

DNA Methylation-based Age Prediction from Blood Samples of French Child and Adult Individuals

Camille Ropert¹, Simon Buré¹, Adèle Sourisce¹, Caroline Gallois¹, Myriam Siffointe¹, Laurent Bartholin², Fabrice Besacier¹, Romain Appourchaux^{1*}, and Mathieu Gabut^{1*}

¹Laboratoire de Police Scientifique de Lyon, Service National de Police Scientifique, Ecully, France

²Bureau de l'Innovation, Sous-direction de la stratégie, de l'innovation et du pilotage, Service National de Police Scientifique, Ecully, France

*equal contribution

Abstract

Objective: Identifying the origin of human biological traces detected at crime scenes by comparing DNA profiles to national or international forensic databases is often key to provide new orientations to police investigations. However, when unknown profiles are established, investigators can benefit from forensic genetics to propose new leads, for example by predicting the physical appearance of individuals. Since blood traces are of primary interest for forensic investigators and often lead to the extraction of usable genetic material, in this study, we developed a methodology to predict the biological age from blood samples based on the analysis of DNA methylation of human genomic regions.

Methods: We first established a cohort of blood samples obtained from 170 French donors aged from 0 to 101 years old. We analyzed the methylation status of 5 age-associated CpG sites using the SNaPshot method, a primer-extension based assay routinely used in the French forensic police laboratories. Using a training set of 136 samples, we generated an age-prediction model based on multiple regression analyses of DNA methylation data and we tested its predictive performances on a validation set.

Results: The SNaPshot assay was adapted to limiting quantities of genomic DNA relevant for forensic investigations. The DNA methylation levels were established for 5 age-related CpG sites in 170 blood samples collected from French male and female donors. We established a statistical model optimized for 5 CpG sites that can explain 97% of age variation with a Mean Absolute Error (MAE) of 3.45 years between the estimated biological and chronological age of individuals.

Conclusion: We developed an approach to predict the biological age of individuals strictly based on the methylation levels of 5 CpG sites from circulating blood samples and that is compatible with routine genetic analyses in French forensic police laboratories.

Keywords: DNA methylation • Bisulfite conversion • SNaPshot • Age prediction • Mathematical model • Chronological and biological age comparison • Blood samples

Introduction

In the early 20th century, Alphonse Bertillon introduced anthropometric measurements as a statistically sound approach for forensic identification in criminal investigations in France [1]. More than a century later, criminal investigations heavily rely on molecular biology and genetics to establish DNA profiles, based on short tandem repeat sequences analysis, that can be compared to national or international forensic DNA databases in order to identify individuals. When established DNA profiles remain unidentified, an alternative genetic approach can be proposed to orientate investigation leads, narrow down potential suspects and help reveal the truth. Indeed, during the last decade, a broad base of forensic genetics studies have focused on predicting phenotypic traits such as gender, skin, eye and hair colour as well as biogeographic origins of individuals, and, to a lower extent predisposition

to baldness and freckles [2-8]. While these phenotypic features were mainly inferred from DNA Polymorphic Nucleotide Region (SNP) analyses [4-6,9,10], predictions of additional features such as the biological age have also emerged from epigenetic studies of DNA methylation [11-16]. DNA methylation is defined as the transfer of a methyl group onto the C5 position of cytosines primarily located at CpG sites throughout the genome. DNA methylation marks undergo highly dynamic patterns to control gene expression, and contribute to define cell and tissue identity as robustly stable marks throughout individual life [12]. Interestingly, several studies have revealed that a number of CpG site subsets rather display variable DNA methylation levels, either gain or loss, during individual aging in humans [17]. These observations at the genomic scale led to the concept of epigenetic clock [12]. The correlation between DNA methylation levels and the chronological age, based on birth declarations, has been driving a large research effort to predict the biological age of individuals from age-related changes in the epigenetic landscape and to evaluate the impact of environmental factors or diseases on human life span [18]. Being capable of predicting the age of an individual also has obvious applications in the forensic field for the analysis of unknown DNA traces collected at crime scenes to facilitate the identification of victims or potential responsibilities.

In the recent years, several studies have proposed age prediction models based on the analysis of DNA methylation changes for different tissues of primary interest in forensic investigations: semen [19], blood [20-22], saliva [22-24], or hard tissues such as teeth and bones [22,25,26]. Concomitantly, different methodologies have been developed to either analyze a limited number of CpG sites (SNaPshot, EpiTYPER, Pyrosequencing) or to perform genome-wide studies (massive parallel sequencing) [2,20,22,24,27-30]. As a

*Address for Correspondence: Mathieu Gabut, Laboratoire de Police Scientifique de Lyon, Service National de Police Scientifique, Ecully, France, E-mail: mathieu.gabut@interieur.gouv.fr, snps-lps69-dip-recherche-developpement@interieur.gouv.fr

Copyright: © 2024 Ropert C, et al. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Received: 02 April, 2024, Manuscript No. jfr-24-131270; Editor Assigned: 04 April, 2024, PreQC No. P-131270; Reviewed: 18 April, 2024, QC No. Q-131270; Revised: 23 April, 2024, Manuscript No. R-131270; Published: 30 April, 2024, DOI: 10.37421/2157-7145.2024.15.606

forensic laboratory, our priority was to select an efficient and cost-effective approach to analyze blood samples that could directly be implemented to the laboratory routine analyses: the SNaPshot assay. Previous studies have reported the development of age-prediction models from blood samples, combining DNA methylation level quantifications by SNaPshot of 5 CpG sites localized in the vicinity of the *ELOVL2*, *FHL2*, *KLF14*, *C1orf132* and *TRIM59* genes, and the use of linear regression models for age predictions [20,24,29-31] (Supplemental Figure 1). These 5 CpG sites were also studied in the context of saliva and buccal swabs and referred to as robust multi-tissue age predictive epigenetic marks [23,32]. In the present study, our main goal was to adapt the SNaPshot assay and statistical age-prediction models to sets of blood samples obtained from the French population. We first challenged the SNaPshot approach using the 5 CpG sites mentioned above for low input DNA samples, to cope with the limiting genomic DNA amounts collected from crime scenes. We analyzed a population of 136 individuals, with equal gender representation and homogeneous age distribution between new borns and centenarians, and we established an optimized age prediction model based on DNA methylation level detection of 5 CpG sites (Supplemental Figure 1).

Materials and Methods

Sample collection

This study was performed in accordance with the recommendations of the French National Ethics Committee legal framework (Comité Consultatif National d'Ethique pour les Sciences de la Vie et de la Santé). All adult participants and parents of minors, under the legal age of 18, signed a written informed consent for research purposes of collected samples. Peripheral blood samples were obtained from 170 living and anonymized French donors, including 78 males, 88 females and 4 donors who did not declare their gender on the consent forms. The donors were evenly distributed among 6 age classes: 0-14 years, 15-29 years, 30-44 years, 45-59 years, 60-74 years and over 75 years. From these samples, 144 (70 females, 70 males and 4 undetermined genders) were used to establish a training population, and the remaining 26 (18 females and 8 males) a validation population. Blood samples were collected and complemented with EDTA and stored at 4 °C until processing or on cotton swabs and processed promptly after collection following standard operating procedures.

DNA extraction and quantification

Genomic DNA was extracted for each blood sample using the Nucleospin™ Plasma XS Kit (Macherey Nagel, Düren, Germany) following manufacturer's instructions. Extracted DNA samples were eluted in 50 µL of 5 mM Tris-HCl and quantified using the Quantifiler™ Trio DNA Quantification Kit (Applied Biosystems, Foster City, CA, USA) following the manufacturer's protocol.

Bisulfite conversion

Genomic DNA samples were resuspended in 20 µL and subjected to bisulfite conversion using the Premium Bisulfite kit (Diagenode, Liège, Belgium) following manufacturer's instructions. The converted DNA samples were eluted in 15 µL of elution buffer from the kit.

To determine the sensitivity and ruggedness of the bisulfite conversion, an

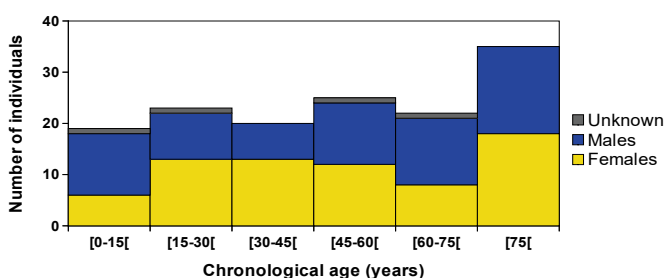


Figure 1. Individual and gender distribution of 144 blood donors across 6 age-classes (n=144). Females, males and undefined genders are indicated in yellow, blue and grey, respectively.

assay was conducted on six different amounts of input DNA (1 ng, 5 ng, 10 ng, 20 ng, 50 ng and 100 ng), obtained by serial dilution of one sample of genomic DNA, and analyzed in triplicates.

To ensure the consistency and reliability of the results for both the training and validation samples, 20 ng of genomic DNA were used to perform the bisulfite conversion and all samples were analyzed in duplicate. A negative control (no DNA) was used to detect a potential contamination. Two additional controls were also included to validate each SNaPshot assay. First, a blood sample with a known profile was systematically included to assess the variability of results between series of SNaPshot analyses. Secondly, a control composed of unconverted and unmethylated human genomic DNA (EpiTect Control DNA Set, Qiagen, Hilden, Germany) was included to assess the bisulfite conversion efficiency.

PCR, multiplex SNaPshot and capillary electrophoresis

The Polymerase Chain Reaction (PCR) steps, multiplex SNaPshot and capillary electrophoresis were carried out under the same conditions as previously described [33,34]. Briefly, five CpG sites, respectively located in the vicinity of the genes *ELOVL2*, *FHL2*, *KLF14*, *C1orf132*, and *TRIM59* were considered in our study. The characteristics of the selected CpG sites are described in Table S1. Converted DNA samples were submitted to multiplex PCR amplification in 20 µL assays containing 2 µL of converted DNA, 4 µL of 5x primers mix (concentrations specified (Table S1), 11.6 µL of pure H₂O, 2 µL of 10X Gold ST[®]R Buffer (Promega Corporation, Madison, WI, U.S.A), and 0.4 µL of AmpliTaq Gold[®] polymerase (Applied Biosystems, Foster City, CA, USA). The PCR amplification program began with an initial denaturation at 95 °C for 11 min, followed by 34 cycles encompassing denaturation at 94 °C for 20 s, annealing at 56 °C for 1 min, and extension at 72 °C for 30 s. Subsequently, the PCR amplification included a final extension step at 72 °C for 7 min.

Next, 2 µL of ExoSAP-IT™ (Applied Biosystems, Foster City, CA, USA) were added to 10 µL of each PCR-amplified products, and the digestion was conducted 45 min at 37 °C followed by 15 min at 80 °C.

The Single Base Extension (SBE) step was performed using the SNaPshot™ multiplex kit (Applied Biosystems, Foster City, CA, USA). 10 µL SBE reactions containing 2 µL of ExoSAP™-treated amplified DNA, 1 µL of a 10X primers mix (concentrations specified in Table S1), 2 µL of 5X Sequencing Buffer BigDye™ termination (Applied Biosystems, Foster City, CA, USA), 1 µL of SNaPshot reaction mix (Applied Biosystems, Foster City, CA, USA), and 4 µL of water were amplified by a PCR sequencing program consisting in 10 s at 96 °C, 5 s at 50 °C and 30 s at 60 °C for 25 cycles.

A final treatment was conducted by adding 1 µL of Shrimp Alkaline Phosphatase (Applied Biosystems, Foster City, CA, USA) to each sample, and incubating the resulting mixes for 45 min at 37 °C followed by 15 min at 8 °C.

The resulting digested SBE products were analyzed using the 3500 xL Genetic Analyzer (Applied Biosystems, Foster City, CA, USA). The methylation rates (0 to 1) at individual CpG sites were determined using the GeneMapper™ Software (version 5) (Applied Biosystems, Foster City, CA, USA). Briefly, we quantified the nucleotide intensities defined by the peak height of the converted and unconverted nucleotides (C or G) and we calculated a ratio of methylated intensities over total peak intensities, as described by Jung SE, et al. [24].

DNA methylation linearity assay

To assess the linearity of DNA methylation detection by the SNaPshot assay for 20 ng of DNA input, completely methylated or completely unmethylated bisulfite converted control DNA samples (EpiTect PCR Control DNA, Qiagen, Hilden, Germany) were mixed to create nine samples with increasing methylation percentages: 0%, 5%, 10%, 25%, 50%, 75%, 90%, 95%, and 100%. These samples were then analyzed in triplicates as previously described, and the measured methylation levels were compared to the expected ratios for each mix.

Data analysis and age prediction models

Independent analyses were conducted for both the training and validation blood sample sets. Data processing was performed using LibreOffice Calc.

(version 7.5.7.1.M1). Statistical analyses were carried out using R (version 4.2.1) and RStudio (version 2023.03.0+386). To predict the biological age of each individual based on the DNA methylation levels of CpG sites, simple and multiple linear regression models were created and the Mean Absolute Error (MAE), the Root Mean Square Error (RMSE), and the Akaike Information Criterion (AIC) were calculated. The coefficient of determination (R^2) were calculated to define the correlation between the chronological and predicted ages. For the training sample set, prediction errors were analyzed in 15-years age classes.

Results

SNaPshot method validation for low input DNA samples

In the recent years, several studies have described the development, for forensic applications, of DNA methylation-based age prediction models from blood samples using the SNaPshot assay (Suppl. Figure 1). Our goal was to propose a methodology to predict the age of unidentified individuals from blood traces discovered at crime scenes, and that was directly compatible with the routine analyses performed in the French police forensic laboratories. To this end, we first re-evaluated the multiplex methylation assay developed by Jung SE, et al. [24] using multi-tissue age methylation CpG sites identified in the vicinity of the *ELOVL2*, *FHL2*, *KLF14*, *C1orf132* and *TRIM59* genes (Suppl. Table 1). We verified the linearity of detection of DNA methylation levels, individually for each CpG site, from already purified and bisulfite converted control unmethylated or methylated genomic DNA samples (Suppl. Figure 2A). Accordingly, performing the SNaPshot assay with DNA sample mixes containing increasing proportions of methylated DNA, ranging from 0 to 100%, revealed a high degree of correlation between the measured and expected methylation levels for the *ELOVL2* ($R^2=0.95$), *FHL2* ($R^2=0.98$), *KLF14* ($R^2=0.97$), *C1orf132* ($R^2=0.99$) and *TRIM59* ($R^2=0.96$) CpG sites (Suppl. Figure 2A).

Since DNA samples collected from crime scenes are often available in limiting amounts, we next assessed the reliability of the bisulfite conversion coupled to the SNaPshot assay for low input genomic DNA samples. Genomic DNA was extracted from a peripheral blood sample and the initial bisulfite conversion was performed on increasing amounts of DNA ranging from 1

ng to 100 ng to detect *KLF14*, *C1orf132*, *ELOVL2*, *FHL2* and *TRIM59* CpGs simultaneously in a multiplex approach (Suppl. Figure 2B). DNA methylation levels measured were consistent with previous studies [20,24,29]. While the *KLF14* and *C1orf132* CpGs displayed low (<10%) and high (>70%) methylation levels, respectively, the *ELOVL2*, *FHL2* and *TRIM59* CpGs were associated to intermediate (35-40%) methylation levels for the blood sample analyzed. This analysis revealed a significantly high variability of the measured DNA methylation levels for the lowest DNA input samples (1 ng, 5 ng and 10 ng) (Suppl. Figure 2B), consistent with previous observations [35]. Interestingly, the methylation values for the 5 CpG sites were however highly consistent when using 20 ng, 50 ng and 100 ng of input DNA for the bisulfite conversion step, with standard deviations ranging from 1% to 2.4%. Therefore, in order to cope with the limiting availability of DNA traces collected in forensic cases while ensuring a highly reproducible detection of DNA methylation for multiple CpG sites, 20 ng of genomic DNA was defined as standard input to perform the bisulfite conversion and multiplex SNaPshot methylation assays in this study.

Establishing a training set of blood samples

The methylation levels of the CpGs in the genes *ELOVL2*, *FHL2*, *KLF14*, *C1orf132* and *TRIM59* were analyzed in 144 out of the 170 peripheral blood samples collected from French individuals between 0 and 101 years old (Figure 1). The blood donors included 4 undeclared genders, 70 female and 70 male participants evenly distributed across 6 age classes (Figure 1). Following genomic DNA extraction and bisulfite conversion of each sample, the methylation levels of the 5 CpGs were simultaneously measured using the SNaPshot assay, as previously described [24]. As expected, all 5 CpG sites showed age-dependant changes in peak distribution for methylated and non-methylated nucleotides on the electrophoregrams (Suppl. Figure 3). The DNA methylation rates of individual CpGs were inferred from the average peak intensity measured in duplicates for each sample. The distribution of methylation values for each CpG site was then analyzed and 8 outlier samples were excluded based on Bonferroni corrected p-values exceeding 0.05. For the remaining 136 samples, changes in DNA methylation rates for the 5 CpG sites were highly correlated with the chronological age of the donors (Figure 2A). The strongest correlations were observed for the CpG sites in *ELOVL2* ($R^2 = 0.92$), *FHL2* ($R^2 = 0.87$), *C1orf132* ($R^2 = 0.84$), *TRIM59* ($R^2 = 0.84$), while the lowest correlation was observed for the *KLF14* CpG site ($R^2 = 0.66$). These results are consistent with previous studies of age determination based on

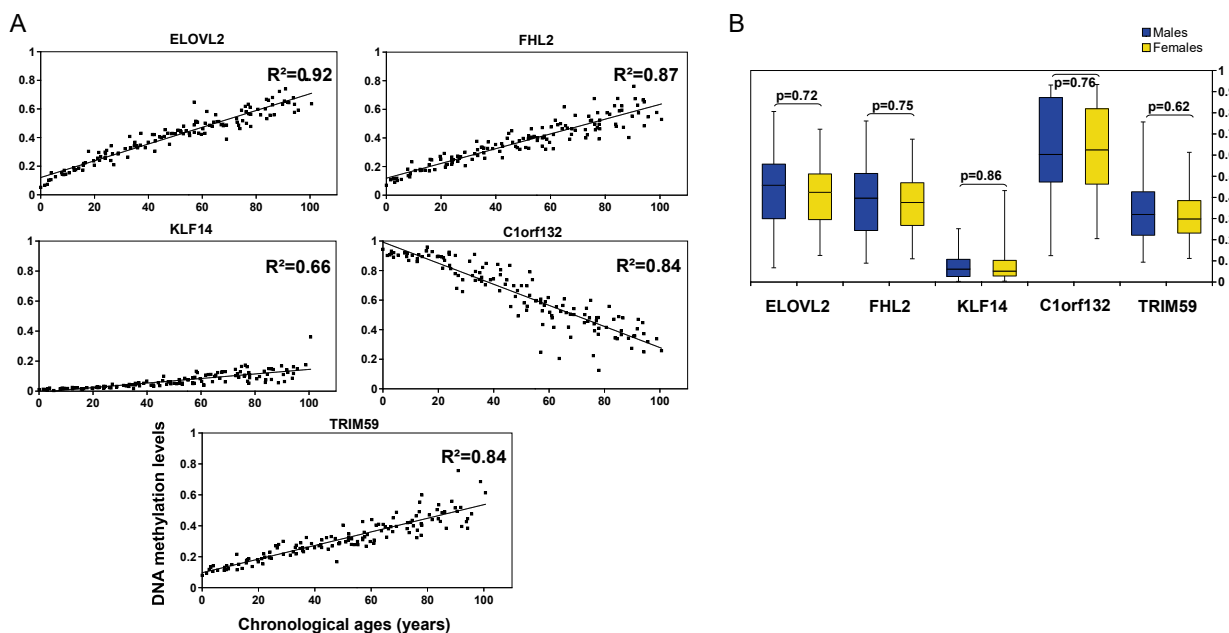


Figure 2. A. Scatter plots representing the correlation between the chronological age and DNA methylation levels at each of the 5 CpG sites analyzed in the *ELOVL2*, *FHL2*, *KLF14*, *C1orf132* and *TRIM59* genes, for a training set composed of 136 blood samples from individuals aged from 0 to 101 years. The lines defining the trend curves and the coefficient of determination (R^2) values are indicated in each graph. B. Box plot representing the distribution of DNA methylation levels for each CpG site in males (blue) and females (yellow) from the training set. The edges of boxes represent the first and the third quartiles respectively, the line within each box represents the median, and the whisker extends represent maximum and minimum values. For each CpG site, the exact p-values of student's t-tests are indicated.

epigenetic DNA modifications in human blood samples [20,24,29,30]. Since the gender distribution is relatively homogeneous across the different age-classes of our sample set (Figure 1), we next confirmed that the distribution of the DNA methylation levels measured was not significantly different between females and males for each of the 5 CpG sites (p-values between 0.62 and 0.86, Figure 2B). In conclusion, we established a training set of 136 blood samples, obtained from 66 females, 66 males and 4 donors of undetermined gender, that can be used to model the French population to assess age-predictive statistical methods (Figure 2).

Development of an age prediction model for blood samples of the French population

Next, we analyzed the predictive potential of each CpG site based on simple linear regression and polynomial regression models (Table 1). Although, a strong correlation was observed between DNA methylation-based predicted ages and chronological ages for all of these models (with R²: 0.81-0.91), their predictive capacities proved to be rather limited with MAE values comprised between 6.18 and 9.05 years, and high AIC values indicative of high prediction error rates (Table 1). To improve the age-predictive capacities for forensic applications, multiple linear regression models including simultaneously the 5 CpG sites as well as several data transformation methods were considered. After comparing multiple modelling methods, an optimized Age Prediction Model (APM) was defined with the following formula:

$$\text{Predicted age (years)} = 55.2403 + 66.0422 \times (\% \text{ ELOVL2 CpG methylation}) + 29.1731 \times (\% \text{ FHL2 CpG methylation}) + 16.4241 \times \log(\% \text{ KLF14 CpG methylation}) + 1.6526 \times \log_2(\% \text{ KLF14 CpG methylation}) - 25.8812 \times (\% \text{ C1orf132 methylation}) + 18.9406 \times (\% \text{ TRIM59 CpG methylation}) \text{ (Table 1).}$$

Using this optimized APM, the biological age of the 136 donors was predicted and compared to their declared chronological age (Figure 3). This APM based on 5 CpGs explained 97% of the total variance observed in the training set (R²=0.97), and its performances were defined with the following metrics: a MAE: ± 3.45 years, a RMSE: ± 4.79 and an AIC of 828. To evaluate the accuracy of the APM predictions, 1000 partial subsets of 43 individuals were randomly generated from the training set. This approach revealed that the model could predict the age of an individual ± 5 years with an accuracy of 75% (Table 2), strictly based on the DNA methylation levels of 5 CpG sites (Table 2 and Figure 3).

Validation of the age prediction model for blood samples

To validate the established APM, an independent validation set composed of the remaining 26 blood samples was analyzed with the SNaPshot approach, similarly to the training sample set. Strong correlations were once again observed between the DNA methylation levels for each of the 5 CpG sites and the chronological age of donors (Figure 4A). Moreover, when applied to the validation set, the APM confirmed a highly significant correlation between the predicted biological and chronological ages with a coefficient of determination of 0.88 (Figure 4B), a MAE of ± 4.49 years and an RMSE of ± 7.17 (Table 2). When combining the training and validation dataset, the APM demonstrated

improved accuracy of age prediction (R²: 0.96, MAE: ± 3.64 and RMSE: ± 5.26) compared to the validation dataset alone (Table 2 and Figure 4).

When analyzing the distribution of the model prediction errors between the chronological and predicted biological ages for the entire 136 individuals of the training set, we noticed that the error range increased along with the age of the donors (Figure 5A). Similar conclusions were described in previously reported studies on blood samples [29,30,35]. Indeed, the median age prediction error was particularly increased for individuals older than 60 years old (Figure 5B). While the APM predictions were quite accurate for the youngest individuals of the training set (0 to 29 years old), we observed a significantly higher dispersion of predicted age error values for volunteers older than 40 years old (Figure 5).

Discussion

The SNaPshot approach has already been described in the past decade to study the relationship between DNA epigenetic modifications (5mC methylation) and chronological ages in blood samples of Korean [24], Portuguese [30], Polish [20], and Italian [29] and Turkish [31] populations, and to propose age-prediction models. Differences of DNA methylation profiles have been reported for specific CpG sites based on the ancestry or biogeographic origin of individuals between Japanese and Germans populations [36] and between Middle East and Central European populations [37]. Consistently, several studies have investigated the potential impact of biogeographic ancestry on DNA methylation based age predictions [36-39]. Altogether, these studies advocate for the development of age-prediction models adapted to the population of interest and to the methodology used. Therefore, in this study, we applied for the first time the SNaPshot approach to study the relationship between DNA methylation and chronological ages in the French population, in the context of a forensic laboratory. Previous studies investigating the importance of DNA methylation changes for age prediction from blood samples relied on 100 ng [31], 40-200 ng [24], 200-400 ng [30], 400 ng [29] and 2 µg [20] of input DNA for bisulfite conversion. However, forensic investigations often depend on limited or even rare biological material collected from crime scenes, therefore impacting the number and type of molecular and genetic analyses that can be performed in comparisons with studies conducted in the frame of fundamental research laboratories. Along this line, our objective was to perform bisulfite conversion coupled to SNaPshot assays from scarce DNA input samples. Our results demonstrate that this methodology can be applied with high reproducibility for as little as 20 ng of genomic DNA, a quantity we decided to use as a standard input for this study to cope with limitations of forensic studies. Interestingly, the results obtained for DNA methylation detection and age prediction accuracy are quite comparable with previous studies performed from significantly larger DNA input samples, therefore suggesting that this limit may actually be surpassed in the future. Accordingly, our data also strongly suggest these levels could be lowered to 5 ng of input DNA (Suppl. Figure 2B). Additional experiments should be performed on a wider sample cohort and by increasing the number of technical replicates per sample to confirm this hypothesis and therefore increase the

Table 1. Correlation analysis between predicted and chronological ages by single linear regression models for the training set (n=136). R²: Coefficient of Determination, MAE: Mean Absolute Prediction Error, RMSE: Root-Mean-Square Error, AIC: Akaike Information Criteria (estimator of prediction error).

CpG associated genes	R ²	MAE (years)	RMSE	AIC	Correct predictions ± 5 years	Age predictive models
<i>ELOVL2</i>	0,91	6,18	8,24	1008	54 %	Simple linear regression
<i>FHL2</i>	0,85	7,73	10,57	1083	46 %	3 rd order polynomial regression
<i>KLF14</i>	0,81	9,05	12,01	1127	36 %	3 rd order polynomial regression
<i>C1orf132</i>	0,87	8,15	10,06	1082	40 %	2 nd order polynomial regression
<i>TRIM59</i>	0,86	7,71	10,36	1082	43 %	2 nd order polynomial regression

Table 2. Correlation analysis between predicted and chronological ages by the multiple regression model (APM) for the different blood sample sets. R²: coefficient of determination, MAE: mean absolute prediction error, RMSE: root-mean-square error, AIC: Akaike information criteria (estimator of prediction error), n.a.: not applicable.

	R ²	MAE (years)	RMSE	AIC	Correct predictions ± 5 years
Training set	0,97	3,45	4,79	828	75 %
Validation set	0,88	4,49	7,17	n.a.	62 %
Combined	0,96	3,64	5,26	n.a.	73%

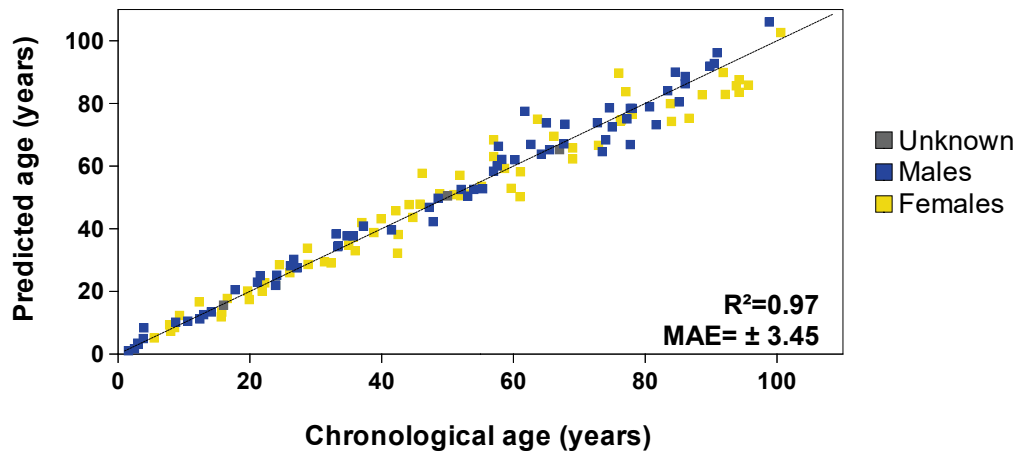


Figure 3. Scatter plot representing the correlation between the chronological age and the age predicted by a multiple regression model (APM) trained on the methylation levels detected at *ELOVL2*, *FHL2*, *KLF14*, *C1orf132*, and *TRIM59* CpG sites for blood samples from new born to 101 years old individuals (n=136). Females, males and undefined individuals are represented by yellow, blue and grey squares, respectively. The line and R^2 value indicate the trend curve and the coefficient of determination, respectively.

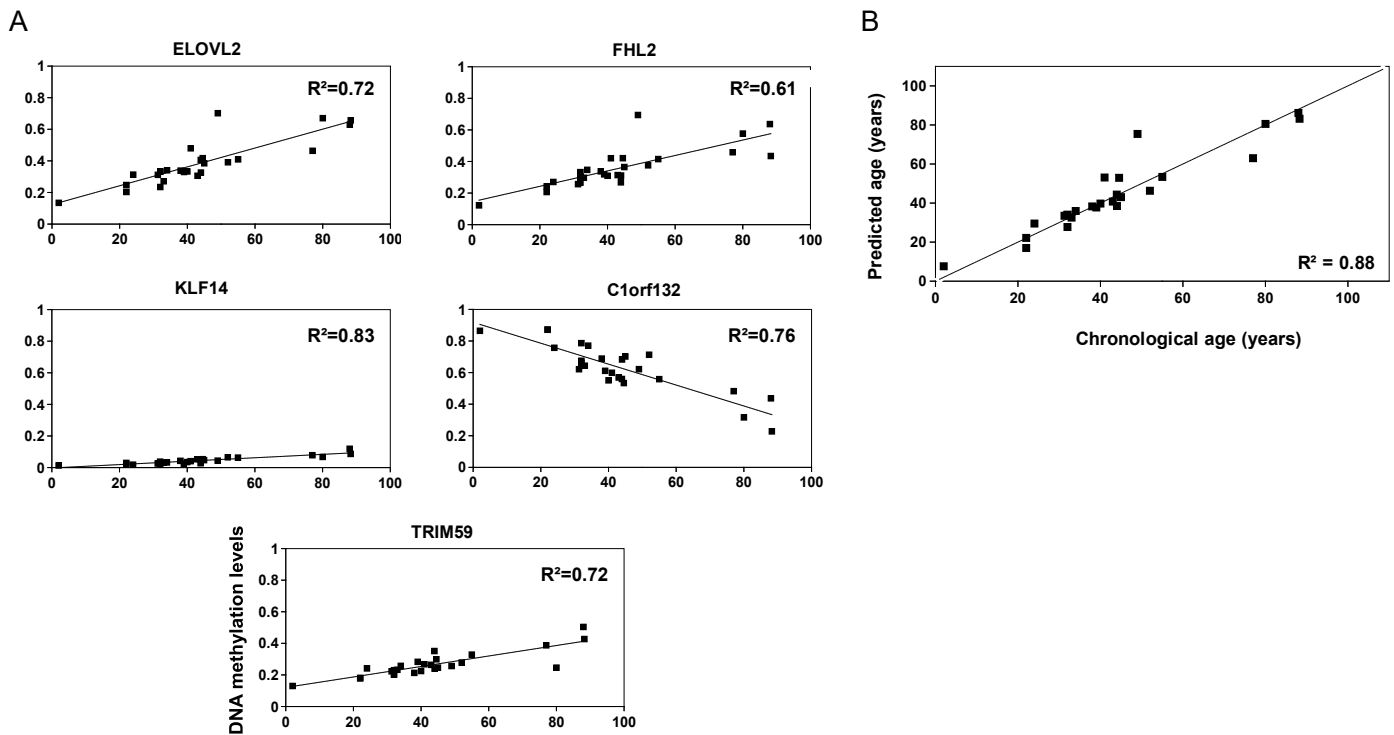


Figure 4. A. Scatter plots representing the correlation between the chronological age and DNA methylation levels at each of the 5 CpG sites analyzed in the *ELOVL2*, *FHL2*, *KLF14*, *C1orf132* and *TRIM59* genes, for a validation set composed of blood samples from individuals aged from 2 to 88 years (n=26). The lines defining the trend curves and the coefficient of determination (R^2) values are indicated in each graph. B. Scatter plot representing the correlation between chronological and predicted biological ages inferred from the age prediction model (APM) for the validation set of blood samples from French individuals (n=26). The line indicates the expected theoretical age for each chronological age.

scope of criminal investigations that could benefit from DNA methylation based age-prediction analyses in the future in France.

Similarly to published studies [20,24,29-31,40], we observed consistent modifications of DNA methylation levels of 5 previously described CpG sites in blood samples depending on the chronological ages of donors. Indeed, *ELOVL2*, *FHL2*, *KLF14* and *TRIM59* CpG sites tend to become more methylated as age increases while the methylation levels of the *C1orf132* site were inversely correlated to age [20,24,29,30,40]. In addition, as previously described in blood samples obtained from different Asian or European populations, predicted biological age based on DNA methylation levels and chronological age tend to display higher correlations for the youngest individuals and increased error rates for individuals older than 45 [20,30,35].

When applied to the validation sample set, the APM displayed significant yet lower correlations ($R^2 = 0.88$ vs. 0.97) between the biological and predicted ages, increased MAE (4.49 vs. 3.45 years) and RMSE (7.17 vs.

5.26) values and reduced 5-years prediction accuracy compared to the training set (62% vs. 75%) (Table 2). This observation is in line with the conclusions related by numerous studies using different APM and sample sets [20,24,29-31,40]. Yet, it likely also results from a more restricted number of samples in the validation set (n= 26) compared to the training set (n= 136). Increasing the number of blood samples analyzed to validate APM should therefore be considered in future studies.

A current limitation to age prediction from DNA methylation analysis remains the representativeness of the training populations for mathematical models, including the age and gender distributions, the total number of samples analyzed, as well as anonymized information relative to the health status of each person implicated in these studies. With these parameters taken into consideration, we established a training cohort that included 136 participants and covering a large age-distribution from new born to 101 years old, with a rather similar representativeness of the different age classes. In comparison,

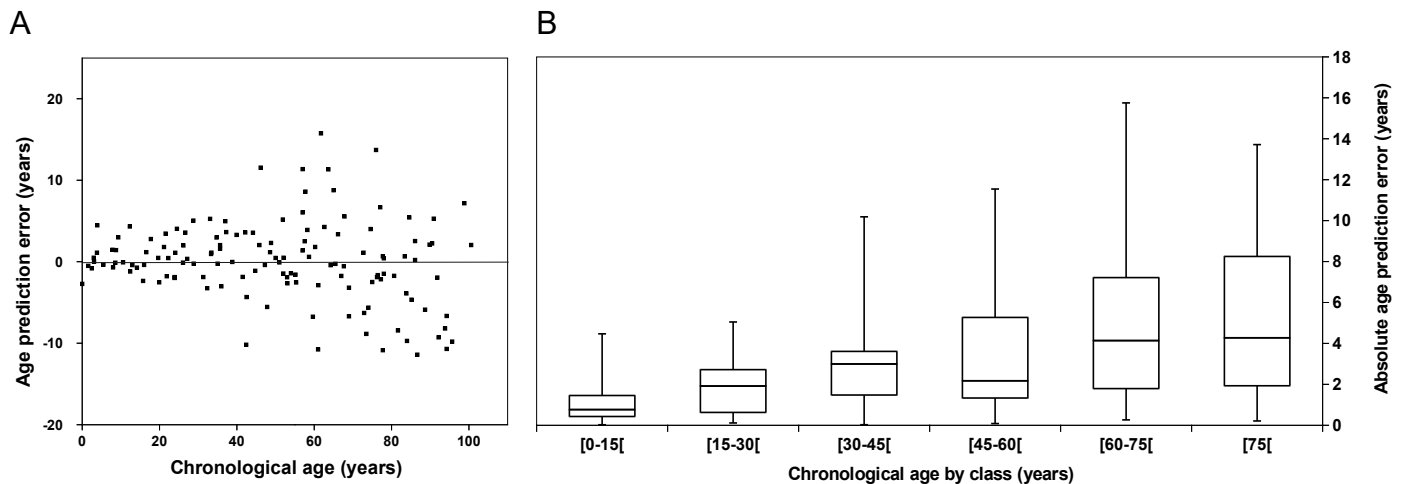


Figure 5. A. Biological age prediction errors versus chronological ages represented by individuals for the training set samples (n=136). B. Boxplots represent the distribution of absolute errors in each 15-year age-class. The edges of the box represent the first and the third quartiles respectively, the line within each box represents the median, and the whisker extends represent maximum and minimum values.

except for the Dias HC, et al. report (59 individuals aged 1-94) [30], previous studies focused on more limited age distributions, including 18-65 years [29], 18-74 years [24], 20-83 [31] or 2-75 years [20]. While our study covers a broad age distribution with equal gender representation, one limitation remains the total number of individuals analyzed to train and validate the age prediction model (APM). Increased numbers of DNA methylation measurements should be considered in future studies to support more robust biological age prediction models, both by increasing the number of participants and by increasing the number of CpG sites analyzed [2,22]. Additional studies are also needed to establish whether a single panel of age-related CpG sites should be considered for these models, or whether two or more panels should be used to predict the biological age from different age classes. Different CpG loci in the vicinity of genes of interest can undergo multiple epigenetic modifications: e.g. 7, 10 and 4 distinct cytosines have been shown to be methylated in the vicinity of *ELOVL2*, *FHL2* and *KLF14* genes, respectively. Taking this complexity and heterogeneity into consideration should also help building more reliable age-predictive models [2,22,28,41]. In addition, different types of models such as quantile regression or constitutional neural networks should be considered in the future to analyze the relationship between DNA methylation and age, as a non-linear relationship [35,40-42].

Conclusion

In conclusion, we established a methodology to predict the biological age of individuals from the French population based on a DNA methylation analysis method compatible with forensic laboratory routines (MAE= 3.64 years with a 5-years prediction rate of 73%). Although criminal investigations by the French police services could already benefit from this biological age prediction model, we anticipate that this reliable methodology should be further improved in a near future prior to adapting it to the French police forensic laboratories. Indeed, defining additional phenotypic features describing unidentified individuals, such as age prediction, remains an interesting perspective and should rely on additional valuable age-associated CpG sites, on more robust statistical models coping for the intrinsic variability of DNA methylation measurements within a complex biological population, or on age-category specific APM to provide the most accurate predictions to support strong investigation leads and help reveal the truth in a near future.

Acknowledgement

The authors thank Emmanuelle Sciacca, Magali Faivre, Joanna Fombonne and Emilie Lessoud for their valuable comments on this project,

and Isabelle Vignon from the Novelab Ingels Vignon laboratory (Société NOVELAB S.E.L.A.S., 69400 Villefranche sur Saône, France) for coordinating the collection of blood samples from donors.

This research was funded by the Service National de Police Scientifique of the French National Police.

Disclosure Statement

The authors declare no conflicts of interest.

References

- Rhodes, Henry Taylor Fowkes. "Alphonse Bertillon: Father of scientific detection." (1956).
- Heidegger, A., Catarina Xavier, H. Niederstätter and M. De la Puente, et al. "Development and optimization of the VISAGE basic prototype tool for forensic age estimation." *Forensic Sci Int Genet* 48 (2020): 102322.
- Xavier, Catarina, Maria de la Puente, Maja Sidstedt and Klara Junker, et al. "Evaluation of the VISAGE basic tool for appearance and ancestry inference using ForenSeq® chemistry on the MiSeq FGx® system." *Forensic Sci Int Genet* 58 (2022): 102675.
- Breslin, Krystal, Bailey Wills, Arwin Ralf and Marina Ventayol Garcia, et al. "HlrisPlex-S system for eye, hair, and skin color prediction from DNA: Massively parallel sequencing solutions for two common forensically used platforms." *Forensic Sci Int Genet* 43 (2019): 102152.
- Chaitanya, Lakshmi, Krystal Breslin, Sofia Zuriaga and Laura Wirken, et al. "The HlrisPlex-S system for eye, hair and skin colour prediction from DNA: Introduction and forensic developmental validation." *Forensic Sci Int Genet* 35 (2018): 123-135.
- Ruiz-Ramírez, Jorge, M. de la Puente, Catarina Xavier and Adrián Ambroa-Conde, et al. "Development and evaluations of the ancestry informative markers of the VISAGE Enhanced Tool for Appearance and Ancestry." *Forensic Sci Int Genet* 64 (2023): 102853.
- Chen, Yan, Pirro Hysi, Carlo Maj and Stefanie Heilmann-Heimbach, et al. "Genetic prediction of male pattern baldness based on large independent datasets." *Eur J Hum Genet* 31 (2023): 321-328.
- Elkins, Kelly M., Alexis T. Garloff and Cynthia B. Zeller. "Additional predictions for forensic DNA phenotyping of externally visible characteristics using the ForenSeq and Imagen kits." *JFS* 68 (2023): 608-613.
- Xavier, C., M. de la Puente, A. Mosquera-Miguel and A. Freire-Aradas, et al. "Development and inter-laboratory evaluation of the VISAGE enhanced tool for appearance and ancestry inference from DNA." *Forensic Sci Int Genet* 61 (2022): 102779.

10. Xavier, Catarina, Maria de la Puente, Ana Mosquera-Miguel and Ana Freire-Aradas, et al. "Development and validation of the VISAGE AmpliSeq basic tool to predict appearance and ancestry from DNA." *Forensic Sci Int Genet* 48 (2020): 102336.
11. Garagnani, Paolo, Maria G. Bacalini, Chiara Pirazzini and Davide Gori, et al. "Methylation of ELOVL 2 gene as a new epigenetic marker of age." *Aging cell* 11 (2012): 1132-1134.
12. Horvath, Steve. "DNA Methylation Data Involving Healthy (Non-Cancer) Tissue." (2013).
13. Boks, Marco P., Eske M. Derks, Daniel J. Weisenberger and Erik Strengman, et al. "The relationship of DNA methylation with age, gender and genotype in twins and healthy controls." *PLoS One* 4 (2009): e6767.
14. Bocklandt, Sven, Wen Lin, Mary E. Sehl and Francisco J. Sánchez, et al. "Epigenetic predictor of age." *PLoS One* 6 (2011): e14821.
15. Rakyan, Vardhman K., Thomas A. Down, Siarhei Maslau and Toby Andrew, et al. "Human aging-associated DNA hypermethylation occurs preferentially at bivalent chromatin domains." *Genome Res* 20 (2010): 434-439.
16. Varshavsky, Miri, Gil Harari, Benjamin Glaser and Yuval Dor, et al. "Accurate age prediction from blood using a small set of DNA methylation sites and a cohort-based machine learning algorithm." *Cell Rep Methods* 3 (2023).
17. Jung, Marc and Gerd P. Pfeifer. "Aging and DNA methylation." *BMC Biol* 13 (2015): 1-8.
18. Yousefi, Paul D., Matthew Suderman, Ryan Langdon and Oliver Whitehurst, et al. "DNA Methylation-based Predictors of Health: Applications and statistical considerations." *Nat Rev Genet* 23 (2022): 369-383.
19. Lee, Hwan Young, Sang-Eun Jung, Yu Na Oh and Ajin Choi, et al. "Epigenetic age signatures in the forensically relevant body fluid of semen: A preliminary study." *Forensic Sci Int Genet* 19 (2015): 28-34.
20. Zbieć-Piekarska, Renata, Magdalena Spólnicka, Tomasz Kupiec and Agnieszka Parys-Proszek, et al. "Development of a forensically useful age prediction method based on DNA methylation analysis." *Forensic Sci Int Genet* 17 (2015): 173-179.
21. Freire-Aradas, A., C. Phillips, A. Mosquera-Miguel and L. Girón-Santamaría, et al. "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 (2016): 65-74.
22. Woźniak, Anna, Antonia Heidegger, Danuta Piniewska-Róg and Ewelina Pośpiech, et al. "Development of the VISAGE enhanced tool and statistical models for epigenetic age estimation in blood, buccal cells and bones." *Aging* 13 (2021): 6459.
23. Hong, Sae Rom, Sang-Eun Jung, Eun Hee Lee and Kyoung-Jin Shin, et al. "DNA methylation-based age prediction from saliva: High age predictability by combination of 7 CpG markers." *Forensic Sci Int Genet* 29 (2017): 118-125.
24. Jung, Sang-Eun, Seung Min Lim, Sae Rom Hong and Eun Hee Lee, et al. "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 (2019): 1-8.
25. Bekaert, Bram, Aubeline Kamalandua, Sara C. Zapico and Wim Van de Voorde, et al. "Improved age determination of blood and teeth samples using a selected set of DNA methylation markers." *Epigenetics* 10 (2015): 922-930.
26. Lee, Hwan Young, Sae Rom Hong, Ji Eun Lee and In Kwan Hwang, et al. "Epigenetic age signatures in bones." *Forensic Sci Int Genet* 46 (2020): 102261.
27. Poussard, Alexandre, Jean-Yves Curci, Christian Siatka and Francis Hermitte, et al. "Evaluation of DNA Methylation-based age-prediction models from saliva and buccal swab samples using pyrosequencing data." *Forensic Sci* 3 (2023): 192-204.
28. Daunay, Antoine, Laura G. Baudrin, Jean-François Deleuze and Alexandre How-Kit. "Evaluation of six blood-based age prediction models using DNA methylation analysis by pyrosequencing." *Sci Rep* 9 (2019): 8862.
29. Onofri, Martina, Arianna Delicati, Beatrice Marcante and Luigi Carlini, et al. "Forensic age estimation through a DNA methylation-based age prediction model in the Italian population: A pilot study." *Int J Mol Sci* 24 (2023): 5381.
30. Dias, Helena Correia, Cristina Cordeiro, Janet Pereira and Catarina Pinto, et al. "DNA methylation age estimation in blood samples of living and deceased individuals using a multiplex SNaPshot assay." *Forensic Sci Int* 311 (2020): 110267.
31. Filoglu, Gonul, Sumeyye Zula Simsek, Gokhan Ersoy and Kadriye Can, et al. "Epigenetic-based age prediction in blood samples: Model development." *JFS* (2024).
32. Hannum, Gregory, Justin Guinney, Ling Zhao and L. I. Zhang, et al. "Genome-wide methylation profiles reveal quantitative views of human aging rates." *Mol Cell* 49 (2013): 359-367.
33. Jung, Sang-Eun, Kyoung-Jin Shin and Hwan Young Lee. "DNA methylation-based age prediction from various tissues and body fluids." *BMB Rep* 50 (2017): 546.
34. Lee, Ji Eun, Jeong Min Lee, Jana Naue and Jan Fleckhaus, et al. "A collaborative exercise on DNA methylation-based age prediction and body fluid typing." *Forensic Sci Int Genet* 57 (2022): 102656.
35. Aliferi, Anastasia, Sudha Sundaram, David Ballard and Ana Freire-Aradas, et al. "Combining current knowledge on DNA Methylation-based age estimation towards the development of a superior forensic DNA intelligence tool." *Forensic Sci Int Genet* 57 (2022): 102637.
36. Becker, J., P. Böhme, A. Reckert and S. B. Eickhoff et al. "Evidence for differences in DNA methylation between Germans and Japanese." *Int J Legal Med* 136 (2022): 405-413.
37. Fleckhaus, J., P. Bugert, N. A. M. Al-Rashedi and M. A. Rothschild. "Investigation of the impact of biogeographic ancestry on DNA methylation based age predictions comparing a Middle East and a Central European population." *Forensic Sci Int Genet* 67 (2023): 102923.
38. Fleckhaus, Jan, Ana Freire-Aradas, Markus A. Rothschild and Peter M. Schneider. "Impact of genetic ancestry on chronological age prediction using DNA methylation analysis." *Forensic Sci Int Genet Suppl Ser* 6 (2017): e399-e400.
39. Thong, Zhonghui, Jolena Ying Ying Tan, Eileen Shuzhen Loo and Yu Wei Phua, et al. "Artificial neural network, predictor variables and sensitivity threshold for DNA methylation-based age prediction using blood samples." *Sci Rep* 11 (2021): 1744.
40. Freire-Aradas, A., Lorena Girón-Santamaría, Ana Mosquera-Miguel and Adrián Ambroa-Conde, et al. "A common epigenetic clock from childhood to old age." *Forensic Sci Int Genet* 60 (2022): 102743.
41. Yamagishi, Takayuki, Wataru Sakurai, Ken Watanabe and Kochi Toyomane, et al. "Development and comparison of forensic interval age prediction models by statistical and machine learning methods based on the methylation rates of ELOVL2 in blood DNA." *Forensic Sci Int Genet* 69 (2024): 103004.
42. Freire-Aradas, A., Ewelina Pośpiech, A. Aliferi and L. Girón-Santamaría, et al. "A comparison of forensic age prediction models using data from four DNA methylation technologies." *Front Genet* 11 (2020): 557076.

How to cite this article: Camille Ropert, Simon Buré, Adèle Sourisce and Caroline Gallois, et al. "DNA Methylation-based Age Prediction from Blood Samples of French Child and Adult Individuals." *J Forensic Res* 15 (2024): 606.