International Journal of Legal Medicine
https://doi.org/10.1007/s00414-024-03365-2

ORIGINAL ARTICLE

A methylation panel of 10 CpGs for accurate age inference via stepwise conditional epigenome-wide association study

Yu Qian[1,2] · Qianqian Peng[3] · Qili Qian[3] · Xingjian Gao[4] · Xinxuan Liu[1] · Yi Li[3] · Xiu Fan[1] · Yuan Cheng[1] · Na Yuan[1] · Sibte Hadi[5] · Li Jin[6,7,8] · Sijia Wang[3,9] · Fan Liu[5]

Received: 29 May 2024 / Accepted: 31 October 2024

Abstract
Estimating individual age from DNA methylation at age associated CpG sites may provide key information facilitating forensic investigations. Systematic marker screening and feature selection play a critical role in ensuring the performance of the final prediction model. In the discovery stage, we screened for 811876 CpGs from whole blood of 2664 Chinese individuals ranging from 18 to 83 years of age based on a stepwise conditional epigenome-wide association study (SCEWAS). The SCEWAS identified 28 CpGs showing genome-wide significant and independent effects. Further restricting this panel to 10 most informative CpGs showed a tolerable loss of information. A linear model consisting of these 10 CpGs could explain 93% of the age variance ($R^2 = 0.93$) in the training set (n = 2664). In an independent test set of Chinese individuals (n = 648), this model also provided highly accurate predictions ($R^2 = 0.85$, mean absolute deviation, MAD = 3.20 years). The model was additionally validated in a public dataset of multiple ancestral origins (86 Europeans, 14 Asians, and 273 Africans) and the prediction accuracy reduced significantly ($R^2 = 0.85$, MAD = 6.21 years), as might be expected due to different genomic backgrounds, sample sizes, and age ranges. Our 10 CpG model also outperformed the recently proposed 9-CpG model constructed in 390 Chinese males ($R^2 = 0.79$ in test set). We also demonstrated that our SCEWAS approach outperformed the traditional EWAS and the elastic net approach in obtaining a small set of most age informative CpGs. Overall, our systematic genome-wide feature selection identified a small panel of 10 CpGs for accurate age estimation with high potential in forensic applications.

Keywords DNA methylation · Age prediction · Forensic DNA phenotyping · Epigenome-Wide Association Study · East Asian population

Introduction

The utilization of age-associated CpG markers for accurate age estimation holds promise in various contexts. Such information proves valuable in narrowing the focus of investigations involving suspects, thereby enhancing the efficiency of forensic inquiries [1–3]. Over the past few decades, a multitude of studies with a forensic orientation have centered on CpG-based age prediction.

Weidner et al. investigated three CpG sites (*ITGA2B*, *ASPA*, and *PDE4C*) using pyrosequencing technology in a cohort of 69 European subjects, achieving a mean absolute deviation (MAD) of 4.5 years [4]. Similarly, Zbiec-Piekarska et al. examined seven CpG sites within *ELOVL2*, obtaining a MAD of 5.03 years in a discovery set of 303 samples and 5.75 years in a validation set of 124 samples from a Polish cohort [5]. The same group further developed a model comprising 41 CpG sites, trained on 300 samples and tested on 120 Polish samples, achieving a MAD of 3.9 years [6]. Turning attention to three CpG sites (*ELOVL2*, *ZNF423*, *CCDC102B*) in 535 Korean subjects, Park et al. attained a MAD of 3.16 years [7]. Jung et al. developed a multiplex methylation SNaPshot assay targeting five CpG sites within the *ELOVL2*, *FHL2*, *KLF14*, *MIR29B2CHG/C1orf132* and *TRIM59* genes. Using a combined model trained on 100 Korean samples each from blood, saliva, and buccal swabs, the assay was tested on an additional 50 Korean samples per tissue type, achieving a MAD of 3.844 years [8]. In a study

Yu Qian, Qianqian Peng, and Qili Qian contributed equally to this work.

Sijia Wang and Fan Liu contributed equally to the supervision of this work.

Extended author information available on the last page of the article

Published online: 05 December 2024

by Feng et al., evaluation of nine age-related CpG sites using EpiTYPER technology on 65 Chinese males resulted in a linear regression model for age prediction, achieving a MAD of 2.49 years [9]. In a broader approach, Li et al. proposed a regression model involving 83 CpG sites in a cohort of 90 Chinese children and adolescents, validated with a remarkable MAD of 0.62 across 89 test datasets [10].

These investigations highlight the strong potential of age-associated CpGs for forensic age estimation. However, methodological improvements are needed, as many studies have focused on a limited set of candidate sites rather than conducting a comprehensive genome-wide analysis. Furthermore, the issue of redundant information among markers has seldom been addressed. The preponderance of marker discovery studies relied on data from the Illumina 450 K array chip, which has notably fewer markers compared to the 850 K array. Moreover, the geographical bias in the populations studied, primarily of European or American origin, warrants attention [11]. The representation of well-sized studies in East Asian populations, encompassing a wide age spectrum and both genders, has been notably limited.

Materials and methods

The national survey of physical traits cohort

The National Survey of Physical Traits cohort (NSPT) is a population-based prospective cohort study to explore the environmental and genetic factors associated with physical traits and diseases. The NSPT cohort study was conducted with the official approval of the Shanghai Institutes for Biological Sciences (ER-SIBS-261410). The NSPT totally collected samples of 3,523 Han Chinese individuals from three sites (i.e., Taizhou, Nanning, and Zhengzhou). All individuals provided written informed consent. Phenotype quality controls were conducted together with other projects, after which, this study included a total of 3312 individuals. All samples were measured for DNA methylation using the Illumina Infinium HumanMethylation850 BeadChip but in three different batches, where the first batch (n = 648) was measured in 2018 and second batch (n = 732) and third batch (n = 1932) were separately measured in 2019. In this study, we used the combined sets of the second and the third batches as the training set (n = 2664) and used the first batch as the test set (n = 648). Therefore, the samples used in our model building and test are completely independent.

Blood samples from three Chinese cities (Zhengzhou, Taizhou, Nanning) were sent to Fudan University Taizhou Institute of Health Sciences for storage at − 80 °C until DNA extraction. DNA extraction was performed using a TGuide M48 Automated nucleic acid extractor (MGBio, Shanghai, China). Genome-wide DNA methylation was profiled using the Infinium MethylationEPIC BeadChips (Illumina). Five hundred nanogram of genomic DNA from each whole blood sample was bisulfite converted using the EZ DNA Methylation Kit (Zymo Research). BeadChips were processed following the manufacturer guide and protocol for Infinium MethylationEPIC array. DNA was hybridized to BeadChips and single base extension were performed using a Freedom EVO robot (Tecan). BeadChips were subsequently imaged using the iScan Microarray Scanner (Illumina). Illumina.idat files were then processed with the minfi[12] Bioconductor package1 without background correction (although background correction reduces bias it does so at the expense of increased variance, which is generally something to be avoided, unless the DNAm data are used for copy-number estimation). Probes with SNPs were removed using the dropLociWithSnps function from minfi. This function uses the SNP information provided by Illumina and UCSC Common SNP tables (including version 132, 135, 137, 138, 141, 142, 144, 146, and 147) with preset MAF (0 is the default value and was used here) to filter SNP CpGs. X and Y chromosome data are often excluded from large-scale genomic and epigenomic analyses due to the analytical complexities arising from dosage differences between XX and XY individuals and the effects of X-chromosome inactivation (XCI) on the epigenome [13]. For the purpose of age prediction modeling, the large pool of autosomal CpG markers is sufficient. Therefore, we excluded probes from the X and Y chromosomes in our analysis. We further used the Illumina definition of β values and derived P values of detection for the rest of probes by comparing the total intensity U + M to that of the background distribution (given by negative control probes), as implemented in minfi. β values with P values of detection greater than 0.01 were set to NA. Of note, the threshold of detection ($P < 0.01$) is more stringent than the $P < 0.05$ threshold used in the other cohorts, partly because sample coverages were very high, allowing for a more stringent threshold while also retaining a high coverage over probes. Only probes with less than 5% missing values were retained. The missing β values were then imputed with the impute.knn function (using k = 5) in R. Type-2 probe bias was corrected using Beta-Mixture Quantile Normalization (BMIQ) [14]. Based on principal component analyses, we found a significant slide/beadchip effect. Therefore, we used ComBat[15] on M-values (logit of β values) to correct for the slide effect and then transformed the M-values back to β values. After quality control, 811,876 CpGs were retained. Our methylation data has been uploaded to both the OMIX platform (https://ngdc.cncb.ac.cn/omix/release/OMIX004363) and the NODE platform (https://www.biosino.org/node) under accession number OEP002902. Due

to the Regulations on the Management of Human Genetic Resources in China, the data is currently listed as "Unavailable" on OMIX while the registration process with the Human Genetic Resource Management Platform of MOST is ongoing. If access to the data is required before this process is complete, please contact the corresponding author via email. We ensure that all data usage requests will be managed in strict compliance with the Regulations on the Management of Human Genetic Resources in China.

EWAS Atlas data

In order to validate our prediction model using external data, we downloaded the data from EWAS Atlas (https://ngdc.cncb.ac.cn/ewas/atlas). This data contains blood methylation data of 373 individuals of mixed ancestral origins (86 Europeans, 14 Asians, and 273 Africans). The methylation data was generated using Infinium MethylationEPIC BeadChips. The quality controls of methylation data have been described in details previously [16, 17]. In brief, signal intensities of type I probes between arrays were normalized using an in-house reference-based method called GMQN2. BMIQ was used to correct the bias associated with technical differences between Type I and Type II array designs. Probes with high detection P-values (by default, the threshold is set at 2.2e-16, which is the smallest number that can be stored by the floating system in R program) were removed and samples with more than 20% of the probes with high detection P-values were removed.

Stepwise conditional epigenome-wide association studies (SCEWAS)

The initial epigenome-wide associated study (EWAS) of chronological age was performed using linear model using limma [18]. Methylation beta values were rank normalized prior to the EWAS to ensure that all CpGs follow the normal distributions using the qnorm function of R,qnorm($\frac{\text{rank(x)}-.5}{\text{length(x)}}$). We included 5 genomic CpG PCs as covariates in our EWAS due to the observation that genomic CpG PCs showed highly significant association with age (Table S1), which might be expected as a large proportion of CpGs in the genome are significantly associated with age. Additional covariates included sex, BMI, cell fractions (CD8 + T cells, CD4 + T cells, NK, B cells, monocytes, Neutrophils), sampling location, slide, and batch. Genome-wide significance threshold was set as $p < 6.16e\text{-}8$ based on Bonferroni correction of 0.8 million CpGs.

The initial EWAS provided a large number of age associated CpGs. To obtain a small panel of CpGs for accurate age prediction, which would require each CpG in this panel has independent contribution to the prediction, we developed a computational pipeline of SCEWAS. SCEWAS is carried out in an iterative manner, where the in next round EWAS, the most significant CpG from the previous round EWAS is added as a covariate, until no CpGs can be identified at the genome-wide significance level. In total, SCEWAS iterated 28 times. In this way, SCEWAS guarantees the most significant CpG from a next round EWAS has an effect that is independent of the CpGs identified from all previous rounds of EWAS, thus minimizing the redundant information between CpGs. In the forensic application, the desired trade-off between the number of markers and the model performance often differs from the theoretical optimal, i.e., fewer markers are preferred once the loss of accuracy (R^2) is tolerable. This is mainly because of practical reasons as DNA obtained at a crime scene is often trace, contaminated or mixed. Such DNA is preferably analyzed using platforms such as EpiTYPER[19] or bisulfite multiplex amplicon sequencing [20] by targeting on a small set of CpG markers. Therefore, we decide to stop the SCEWAS if the loss of R^2 ($\Delta R^2 = 0.2\%$) between the current model and the model in a previous EWAS iteration was tolerable.

Simply selecting significant CpGs based on their association p-values from the initial EWAS can also results in a small panel. Another commonly used method is elastic net, which linearly combines the L1 and L2 penalties of the lasso and ridge methods. In our study, we compared the performance of SCEWAS with traditional EWAS and elastic net. The elastic net analysis was carried out only for the CpGs which were genome-wide significant in our initial EWAS, using the R package glmnet and the same set of covariates. The optimal value of Lambda in the elastic net was determined by tenfold cross-validations ($\lambda = 0.443$, alpha = 0.4).

Prediction analysis

Multiple regression (MR) is the most used statistical technique in prediction modeling to examine the linear relationship between multiple independent variables and a dependent variable, enabling the assessment of their combined effects on prediction. MR was constructed using lm function in R.We used Akaike Information Criterion (AIC), AIC~2 k + nln($\sum$i = 1n(yi − E(yi))2) to rank all significant CpGs from SCEWAS, where k is the number of CpGs, n is the sample size, and E(yi) is the fitted value. For comparing different modeling methods, a support vector machine (SVM) was constructed using the svm function in R package 'e1071' with fine-tuned parameters (cost = 100, gamma = 0.001) and an artificial neural network (ANN) was constructed using the nnet function in R package 'nnet' with parameters (size = 10, decay = 0.01, maxit = 1000, linout = T, trace = F). Model performance was evaluated using a range of accuracy

measurements for comparison in previous works, including median absolute deviation (MEAD), MAD, age variance explained by the methylation markers (R^2), correlation between the predicted and observed values (r), root mean square error (RMSE), and the proportions of correct prediction within an error range of ± 5 and/or 6 years. All statistical analyses were conducted using R scripting unless otherwise specified.

Results

SCEWAS

The discovery cohort comprised 2,664 Chinese individuals aged 18–83 years (mean age = 48.6 ± 13.1 years, 33.8% male, Table S2). The initial EWAS in this cohort (N = 2,664) identified 51,348 CpG sites significantly associated with age at the genome-wide level (6.32%, P ≤ 6.16e-8, Fig. S1-S2). The high proportion of age-associated CpGs and a notable genomic inflation factor (λ = 2.1) are consistent with prior studies [7, 10, 21, 22], indicating extensive age-related changes in methylation across the genome. By conditioning on the most significant CpGs from the initial EWAS, our subsequent SCEWAS analysis identified 28 CpGs across 28 distinct genomic regions that independently and significantly associated with age, explaining 94.8% of the variance in age (Fig. 1A, Fig. S1-S2, Table S3). Repeating the SCEWAS analysis without including epigenomic principal components as covariates identified 31 significant CpGs (Table S5), 12 of which overlapped with the initial 28 CpGs. Since the explanatory power of these 31 CpGs (94.8%) was comparable to that of the initial 28 CpGs, we opted to use the 28 CpGs for subsequent modeling to maintain a conservative approach.

In our SCEWAS, the most significant CpG is identified in *ELOVL2* (cg16867657, $P < 1E\text{-}300$), which showed a strong positive correlation with age and explained 83.2% age variance (Fig. S4-S5, Table S3). This CpG has been repeatedly reported as the strongest marker associated with age in previous studies [5, 7, 10, 23–26]. The other 27 CpGs were located in or nearby *AUH, C1orf96, MIR29B2CHG/C1orf132, C1orf201, CCDC102B, EXD2, FAM118B, FHL2, FIGN, GRM2,KIAA0430, KLF14, PAK6, PIGU, PRICKLE2,,PTH2R, RNF180, RP11-398F12.1, SFMBT1, SST, SPAG6* and *TRIM59,* in the descending order of their association significance in the SCEWAS (Table S3). The majority of these loci (*KLF14, PAK6, GRM2, FHL2, MIR29B2CHG/C1orf132, CCDC102B, RNF180, KIAA0430, TRIM59, PIGU, PTH2R, PRICKLE2, AUH, C1orf201, SFMBT1, C1orf96, RP11-398F12.1, FAM118B, EXD2, SST, FIGN, RP11-573G6.8*, 25/28) has been associated with age in previous EWASs, confirming the reliability of our

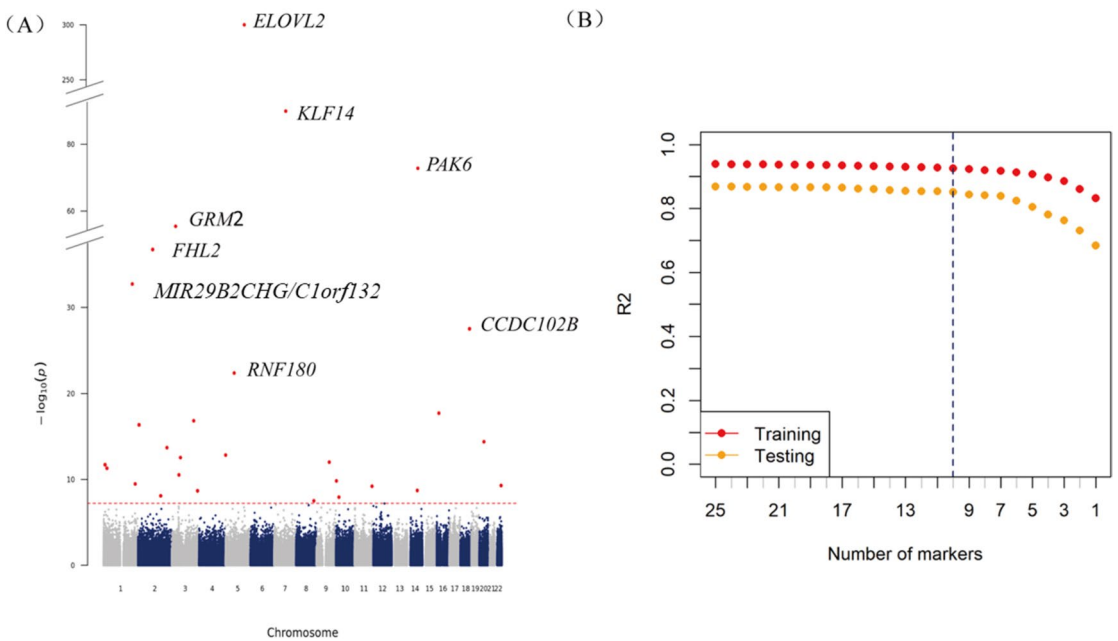

Fig. 1 A) Manhattan plots of the conditional epigenome-wide association of chronological age. The red points represent the independent and significant CpG sites. The red line represents the significance threshold ($P < 6.16e\text{-}8$). B) Feature selection outcomes from the backward stepwise regression (BSR). The BSR selection was conducted beginning with a total of 25 markers that were proposed by conditional EWAS and multiple regression as the theoretical marker set. The red and orange dots represent the model fitting results from the BSR selecting in training and validation set, respectively. The purple dashed line represents the iteration in which the $\Delta r^2 > 0.2\%$

findings [21, 23, 24, 27–33]. In a multiple regression analysis, 25 out of the 28 CpGs showed nominally significant ($P < 0.05$) association with age, which together explained 93.9% age variance (Table S5). This confirms that most of our identified CpGs indeed having independent effects. Note that sex was not significant in simple and multiple regression analysis, thus not further considered in the subsequent analysis.

Prediction model building

In forensic applications, the desired trade-off between the number of CpG markers and the model performance often differs from the theoretical optimal, i.e., fewer markers are preferred as long as the loss of accuracy (R^2) is tolerable. We therefore further reduced our panel to 10 most age-informative CpGs (see method), which together in a MR model explained 92.6% age variance in the discovery cohort (Fig. 1B, Table 1). These 10 CpGs were located in or close to *ELOVL2*, *KLF14*, *CCDC102B, MIR29B2CHG/C1orf132*, *FHL2*, *GRM2*, *RNF180*, *RP11-573G6.8*, *PTH2R*, and *STPG1*, in the descending order of significance. Half of these CpGs showed positive correlations and others showed negative correlations with age (Fig. S3-S5). Although 9 out of these 10 CpGs have been reported in previous EWAS, our model has a novelty. This is because a large number of CpGs over the genome are associated with age, as demonstrated in previous EWASs and our current study, i.e., more than 6% of the CpGs over the genome showed genome-wide significant association with age. It is thus not surprising that 9 out of the 10 CpGs in our age prediction model have been reported for association with age in previous EWASs. However, considering age prediction modeling, five out of the 10 CpGs in our age prediction model have been used in previous age prediction models, while the remaining 5 represent novel markers for constructing age prediction models (Table S6).

We compared the fitness of the linear models constructed using the CpGs selected from our proposed SCEWAS, top-associated CpGs of EWAS, and elastic net under different numbers of CpGs. Constructing linear models by simply accumulating the CpGs according to their association p-values from EWAS performed the worst (Figure S5, Table S7), as it ignores the correlations between CpGs. It was obvious that the fitness of the elastic net was lower than that of SCEWAS for all models ranging from 2 to 25 CpGs and higher than that of EWAS for most models within this range (Figure S5). For example, considering 10 selected CpGs, the model using CpGs from SCEWAS ($R^2 = 0.93$) showed higher fitness than the model using CpGs from elastic net ($R^2 = 0.90$) and EWAS ($R^2 = 0.90$). These results convincingly demonstrated the efficiency of SCEWAS in obtaining a small set of most age informative markers.

Using 10 CpGs as predictive markers, MR, SVM, and ANN-based models were established in the training set (n = 2664). A model fitting analysis showed that the MR and the machine learning models had similar fitness ($R^2_{MR} = 0.93$; $R^2_{SVM} = 0.92$; $R^2_{ANN} = 0.90$; Fig. 2A, Table 2).

In addition, the fitness of our 10-CpG age prediction linear model was assessed in samples of different age groups, genders, and sampling locations (Table S8-S10). The results showed that the fitness reduced slightly in the elderly individuals (≥ 56 years old, $R^2 = 0.32$, Table S8), which is consistent with the previous findings [6, 7, 9]. Sex and different sampling locations had no significant impact on model fitness (Table S9-S10).

Table 1 The top 10 CpGs from SCEWAS in the training set (N = 2664)

Rank CpGs	Region	Gene	Simple regression		Multiple regression		
			Beta	P	Beta	P	Accumulative R^2
(Intercept)	–	–	–	–	20.99	5.12E-27	–
1 cg16867657	6p24.2	*ELOVL2*	163.32	0.00E + 00	54.11	4.14E-115	83.16%
2 cg08097417	7q32.2	*KLF14*	244.71	0.00E + 00	44.68	2.37E-56	85.14%
3 cg13552692	18q22.1	*CCDC102B*	−128.48	0.00E + 00	−24.80	7.27E-55	88.52%
4 cg10501210	1q32.2	*C1orf132*	−97.66	0.00E + 00	−18.16	1.39E-54	89.87%
5 cg06639320	2q12.2	*FHL2*	148.18	0.00E + 00	27.12	1.44E-49	90.66%
6 cg26079664	3p21.2	*GRM2*	88.15	2.29E-316	10.20	1.69E-24	91.23%
7 cg07850154	5q12.3	*RNF180*	−117.16	1.76E-228	−17.39	4.44E-24	91.75%
8 cg18537454	10p12.2	*RP11-573G6.8*	51.32	5.26E-37	13.65	4.01E-23	92.03%
9 cg01949324	2q34	*PTH2R*	−110.49	9.83E-227	−15.09	7.65E-23	92.30%
10 cg21531089	1p36.11	*STPG1*	−99.39	0.00E + 00	−11.38	3.43E-21	92.55%

Model test in an independent Chinese sample

The test set included 648 Chinese individuals (19–71 years old, mean age = 55.52 ± 10.27 years, 54.0% males, Table S2). Applying the 28-CpGs MR model to predict age produced highly accurate prediction results (R^2 = 0.86, MAD = 3.07, ±5 years accuracy = 0.84, ±6 years accuracy = 0.89). Applying the 10-CpG MR model to predict age also produced highly accurate prediction results (R^2 = 0.85, MAD = 3.20, ±5 years accuracy = 0.80, ±6 years accuracy = 0.89, Fig. 2B, Table 2). As expected, the top ranked predictor was identified as cg16867657 in *ELOVL2*, which alone explained 68.4% age variance (Fig. 1B). The accuracy of applying the 10-CpG SVM and ANN models to predict age in the validation set resulted in similar or reduced accuracies (SVM: R^2 = 0.85, MAD = 3.15, ±5 years accuracy = 0.80, ±6 years accuracy = 0.86; ANN: R^2 = 0.79, MAD = 3.40, ±5-year accuracy = 0.79, ±6-year accuracy = 0.85; Table 2).

The prediction analysis was conducted in subgroups of age, sex, and sampling locations (Table S8-S10). The prediction accuracy was slightly reduced in elderly individuals (≥65 years old, MAD = 4.16, ±5 year accuracy = 0.64, ±6 year accuracy = 0.76, Table S8), similar between males and females (Male: MAD = 3.13, ±5 year accuracy = 0.82, ±6 year accuracy = 0.89; Females: MAD = 3.26, ±5 year precision = 0.78, ±6 year precision = 0.85, Table S9), and similar between different sampling locations (Nanning: MAD = 3.42, ±5 years accuracy = 0.76, ±6 years accuracy = 0.83; Taizhou: MAD = 2.96, ±5 years accuracy = 0.84, ±6 years Accuracy = 0.91; Zhengzhou: MAD = 3.07, ±5 years accuracy = 0.84, ±6 years accuracy = 0.90, Table S10). These results are also consistent with previous findings [5, 7, 34].

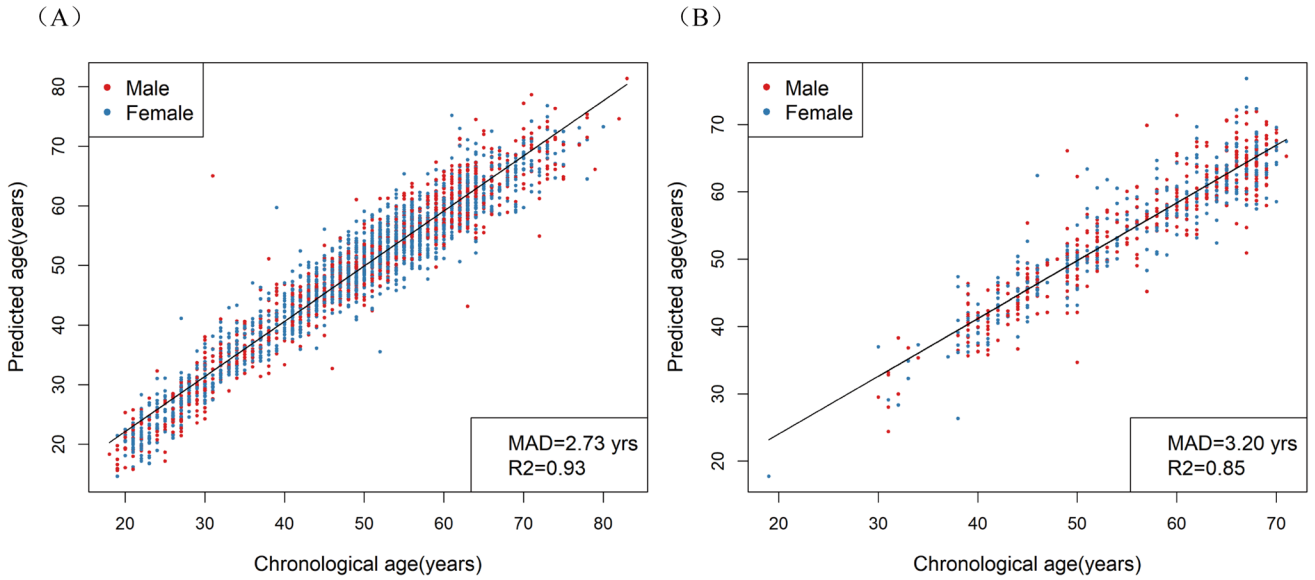

Fig. 2 Scatter plots for chronological age against predicted age. A) Training set (n = 2664), B) Test set (n = 648). The black line is the fitted regression line. Different genders are indicated using colors (red: Male, blue: Female)

Table 2 Performance of the 10-CpG model in predicting age in training (n = 2664) and test (n = 648) sets

Dataset	N	Method	MAD	MEAD	5yrs	6yrs	RMSE	r	R^2
Training	2664	MR	2.73	11.11	0.86	0.91	3.59	0.96	0.93
Testing	648	MR	3.20	10.32	0.80	0.87	4.17	0.92	0.85
Training	2664	SVM	2.64	11.61	0.87	0.92	3.51	0.96	0.92
Testing	648	SVM	3.15	11.86	0.80	0.86	4.17	0.92	0.85
Training	2664	ANN	2.96	11.35	0.84	0.90	4.20	0.95	0.90
Testing	648	ANN	3.40	11.19	0.79	0.85	4.87	0.89	0.79

N: number of samples; MAD: mean absolute deviation; MEAD: median absolute deviation;
RMSE: root mean square error; r: pearson correlation;
R^2: variation explained; 5 yrs: prediction accuracy of ±5 years deviation;
6 yrs: prediction accuracy of ±6 years deviation; BSR: backward step-wise regression;
MR: multiple regression; SVM: support vector machine; ANN: artificial neural network

Model test in individuals of mixed ancestral origins (European, African and Asian)

We also tested our 10-CpG model in a public data set [16, 17]. This test set included 373 individuals (18–69 years old, mean age = 45.00 ± 13.10 years, 55.2% males, 73.2%African, 23.1% European). The prediction accuracy of our 10-CpG MLR model was slightly reduced in individuals of mixed ancestral origins (European, African and Asian) as well as in each ancestral group (R^2 European = 0.84, MAD European = 6.07, R^2 Africa = 0.86, MAD Africa = 6.21, R^2 Asian = 0.95, MAD Asian = 7.35), as might be expected due to different genomic backgrounds, sample sizes and age ranges. Further validations of our model in non-Asian populations with larger sample sizes are warranted in future studies.

Comparison with previous DNAm age prediction studies

To date, there are 20 age prediction models based on methylation markers (Table S11). The prediction accuracies in terms of R^2 ranged from 0.71–0.96 and in terms of MAD ranged from 2.60–7.87 years old [2, 5–7, 9, 23, 24, 26, 29, 33–42]. Besides, the majority (95%) of the test set sample size was between 40 and 583 as well as the gender and age distribution were unknown. The highest R^2 in Europeans was obtained by Freire-Aradas et al. [34] based on 7 methylation sites in 725 training set ($R^2 = 0.96$) and by Vidaki et al.[2] based on 16 methylation sites in 694 training set ($R^2 = 0.96$, Table S11). As for Chinese, the highest R^2 was obtained by Feng et al.[9] ($R^2 = 0.92$) in 390 males.

We compared our 10-CpG model with the 9-CpG model proposed by Feng et al., which represents the currently most accurate model for age prediction of males in Chinese populations. Because the prediction accuracies may differ due to different sample sizes and different phenotype distribution, for a fair comparison we reconstructed the 9-CpG model of Feng et al. in our training set and accessed the accuracy in our test set. The performance of the 9-CpG model in our validation set ($R^2 = 0.78$, MAD = 3.87, Table S12-13) was lower than our 10-CpG model ($R^2 = 0.85$, MAD = 3.20). The performance of the 9-CpG model had similar fitness between males and females (Table S13). This result supports that our 10 age-associated CpGs selected from the genome-wide screening may be more informative in predicting age of unknown samples without information on sex.

Our study aimed to compare the predictive accuracy of different sets of markers on the same platform (array-based), ensuring a fair comparison by controlling for platform-related variability. Specifically, the higher accuracy observed with our 10 CpG model was based on array data. The decreased performance of previously reported markers on a PCR-based platform in our comparisons could be attributed to such platform-related differences. We acknowledge that transforming this model to a PCR-based platform might yield different accuracy results due to differences in platform sensitivity, specificity, and other technical factors. Factors such as primer design, amplification efficiency, and detection methods differ between platforms and could affect the model's performance. Future work should explore these potential differences to further validate our findings.

Discussion

In this study, we present a genome-wide screening analysis of 811,876 CpGs in a large sample of Chinese individuals of both sexes. The SCEWAS appeared an effective approach for screening markers with independent effects, which identified a small panel consisting of 28 age-informative CpGs. A thresholding-based analysis further reduced the panel to 10 CpGs with tolerable loss of information but increased realizability in forensic applications. The 10-CpG model produced high prediction accuracy in a well-sized and independent validation set, demonstrating its high potential in future forensic applications.

While conditional association analysis is an established technique, the novelty of our work lies in its application and methodological integration. We applied SCEWAS for the first time in epigenome-wide association studies, refining a large set of age-associated CpGs to a minimal yet highly predictive panel. This iterative approach optimizes predictive models in contexts where many CpGs are strongly associated with the trait of interest. Regarding CpG marker selection, although many age-associated CpGs have been previously reported, our study identified a unique combination of 10 CpGs, 5 of which have not been used in prior age prediction models (Supplementary Table 5). This specific panel resulted in a MAD of 3.2 years in our test set, highlighting its predictive power. The novelty of our study lies not only in the individual markers but in their synergistic combination and the resulting model's accuracy. Our methodological approach and unique marker combination contribute significantly to advancing age prediction.

In each cycle of SCEWAS, the inclusion of the most significant CpG from the previous cycle as a covariate adjusts the model, thereby changing the statistical landscape for the remaining CpGs. This adjustment can cause shifts in the relative significance of other CpGs. The method ensures that each newly identified CpG provides information independent of previously selected CpGs. As a result, CpGs with initially lower significance might emerge as significant in later cycles once the most dominant CpGs' effects are accounted for. The significance of CpGs is not static; it is recalculated in each

iteration with the updated model. While some CpGs may consistently appear as top hits, others might only reach significance after the most influential CpGs have been included. In addition, epigenetic markers often interact in complex ways. The inclusion of certain CpGs can reveal or obscure the effects of others, leading to a different final set of significant markers than would be identified by simply selecting the top CpGs from initial cycles. Therefore, while the first several CpGs selected in our model were the most significant in the initial cycles, the final set of 10 CpGs resulted from a comprehensive iterative process designed to maximize predictive accuracy by considering conditional dependencies and interactions among all CpGs.

Overviewing the functions of the genes nearby our 10 selected CpGs suggests that these genes play an important role in cell development and differentiation [43–50]. These genes are located on different functional pathways and have different physiological functions, which may make these 10 CpGs have independent predictive effects on age. Five (*ELOVL2* cg16867657, *MIR29B2CHG/C1orf132* cg10501210, *FHL2* cg06639320, *PTH2R* cg01949324, and *KLF14* cg08097417) of the 10 CpGs in our proposed model have been used in previous age prediction models [21, 23, 24, 27–33, 43–47]. Among them, cg16867657, located in the promoter region of the *ELOVL2* gene, is the most extensively reported marker [23, 24, 27–33]. The *ELOVL2* (ELOVL Fatty Acid Elongase 2) gene plays a crucial role in the metabolism of lipids and fatty acids [43]. Age-related changes in methylation levels at *ELOVL2* is robust in most tissues and thus may be widely used for forensic age prediction purpose [26]. *MIR29B2CHG/C1orf132* (MIR29B2 and MIR29C Host Gene) is an RNA gene associated with age-related macular degeneration in a GWAS study [44]. *FHL2* encodes a member of the four-and-a-half-LIM-only protein family, and the protein functions as a link between presenilin-2 and an intracellular signaling pathway. It may also serve as a molecular transmitter, connecting different signaling pathways to transcriptional regulation and playing a role in cell growth [45]. *PTH2R* (Parathyroid Hormone 2 Receptor) encodes a specific receptor for parathyroid hormone, and its mutation is associated with age-related degenerative changes in the lumbar spine [47]. *KLF14* (Kruppel Like Factor 14) encodes a member of the Kruppel-like family of transcription factors which functions as a transcriptional co-repressor, and is associated with multiple metabolic phenotypes [46].

Four out of the 10 CpGs (*C1orf201* cg21531089, *GRM2* cg26079664, *RNF180* cg07850154 and *CCDC102B* cg13552692), although not used in previous age prediction models, have been reported to be significantly associated with age in previous EWASs [21, 23, 24, 28–30]. *C1orf201* (Sperm Tail PG-Rich Repeat Containing 1) plays an important role in the phylogeny of organisms. *GRM2* (Glutamate Metabotropic Receptor 2) encodes glutamate g protein-coupled receptor, which triggers signals through guanine nucleotide-binding protein (g protein) and regulates the activity of downstream effectors (such as adenylate cyclase). It plays an important role in mediating the inhibition of nerve conduction and synapse formation or synapse stabilization. *RNF180* (Ring Finger Protein 180) encodes E3 ubiquitin protein ligase, promoting polyubiquitination and degradation through the ZIC2 proteasome pathway. The methylation level of its promoter was a marker of gastric cancer and atrophic gastritis and was related to the survival rate of gastric cancer patients [48–50]. *CCDC102B* (Coiled-Coil Domain Containing 102B) gene variants are associated with diseases such as vision and autism.

One (*RP11-573G6.8* cg18537454) of the 10 CpGs has not been previously associated with age nor used in age prediction models. The cg18537454 is located on the promoter region of *RP11-573G6.8*. It is a LncRNA Gene and there is no clear function record in the literature for this gene and why it can be used as a predictor to predict age still needs further research.

Our study has a number of advantages. Both the discovery and the replication sets are well sized. The high-density methylation array (Illumina 850 K) has a nearly two-fold increased coverage than the 450 K array, which was typically used in previous EWASs. The SCEWAS approach guarantees the resultant markers have independent effects, which is widely applicable to genome-wide screening studies of similar purposes. Compared with the 9-CpG model proposed by Feng et al., our 10-CpG model appeared more accurate in a fair comparison using the same dataset, and applicable to both sexes.

DNA obtained at a crime scene is often of limited amount. In forensic practice, such DNA cannot be analyzed using genome-wide DNA methylation Epic arrays but preferably using more targeted platforms such as EpiTYPER [19] or bisulfite multiplex PCR followed by sequencing on the MiSeq FGx platform [20]. In our study, the usage of the Infinium MethylationEPIC BeadChips is to select a small set of the most informative CpGs in a systematic manner. Although this approach significantly improved prediction accuracy, model parameters derived from our BeadChip platform may not be directly transferable to other targeted platforms and thus require further validation.

Regarding the limitations of this study, the performance of our prediction model on cross-platform data was not investigated, which should be explored in future studies. For the lack of data on adolescents and children, the applicability of this model in adolescents and children remained unknown and needed further research.

Conclusions

Stepwise conditional EWAS was used for screening CpG markers with independent age effect at the genome-wide level and a systematic feature selection identified 10 CpGs as the optimal subset for age prediction. A linear model consisting of 10 CpGs showed higher accuracy than previous studies in a large and independent validation sets, demonstrating high potential in forensic applications. Competing machine learning models such as ANN and SVM did not outperform the linear model to any obvious degree. Additional analysis showed our model has high prediction accuracy and is applicable for both Chinese males and females with a large age span. Our proposed model is useful in forensic application in East Asian and other populations.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s00414-024-03365-2.

Acknowledgements This project was supported by the Strategic Priority Research Program of Chinese Academy of Sciences (Grant No. XDB38010400, XDC01000000, XDB38020400), Shanghai Municipal Science and Technology Major Project (Grant No. 2017SHZDZX01), National Natural Science Foundation of China (NSFC) (81930056, 32325013, 92249302, 32471216), Science and Technology Service Network Initiative of Chinese Academy of Sciences (KFJ-STS-QYZD-2021-08-001, KFJ-STSZDTP-079), the CAS Youth Innovation Promotion Association (Grant No. 2020276), National Key Research and Development Project 2018YFC0910403, Shanghai Science and Technology Commission Excellent Academic Leaders Program (22XD1424700), CAS Interdisciplinary Innovation Team Project, Max Planck-CAS Paul Gerson Unna Independent Research Group Leadership Award. This work was also supported by Naif Arab University for Security Sciences (Grant No. NAUSS-23-R17).

Author contributions Conceptualization: Fan Liu, Sijia Wang.
Data curation: Yu Qian, Qianqian Peng, Qili Qian, Xingjian Gao, Xinxuan Liu, Yi Li, Xiu Fan, Yuan Cheng, Na Yuan.
Formal analysis: Yu Qian, Qianqian Peng, Qili Qian.
Validation: Yu Qian, Xingjian Gao, Xinxuan Liu, Yi Li, Xiu Fan, Yuan Cheng.
Funding acquisition: Fan Liu, Sijia Wang, Sibte Hadi, Li Jin.
Supervision: Fan Liu, Sijia Wang.
Writing-original draft: Fan Liu, Yu Qian.
Final approval: All coauthors.

Data Availability Individual level data that support the findings of this study are available from the corresponding author upon reasonable request.

Declaration

Competing interests The authors declare that they have no competing interests.

References

1. Parson W (2018) Age estimation with DNA: from forensic DNA fingerprinting to forensic (Epi)Genomics: a mini-review. Gerontology 64:326–332. https://doi.org/10.1159/000486239
2. Vidaki A, Kayser M (2018) Recent progress, methods and perspectives in forensic epigenetics. Forensic Sci Int Genet 37:180–195. https://doi.org/10.1016/j.fsigen.2018.08.008
3. Nie YC, Yu LJ, Guan H et al (2017) Research progress on the detection method of DNA methylation and its application in forensic science. Fa Yi Xue Za Zhi 33:293–300. https://doi.org/10.3969/j.issn.1004-5619.2017.03.017
4. Weidner CI, Lin Q, Koch CM et al (2014) Aging of blood can be tracked by DNA methylation changes at just three CpG sites. Genome Biol 15:R24. https://doi.org/10.1186/gb-2014-15-2-r24
5. Zbiec-Piekarska R, Spolnicka M, Kupiec T et al (2015) Examination of DNA methylation status of the ELOVL2 marker may be useful for human age prediction in forensic science. Forensic Sci Int Genet 14:161–167. https://doi.org/10.1016/j.fsigen.2014.10.002
6. Zbiec-Piekarska R, Spolnicka M, Kupiec T et al (2015) Development of a forensically useful age prediction method based on DNA methylation analysis. Forensic Sci Int Genet 17:173–179. https://doi.org/10.1016/j.fsigen.2015.05.001
7. Park JL, Kim JH, Seo E et al (2016) Identification and evaluation of age-correlated DNA methylation markers for forensic use. Forensic Sci Int Genet 23:64–70. https://doi.org/10.1016/j.fsigen.2016.03.005
8. Jung SE, Lim SM, Hong SR, Lee EH, Shin KJ, Lee HY (2019) DNA methylation of the ELOVL2, FHL2, KLF14, C1orf132/MIR29B2C, and TRIM59 genes for age prediction from blood, saliva, and buccal swab samples. Forensic Sci Int Genet 38:1–8. https://doi.org/10.1016/j.fsigen.2018.09.010
9. Feng L, Peng F, Li S et al (2018) Systematic feature selection improves accuracy of methylation-based forensic age estimation in Han Chinese males. Forensic Sci Int Genet 35:38–45. https://doi.org/10.1016/j.fsigen.2018.03.009
10. Chunxiao Li WG, Gao Y, Canqing Yu, Lv J, Lv R, Duan J, Sun Y, Guo X, Cao W, Li L (2018) Age prediction of children and adolescents aged 6–17 years: an epigenome-wide analysis of DNA methylation. Aging 10:1015–1026
11. Horvath S, Gurven M, Levine ME et al (2016) An epigenetic clock analysis of race/ethnicity, sex, and coronary heart disease. Genome Biol 17:171. https://doi.org/10.1186/s13059-016-1030-0
12. 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 30:1363–1369. https://doi.org/10.1093/bioinformatics/btu049
13. Inkster AM, Wong MT, Matthews AM, Brown CJ, Robinson WP (2023) Who's afraid of the X? Incorporating the X and Y chromosomes into the analysis of DNA methylation array data. Epigenetics Chromatin 16:1. https://doi.org/10.1186/s13072-022-00477-0
14. 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 29:189–196. https://doi.org/10.1093/bioinformatics/bts680
15. Johnson WE, Li C, Rabinovic A (2007) Adjusting batch effects in microarray expression data using empirical Bayes methods. Biostatistics (Oxford, England) 8:118–127. https://doi.org/10.1093/biostatistics/kxj037
16. Xiong Z, Li M, Yang F et al (2020) EWAS Data Hub: a resource of DNA methylation array data and metadata. Nucleic Acids Res 48:D890–D895. https://doi.org/10.1093/nar/gkz840
17. Xiong Z, Li M, Ma Y, Li R, Bao Y (2021) GMQN: A reference-based method for correcting batch effects as well as probes bias in HumanMethylation BeadChip. Preprint https://www.biorxiv.org/content/10.1101/2021.09.06.459116.abstract
18. Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, Smyth GK (2015) limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res 43(7):e47. https://doi.org/10.1093/nar/gkv007
19. Peng F, Feng L, Chen J et al (2019) Validation of methylation-based forensic age estimation in time-series bloodstains on FTA

cards and gauze at room temperature conditions. Forensic Sci Int Genet 40:168–174. https://doi.org/10.1016/j.fsigen.2019.03.006
20. Woźniak A, Heidegger A, Piniewska-Róg D et al (2021) Development of the VISAGE enhanced tool and statistical models for epigenetic age estimation in blood, buccal cells and bones. Aging 13:6459–84. https://doi.org/10.18632/aging.202783
21. Zaghlool SB, Al-Shafai M, Al Muftah WA, Kumar P, Falchi M, Suhre K (2015) Association of DNA methylation with age, gender, and smoking in an Arab population. Clin Epigenetics 7:6. https://doi.org/10.1186/s13148-014-0040-6
22. Bell JT, Tsai PC, Yang TP et al (2012) Epigenome-wide scans identify differentially methylated regions for age and age-related phenotypes in a healthy ageing population. PLoS Genet 8:e1002629. https://doi.org/10.1371/journal.pgen.1002629
23. Hannum G, Guinney J, Zhao L et al (2013) Genome-wide methylation profiles reveal quantitative views of human aging rates. Mol Cell 49:359–367. https://doi.org/10.1016/j.molcel.2012.10.016
24. Florath I, Butterbach K, Müller H, Bewerunge-Hudler M, Brenner H (2014) Cross-sectional and longitudinal changes in DNA methylation with age: an epigenome-wide analysis revealing over 60 novel age-associated CpG sites. Hum Mol Genet 23:1186–1201. https://doi.org/10.1093/hmg/ddt531
25. McCartney DL, Zhang F, Hillary RF et al (2019) An epigenome-wide association study of sex-specific chronological ageing. Genome Med 12:1. https://doi.org/10.1186/s13073-019-0693-z
26. Horvath S (2013) DNA methylation age of human tissues and cell types. Genome Biol 14(10):R115
27. Chitrala KN, Hernandez DG (2020) Race-specific alterations in DNA methylation among middle-aged African Americans and Whites with metabolic syndrome. Epigenetics 15:462–482. https://doi.org/10.1080/15592294.2019.1695340
28. McCartney DL, Zhang F, Hillary RF et al (2019) An epigenome-wide association study of sex-specific chronological ageing. Genome Med 12:1. https://doi.org/10.1186/s13073-019-0693-z
29. Alsaleh H, Haddrill PR (2019) Identifying blood-specific age-related DNA methylation markers on the Illumina MethylationEPIC® BeadChip. Forensic Sci Int 303:109944. https://doi.org/10.1016/j.forsciint.2019.109944
30. Wang Y, Karlsson R (2018) Epigenetic influences on aging: a longitudinal genome-wide methylation study in old Swedish twins. Epigenetics 13:975–987. https://doi.org/10.1080/15592294.2018.1526028
31. Benton MC, Sutherland HG, Macartney-Coxson D, Haupt LM, Lea RA, Griffiths LR (2017) Methylome-wide association study of whole blood DNA in the Norfolk Island isolate identifies robust loci associated with age. Aging 9:753–68. https://doi.org/10.18632/aging.101187
32. Garagnani P, Bacalini MG, Pirazzini C et al (2012) Methylation of ELOVL2 gene as a new epigenetic marker of age. Aging Cell 11:1132–1134. https://doi.org/10.1111/acel.12005
33. Li C, Gao W, Gao Y et al (2018) Age prediction of children and adolescents aged 6–17 years: an epigenome-wide analysis of DNA methylation. Aging (Albany NY) 10:1015–26. https://doi.org/10.18632/aging.101445
34. Freire-Aradas A, Phillips C, Mosquera-Miguel A et al (2016) Development of a methylation marker set for forensic age estimation using analysis of public methylation data and the Agena Bioscience EpiTYPER system. Forensic Sci Int Genet 24:65–74. https://doi.org/10.1016/j.fsigen.2016.06.005
35. Yi SH, Jia YS, Mei K, Yang RZ, Huang DX (2015) Age-related DNA methylation changes for forensic age-prediction. Int J Legal Med 129:237–244. https://doi.org/10.1007/s00414-014-1100-3
36. Xu Y, Li X, Yang Y, Li C, Shao X (2019) Human age prediction based on DNA methylation of non-blood tissues. Comput Methods Programs Biomed 171:11–18. https://doi.org/10.1016/j.cmpb.2019.02.010
37. Huang Y, Yan J, Hou J, Fu X, Li L, Hou Y (2015) Developing a DNA methylation assay for human age prediction in blood and bloodstain. Forensic Sci Int Genet 17:129–136. https://doi.org/10.1016/j.fsigen.2015.05.007
38. Zubakov D, Liu F, Kokmeijer I et al (2016) Human age estimation from blood using mRNA, DNA methylation, DNA rearrangement, and telomere length. Forensic Sci Int Genet 24:33–43. https://doi.org/10.1016/j.fsigen.2016.05.014
39. Cho S, Jung SE, Hong SR et al (2017) Independent validation of DNA-based approaches for age prediction in blood. Forensic Sci Int Genet 29:250–256. https://doi.org/10.1016/j.fsigen.2017.04.020
40. Naue J, Hoefsloot HCJ, Mook ORF et al (2017) Chronological age prediction based on DNA methylation: massive parallel sequencing and random forest regression. Forensic Sci Int Genet 31:19–28. https://doi.org/10.1016/j.fsigen.2017.07.015
41. Xu C, Qu H, Wang G et al (2015) A novel strategy for forensic age prediction by DNA methylation and support vector regression model. Sci Rep 5:17788. https://doi.org/10.1038/srep17788
42. Pan C, Yi S, Xiao C, Huang Y, Chen X, Huang D (2020) The evaluation of seven age-related CpGs for forensic purpose in blood from Chinese Han population. Forensic Sci Int Genet 46:102251. https://doi.org/10.1016/j.fsigen.2020.102251
43. Slieker RC, Relton CL, Gaunt TR, Slagboom PE, Heijmans BT (2018) Age-related DNA methylation changes are tissue-specific with ELOVL2 promoter methylation as exception. Epigenetics Chromatin 11:25. https://doi.org/10.1186/s13072-018-0191-3
44. Holliday EG, Smith AV, Cornes BK et al (2013) Insights into the genetic architecture of early stage age-related macular degeneration: a genome-wide association study meta-analysis. PLoS ONE 8:e53830. https://doi.org/10.1371/journal.pone.0053830
45. Wang C, Lv X, He C, Davis JS, Wang C, Hua G (2020) Four and a half LIM domains 2 (FHL2) contribute to the epithelial ovarian cancer carcinogenesis. Int J Mol Sci 21:7751. https://doi.org/10.3390/ijms21207751
46. Small KS, Hedman AK, Grundberg E et al (2011) Identification of an imprinted master trans regulator at the KLF14 locus related to multiple metabolic phenotypes. Nat Genet 43:561–564. https://doi.org/10.1038/ng.833
47. Åkesson K, Tenne M, Gerdhem P, Luthman H, McGuigan FE (2015) Variation in the PTH2R gene is associated with age-related degenerative changes in the lumbar spine. J Bone Miner Metab 33:9–15. https://doi.org/10.1007/s00774-013-0550-x
48. Deng J, Guo J, Guo X et al (2016) Mediation of the malignant biological characteristics of gastric cancer cells by the methylated CpG islands in RNF180 DNA promoter. Oncotarget 7:43461–74. https://doi.org/10.18632/oncotarget.9494
49. Han F, Sun LP, Liu S et al (2016) Promoter methylation of RNF180 is associated with H.pylori infection and serves as a marker for gastric cancer and atrophic gastritis. Oncotarget 7:24800–9. https://doi.org/10.18632/oncotarget.8523
50. Xie XM, Deng JY, Hou YC et al (2015) Evaluating the clinical feasibility: the direct bisulfite genomic sequencing for examination of methylated status of E3 ubiquitin ligase RNF180 DNA promoter to predict the survival of gastric cancer. Cancer Biomarkers : Section Dis Markers 15:259–265. https://doi.org/10.3233/cbm-150466

Authors and Affiliations

Yu Qian[1,2] · Qianqian Peng[3] · Qili Qian[3] · Xingjian Gao[4] · Xinxuan Liu[1] · Yi Li[3] · Xiu Fan[1] · Yuan Cheng[1] · Na Yuan[1] · Sibte Hadi[5] · Li Jin[6,7,8] · Sijia Wang[3,9] · Fan Liu[5]

✉ Fan Liu
fliu@nauss.edu.sa

Sijia Wang
wangsijia@sinh.ac.cn

1 Key Laboratory of Genomic and Precision Medicine, Beijing Institute of Genomics, China National Center for Bioinformation, Chinese Academy of Sciences, Beijing, China

2 Beijing No.8 High School, Beijing, China

3 CAS Key Laboratory of Computational Biology, Shanghai Institute of Nutrition and Health, University of Chinese Academy of Sciences, Chinese Academy of Sciences, 320 Yue Yang Road, Shanghai 200031, China

4 National Clinical Research Center of Kidney Diseases, Jinling Hospital, Nanjing, Jiangsu, China

5 Department of Forensic Sciences, College of Criminal Justice, Naif Arab University of Security Sciences, Riyadh 11452, Kingdom of Saudi Arabia

6 Human Phenome Institute, Fudan University, Shanghai, China

7 Taizhou Institute of Health Sciences, Fudan University, Taizhou, Jiangsu, China

8 State Key Laboratory of Genetic Engineering and Ministry of Education Key Laboratory of Contemporary Anthropology, Collaborative Innovation Center for Genetics and Development, School of Life Sciences, Fudan University, Shanghai, China

9 Center for Excellence in Animal Evolution and Genetics, Chinese Academy of Sciences, Kunming 650223, China