells. *Developmental Biology*, 307(2):368-379. https://doi.org/10.1016/j.ydbio.2007.05.002

Onofri M, Delicati A, Marcante B, et al., 2023. Forensic age estimation through a DNA methylation-based age prediction model in the italian population: A pilot study. *International Journal of Molecular Sciences*, 24(6):5381. https://doi.org/10.3390/ijms24065381

Opstad TB, Pettersen AÅ, Arnesen H, et al., 2011. Circulating levels of il-18 are significantly influenced by the il-18 +183 a/g polymorphism in coronary artery disease patients with diabetes type 2 and the metabolic syndrome: An observational study. *Cardiovascular Diabetology*, 10:110. https://doi.org/10.1186/1475-2840-10-110

Ortega-Recalde O, Peat JR, Bond DM, et al., 2021. Estimating global methylationmethylation and erasure using low-coverage whole-genome bisulfite sequencingbisulfite sequencing (bs-seq) (wgbswhole-genome bisulfite sequencing (wgbs)). In: Bogdanovic, O., Vermeulen, M. Eds.), Tet proteins and DNA demethylation: Methods and protocols. Springer US, New York, NY, p.29-44.

Pidsley R, Zotenko E, Peters TJ, et al., 2016. Critical evaluation of the illumina methylationepic beadchip microarray for whole-genome DNA methylation profiling. *Genome Biology*, 17(1):208. https://doi.org/10.1186/s13059-016-1066-1

Pilsner JR, Saddiki H, Whitcomb BW, et al., 2022. Sperm epigenetic clock associates with pregnancy outcomes in the general population. *Human Reproduction*, 37(7):1581-1593. https://doi.org/10.1093/humrep/deac084

Pisarek A, Pośpiech E, Heidegger A, et al., 2021. Epigenetic age prediction in semen – marker selection and model development. *Aging*, 13(15):19145-19164. https://doi.org/10.18632/aging.203399

Rajkumari S, Nirmal M, Sunil PM, et al., 2013. Estimation of age using aspartic acid racemisation in human dentin in indian population. *Forensic Science International*, 228(1-3):38-41. https://doi.org/10.1016/j.forsciint.2013.02.021

Reed K, Poulin ML, Yan L, et al., 2010. Comparison of bisulfite sequencing pcr with pyrosequencing for measuring differences in DNA methylation. *Analytical Biochemistry*, 397(1):96-106. https://doi.org/10.1016/j.ab.2009.10.021

Ro LS, Lai SL, Chen CM, et al., 2003. Deleted 4977-bp mitochondrial DNA mutation is associated with sporadic amyotrophic lateral sclerosis: A hospital-based case-control study. 28(6):737-743.

Ruiz RJ, Trzeciakowski J, Moore T, et al., 2017. Acculturation predicts negative affect and shortened telomere length. *Biological Research for Nursing*, 19(1):28-35. https://doi.org/10.1177/1099800416672005

Sahoo K, Sundararajan V, 2024. Methods in DNA methylation array dataset analysis: A review. *Computational and Structural Biotechnology Journal*, 23:2304-2325. https://doi.org/10.1016/j.csbj.2024.05.015

Sanders JL, Newman AB, 2013. Telomere length in epidemiology: A biomarker of aging, age-related disease, both, or neither? *Epidemiologic Reviews*, 35(1):112-131. https://doi.org/10.1093/epirev/mxs008

Schütte B, El Hajj N, Kuhtz J, et al., 2013. Broad DNA methylation changes of spermatogenesis, inflammation and immune response-related genes in a subgroup of sperm samples for assisted reproduction. *Andrology*, 1(6):822-829. https://doi.org/10.1111/j.2047-2927.2013.00122.x

Schwender K, Holländer O, Klopfleisch S, et al., 2021. Development of two age estimation models for buccal swab samples based on 3 cpg sites analyzed with pyrosequencing and minisequencing. *Forensic Science International Genetics*, 53:102521. https://doi.org/10.1016/j.fsigen.2021.102521

Se J, Kj S, Hy L, 2017. DNA methylation-based age prediction from various tissues and body fluids. *BMB reports*, 50(11):546-553. https://doi.org/10.5483/bmbrep.2017.50.11.175

Siebert-Kuss LM, Dietrich V, Di Persio S, et al., 2024. Genome-wide DNA methylation changes in human spermatogenesis. *American Journal of Human Genetics*, 111(6):1125-1139. https://doi.org/10.1016/j.ajhg.2024.04.017

Suzuki M, Liao W, Wos F, et al., 2018. Whole-genome bisulfite sequencing with improved accuracy and cost. *Genome Research*, 28(9):1364-1371. https://doi.org/10.1101/gr.232587.117

Tengan CH, Ferreiro-Barros C, Cardeal M, et al., 2002. Frequency of duplications in the d-loop in patients with mitochondrial DNA deletions. *Biochimica Et Biophysica Acta*, 1588(1):65-70. https://doi.org/10.1016/s0925-4439(02)00140-0

Teschendorff AE, Menon U, Gentry-Maharaj A, et al., 2010. Age-dependent DNA methylation of genes that are suppressed in stem cells is a hallmark of cancer. *Genome Research*, 20(4):440-446. https://doi.org/10.1101/gr.103606.109

Teschendorff AE, Marabita F, Lechner M, et al., 2013. A beta-mixture quantile normalization method for correcting probe design bias in illumina infinium 450 k DNA methylation data. *Bioinformatics*, 29(2):189-196.

Theda C, Hwang SH, Czajko A, et al., 2018. Quantitation of the cellular content of saliva and buccal swab samples. *Scientific Reports*, 8(1):6944. https://doi.org/10.1038/s41598-018-25311-0

Thodberg HH, Van Rijn RR, Jenni OG, et al., 2017. Automated determination of bone age from hand x-rays at the end of puberty and its applicability for age estimation. *International Journal of Legal Medicine*, 131(3):771-780. https://doi.org/10.1007/s00414-016-1471-8

Vasu V, Turner KJ, George S, et al., 2017. Preterm infants have significantly longer telomeres than their term born counterparts. *PloS One*, 12(6):e0180082. https://doi.org/10.1371/journal.pone.0180082

Vidaki A, Kayser M, 2018. Recent progress, methods and perspectives in forensic epigenetics. *Forensic Science International-Genetics*, 37:180-195. https://doi.org/10.1016/j.fsigen.2018.08.008

Wang J, Wang C, Wei Y, et al., 2022. Circular rna as a potential biomarker for forensic age prediction. *Frontiers in Genetics*, 13:825443. https://doi.org/10.3389/fgene.2022.825443

Wang Q, Gu L, Adey A, et al., 2013. Tagmentation-based whole-genome bisulfite sequencing. *Nature Protocols*, 8(10):2022-2032. https://doi.org/10.1038/nprot.2013.118

Wang Q, Zhan Y, Pedersen NL, et al., 2018. Telomere length and all-cause mortality: A meta-analysis. *Ageing Research Reviews*, 48:11-20. https://doi.org/10.1016/j.arr.2018.09.002

Xiao C, Yi S, Huang D, 2021. Genome-wide identification of age-related cpg sites for age estimation from blood DNA of han chinese individuals. *ELECTROPHORESIS*, 42(14-15):1488-1496. https://doi.org/10.1002/elps.202000367

Xiao C, Li Y, Chen M, et al., 2023. Improved age estimation from semen using sperm-specific age-related cpg markers. *Forensic Science International: Genetics*, 67:102941. https://doi.org/10.1016/j.fsigen.2023.102941

Zapico SC, Ubelaker DH, 2016. Relationship between mitochondrial DNA mutations and aging. Estimation of age-at-death. *The Journals of Gerontology Series A, Biological Sciences and Medical Sciences*, 71(4):445-450. https://doi.org/10.1093/gerona/glv115

Zhang Q, Vallerga CL, Walker RM, et al., 2019. Improved precision of epigenetic clock estimates across tissues and its implication for biological ageing. *Genome Medicine*, 11(1) https://doi.org/10.1186/s13073-019-0667-1

Zubakov D, Liu F, Kokmeijer I, et al., 2016. Human age estimation from blood using mrna, DNA methylation, DNA rearrangement, and telomere length. *Forensic Science International Genetics*, 24:33-43. https://doi.org/10.1016/j.fsigen.2016.05.014

Supplementary information
Tables S1-S6; Figs. S1-S12