Forensic Science International: Genetics 61 (2022) 102770

Contents lists available at ScienceDirect

Forensic Science International: Genetics

journal homepage: www.elsevier.com/locate/fsigen

Research paper

Epigenetic age estimation in saliva and in buccal cells

A. Ambroa-Conde [a], L. Girón-Santamaría [a], A. Mosquera-Miguel [a], C. Phillips [a], M.A. Casares de Cal [b], A. Gómez-Tato [b], J. Álvarez-Dios [c], M. de la Puente [a], J. Ruiz-Ramírez [a], M.V. Lareu [a], A. Freire-Aradas [a,*]

[a] Forensic Genetics Unit, Institute of Forensic Sciences, University of Santiago de Compostela, Spain
[b] CITMAga (Center for Mathematical Research and Technology of Galicia), University of Santiago de Compostela, Spain
[c] Faculty of Mathematics, University of Santiago de Compostela, Spain

ARTICLE INFO

Keywords:
DNA methylation
Forensic age estimation
Logistic regression
Quantile regression
SNaPshot
Saliva
Buccal swab
Buccal cells

ABSTRACT

Age estimation based on epigenetic markers is a DNA intelligence tool with the potential to provide relevant information for criminal investigations, as well as to improve the inference of age-dependent physical characteristics such as male pattern baldness or hair color. Age prediction models have been developed based on different tissues, including saliva and buccal cells, which show different methylation patterns as they are composed of different cell populations. On many occasions in a criminal investigation, the origin of a sample or the proportion of tissues is not known with certainty, for example the provenance of cigarette butts, so use of combined models can provide lower prediction errors.

In the present study, two tissue-specific and seven age-correlated CpG sites were selected from publicly available data from the Illumina HumanMethylation 450 BeadChip and bibliographic searches, to help build a tissue-dependent, and an age-prediction model, respectively. For the development of both models, a total of 184 samples (N = 91 saliva and N = 93 buccal cells) ranging from 21 to 86 years old were used. Validation of the models was performed using either k-fold cross-validation and an additional set of 184 samples (N = 93 saliva and N = 91 buccal cells, 21–86 years old).

The tissue prediction model was developed using two CpG sites (*HUNK* and *RUNX1*) based on logistic regression that produced a correct classification rate for saliva and buccal swab samples of 88.59 % for the training set, and 83.69 % for the testing set. Despite these high success rates, a combined age prediction model was developed covering both saliva and buccal cells, using seven CpG sites (cg10501210, *LHFPL4*, *ELOVL2*, *PDE4C*, *HOXC4*, *OTUD7A* and *EDARADD*) based on multivariate quantile regression giving a median absolute error (MAE): ± 3.54 years and a correct classification rate (%CP±PI) of 76.08 % for the training set, and an MAE of ± 3.66 years and a %CP±PI of 71.19 % for the testing set. The addition of tissue-of origin as a co-variate to the model was assessed, but no improvement was detected in age predictions. Finally, considering the limitations usually faced by forensic DNA analyses, the robustness of the model and the minimum recommended amount of input DNA for bisulfite conversion were evaluated, considering up to 10 ng of genomic DNA for reproducible results. The final multivariate quantile regression age predictor based on the models we developed has been placed in the open-access *Snipper* forensic classification website.

1. Introduction

Age estimation can provide key information in criminal, legal and anthropological investigations [1]. In cases where there are no suspects and the DNA profiles recovered from forensic biological samples do not match with any profile stored in national DNA databases, age prediction can play an important role guiding police investigations, which can reduce the number of potential suspects [2]. Age estimation may also improve the prediction of phenotypic characteristics related to aging, e.g. hair colour [3] or male pattern baldness [4]. Additionally, if the prediction models develop enough accuracy, legal disputes could potentially be supported by age estimation [5]. In all these cases, chronological age rather than biological age needs to be inferred [6].

DNA methylation has become the gold standard biomarker for

* Corresponding author.
E-mail address: ana.freire@usc.es (A. Freire-Aradas).

https://doi.org/10.1016/j.fsigen.2022.102770
Received 23 June 2022; Received in revised form 22 August 2022; Accepted 24 August 2022
Available online 27 August 2022

human age estimation. This epigenetic signature consists of the addition of a methyl group (-CH_3) to the 5′ carbon of cytosines positioned next to guanines (CpG nucleotides) [7]. Age correlation with DNA methylation has been largely confirmed by a broad range of epigenetic studies [8–16]. Based on the DNA methylation values of age correlated CpG sites, multiple forensic age prediction models have been developed to date, reviewed in [1]. Since DNA methylation is tissue-specific [17], most of these epigenetic clocks have been based on specific forensic tissues, including blood [18–21], buccal swabs [22–24], saliva [23,25] and semen [26,27]. More recently, skeletal remains, e.g., bones and teeth have been studied [28,29].

Whole blood is not uniformly composed of identical cell types, but consists of distinct cell populations in varying proportions. As methylation profiles of peripheral blood mononuclear cells and granulocytes have been identified [30], cell heterogeneity could act as a confounder. However, studies have observed that DNA methylation for age correlated CpG sites does not vary significantly across sorted blood cells from healthy subjects [13], and subsequently, most forensic age prediction models were based on whole blood treated as a homogeneous tissue.

Another tissue source that lacks cellular homogeneity is the oral cavity, where saliva and buccal swabs have different varied proportions of leucocytes and epithelial cells [31]. This difference in cell content could potentially create differences in DNA methylation for specific CpG sites, and this phenomenon was previously observed for *ELOVL2* and *FHL2* [23], indicating that both sample types cannot be considered a single biological source beforehand.

Nevertheless, considering that deconvolution to assign the specific biological source - saliva or buccal swabs - to forensic oral cavity specimens is difficult to achieve, e.g., cigarette butts, the development of a single age prediction model covering both tissues represents a practical approach.

A similar approach has already been proposed by Horvath et al. [32], developing the “skin & blood clock”, an epigenetic clock based on 391 CpGs that covers samples originating from blood, skin, saliva, buccal cells, as well as from four additional somatic tissues. The age prediction model reported by Jung et al., is more focused on forensic specimens, and is based on 5 CpG sites applicable to either blood, saliva or buccal cells [23].

In the present study, we focused on specimens from the oral cavity aiming to develop a tissue prediction model that can differentiate saliva from buccal cells, as well as an age prediction model covering both tissues, since most forensic samples related to the oral cavity will comprise a mixture of saliva and buccal cells. Additionally, to include the tissue-of-origin as a co-variable do not improve age predictions. Selection of candidate tissue-specific and age correlated CpG sites was based on the assessment of public data from Illumina HumanMethylation 450 K. Then, 184 volunteers (21–86 years old) were analyzed using SNaPshot™, after collection of either saliva and buccal swabs from the same individual (N = 368). A proportion of the analyzed samples were used to develop the training set (N = 184), while an additional part was used as a testing set for model validation purposes (N = 184). As a result, a tissue prediction model (saliva vs buccal cells) using logistic regression and based on 2 CpG sites was developed. In parallel, an age prediction model covering these tissues together and based on multivariate quantile regression analysis was developed for 7 CpG sites showing the highest correlation with age. Since SNaPshot™ needs a preliminary step of bisulfite conversion that degrades the DNA, requiring high levels of input DNA, we made an evaluation of serial dilutions with this detection system to determine the limits of the assay.

2. Material and methods

2.1. Samples, DNA extraction and quantification

A total of 368 samples, 184 total saliva and 184 buccal cells, were collected from 184 healthy Spanish volunteers from 21 to 86 years old. Based on this set of samples, for the saliva-specific and buccal swab-specific age prediction models, the whole set of 184 saliva and 184 buccal swabs, respectively, were directly used as training sets. For the tissue-combined age prediction model, a random selection was made to generate training and test sets balanced in terms of sample size, distribution of ages and represented tissues. Each group had 184 individuals with the full age range 21–86 years. The training set consisted of 91 saliva and 93 buccal cell samples, while the testing group had 93 saliva and 91 buccal cell samples.

All samples were taken with written informed consent obtained from the donors. Ethical approval was obtained from the ethics committee of investigation in Galicia, Spain (CAEI: 2013/543). Buccal swabs were air-dried and stored at room temperature and total saliva was collected with 15 mL falcon tubes and frozen at − 20 °C until DNA extraction. Genomic DNA was extracted from the whole swab and from 500 μL of total saliva with phenol/chloroform extraction [33]. All DNA samples were quantified by Qubit® dsDNA High Sensitivity (HS) or dsDNA Broad Range (BR) Assay kits (Thermo Fisher) following manufacturer’s guidelines.

2.2. CpG site selection

Selection of candidate CpG sites was based on both bibliographic searches as well as statistical assessment of NCBI GEO methylation studies using public data from the Illumina HumanMethylation450KBeadChip. Tissue-specific CpG site selection was based on the statistical assessment of the methylation β-values from GSE48472 [34] (blood, saliva and buccal cells). To check for absence of correlation with age for the selected tissue-specific markers; GSE87571 [14] GSE92767 [25] and GSE50586 [35] were used. Furthermore, the bibliographic review was focused on publications from 2011 to 2019, and searched for markers presenting a high correlation with age in different tissues: blood [18,20,28,36], saliva [9,37], and buccal cells [22,38]. Additionally, methylation β-values from GSE92767 [25] were statistically assessed to seek to identify additional age-correlated CpG sites.

2.3. Primer design

The flanking regions of the selected CpGs were screened using the UCSC genome browser (https://genome.ucsc.edu/) for the current human genome assembly (GRCh38/hg38), covering 150 bp upstream and downstream of the target CpG. The PCR primer and Single Base Extension (SBE) primer designs were made using BatchPrimer 3 v1.0 [39] applying the following parameters for PCR primers: optimal melting temperature 58 °C, optimal primer length 20 bp and optimal amplicon length 90 bp; and for the SBE primer design: optimal melting temperature 50 °C and optimal probe length 20 bp. Poly-CT tails were added to the SBE primers for size separation.

2.4. Bisulfite conversion, PCR conditions and purification of PCR products

Bisulfite conversion of 100 ng of extracted genomic DNA was carried out with the MethylEdge™ Bisulfite Conversion System (Promega) following manufacturer’s guidelines, obtaining an elution volume of 20 μL. A PCR multiplex amplification in 10.7 μL reaction volume adding 1.5 μL of converted DNA was carried out using 0.3 μL of 250 U AmpliTaq Gold™ DNA Polymerase, 1.5 μL of 10X Buffer II, 3.9 μL of 25 mM $MgCl_2$ (all from Applied Biosystems, AB), 1.5 μL of 32 ng/μL bovine serum albumin, 1 μL of 10 mM GeneAmp® dNTP Mix with dTTP (AB) and 1 μL of primer mix (0.083–5 μM of each primer, Metabion International). PCR cycling used a GeneAmp® PCR system 2720 (AB) with cycling conditions: 95ºC for 11 min; 34 cycles of 94ºC for 20 s, 56ºC for 60 s and 72ºC for 30 s, and a final extension of 72ºC for 7 min

After checking amplification yields in 1 % agarose gels, a purification of 2.5 μL of PCR product was performed adding 1 μL of ExoSAP-IT™ PCR

Product Cleanup Reagent (AB) at 37 °C for 45 min and 80 °C for 15 min

2.5. Single base extension and capillary electrophoresis

Multiplex SBE reactions were performed in a total volume of 6 μL using 2 μL of purified PCR product, 2.5 μL of SNaPshot™ kit (AB) and 1.5 μL of SBE primers (0.51–6 μM of each primer, Metabion International) with cycling conditions: 30 cycles of 96 °C for 10 s, 55 ° for 5 s, and 60 °C for 30 s

After the SNaPshot reaction, extension products were purified by adding 1 μL of Shrimp Alkaline Phosphatase Recombinant (AB) to the total SNaPshot reaction and incubating at 37 °C for 80 min with inactivation at 85 °C for 15 min

Capillary electrophoresis was performed with an ABI3130xl Genetic Analyzer (AB) using 0.1 μL of GeneScan™ 120 LIZ™ dye Size Standard (Thermo Fisher) and 10 μL of HiDi™ Formamide (AB) per sample, adding 9.5 μL of load mix and 1.5 μL of purified SNaPshot product. Results were analyzed with GeneMapperID v3.2 (AB) and the DNA methylation level at each CpG was calculated by dividing the height of the methylated peak by the sum of the heights of the methylated and unmethylated peaks. The latter values were multiplied by a correction factor of 2, when working with reverse primers and 1.6 for forward primers, to overcome differences at fluorochrome signal intensities.

2.6. Statistical analyses

All samples were run in duplicate. The average of the DNA methylation levels in both replicates was used for the statistical analyses. Correlations between age and DNA methylation levels were evaluated using the Spearman Correlation test (r_s). To analyze the reproducibility of the dilutions and the inter-individual variability, the standard deviation (SD) was used (threshold SD > 0.1). Normality was assessed using the Shapiro-Wilk test applied to the residuals of the independent linear regression models tested for each CpG (p-value < 0.05). Logistic regression was used to develop the tissue prediction model using the *pROC* R package [40]. A multivariate quantile regression model was used to build the age prediction model using the *quantreg* R package [41]. Cross-validation of the prediction models was performed with a k-fold cross-validation (k = 10) using the *cvTools* R package [42]. The corresponding predictive accuracy was measured with the following performance metrics: sensitivity, specificity, area under the curve (AUC) and percentage of correct classifications for tissue prediction; and the median absolute error (MAE), the mean absolute error (MAE_{mean}), the root-mean-square error (RMSE) and percentage of correct classifications within the prediction intervals (%CP±PI) for age prediction. The representation of predicted versus chronological age was made using the *ggplot2* R package [43]. All statistical analyses were carried out using R software v.4.0.3 [44] with scripts developed in-house. The sensitivity analysis was carried out using input DNA quantities for bisulfite conversion of 100 ng, 75 ng, 50 ng, 25 ng, 10 ng and 1 ng.

3. Results

3.1. Selection of candidate CpGs

The selection of candidate CpGs was divided into tissue-specific CpGs and age-correlated CpGs.

For selection of tissue-specific CpGs, the GSE48472 dataset was assessed [34]. From this dataset, samples from saliva (N = 5), buccal cells (N = 5) and blood (N = 5) were selected and differences in the corresponding DNA methylation values calculated. A total of 17 CpG sites with the highest differences in DNA methylations levels were found (Table 1): 5 CpGs presenting the highest differences between blood and buccal cells (>|0.72|); 6 CpGs between blood and saliva (>|0.45|) and 6 CpGs between saliva and buccal cells (≥|0.5|).

Once the markers had been selected, absence of correlation with age was evaluated using the following datasets: GSE92767 (saliva) [25], GSE50586 (buccal cells) [35] and GSE87571 (blood) [14]. From the 17 selected tissue-specific CpGs, three displayed correlations with age (r_s >|0.5|): cg01680010 (r_s =−0.607) in buccal cells and cg13408086 (r_s =0.609) and cg08466792 (r_s =0.575) in blood, so were discarded. Based on these results, one CpG site per tissue combination was selected. This selection was initially based on the highest difference displayed by the DNA methylation values observed in pairs of tissues. However, several failures in PCR primer design led to a final selection of cg04915566 (*RUNX1*) for blood-buccal cells, cg16606773 (*RIN2*) for blood-saliva and cg03044684 (*HUNK*) for saliva-buccal cells.

Selection of age-correlated CpGs was based on the assessment of DNA methylation values from GSE92767 (saliva samples, N = 54, 18–73 years old) [25]. In order to select the method to be used for marker selection, normality was evaluated for GSE92767 data, obtaining that 15 % of the residuals of the models (independent linear regression models for each CpG) presented a lack of normality (p-value < 0.05), therefore, the Spearman test was used. For this analysis, CpG sites presenting a Spearman correlation coefficient equal to or greater than |0.8| were selected, providing 49 CpG sites correlated with age (Supplementary Table S1). From this preliminary set of sites, those CpGs with a minimum difference of 0.3 between the highest and lowest methylation

Table 1
Summary of the 17 selected tissue-specific CpG sites based on the statistical assessment of GSE48472. CpG sites correlated with age (r_s >|0.5|) are marked in bold.

Tissue's comparison	Gene	CpG_ID	GRCh38 chromosome position	Differences between pairs of tissues	Correlation with age (r_s blood)	Correlation with age (r_s buccal cells)	Correlation with age (r_s saliva)
Blood-Buccal cells	*RUNX1*	cg04915566	chr21:35049175	0.723	0.023	0.006	-0.309
	MAML2	cg08141395	chr11:96254218	0.748	0.006	0.043	-0.306
	RGS1	cg10861751	chr1:192575586	0.733	-0.024	0.055	-0.266
	EXD3	cg13408086	chr9:137326945	0.724	**0.609**	-0.337	-0.481
	NCKAP1L	cg16509569	chr12:54497850	0.721	-0.019	-0.190	-0.318
Blood-Saliva	*CDC25B*	cg02737268	chr20:3799535	0.483	0.287	-0.079	0.439
	DOT1L	cg04173586	chr19:2167497	0.459	0.001	0.129	0.408
	RIN3	cg15443535	chr14:92687972	0.476	-0.152	0.411	0.399
	none	cg16149628	chr11:1771344	0.471	0.023	0.166	0.270
	RIN2	cg16606773	chr20:19975162	0.459	-0.179	-0.153	-0.049
	WDFY1	cg23363263	chr2:223887272	0.452	-0.102	-0.043	0.218
Saliva-Buccal cells	none	cg01680010	chr7:97017805	0.500	0.148	**-0.607**	-0.079
	none	cg02939659	chr14:101587733	0.500	-0.048	0.472	0.389
	HUNK	cg03044684	chr21:31875719	0.503	-0.065	0.055	0.247
	PAX9	cg07459252	chr14:36661007	0.502	0.334	-0.104	-0.106
	none	cg08466792	chr5:3603113	0.512	**0.575**	-0.362	-0.378
	SIM2	cg25446076	chr21:36710849	0.523	0.379	-0.349	-0.332

values were selected, to give ten candidate CpGs (Table 2) for age prediction in saliva.

As the statistical analysis for selection of age correlated CpG sites was based on saliva samples, but the study also covered buccal cells; a bibliographic search to find genes that show correlation with age in additional somatic tissues was carried out. In the reviewed publications, certain markers were repeatedly found to correlate with age in the tissues of interest (saliva, buccal cells and blood): *PDE4C* [18,20,22,28,37], *EDARADD* [9,28,38] and *ASPA* [18,20,22,28,36,37], and therefore these genes were the focus of further evaluation in our study (included in Table 2).

3.2. Development of an optimized multiplex

From the above analyses, 16 markers were selected: 3 tissue-specific CpG sites (*RUNX1*, *RIN2* and *HUNK*) and 13 age-correlated CpG sites (*OTUD7A*, *FHL2*, *TRIM59*, *RHBDL2*, cg10501210, cg10804656, *LHFPL4*, cg13327545, *ELOVL2*, *HOXC4*, *PDE4C*, *EDARADD* and *ASPA*). PCR and SBE primers were successfully designed for the selected tissue-specific markers, and 11 age-correlated CpGs, with *RHBDL2* and cg10804656 discarded from subsequent analyses. A summary of PCR and SBE primer information is outlined in Supplementary Table S2.

First, each marker was analyzed in singleplex to check for individual amplification performance. Once this initial step was accomplished, a multiplex covering all 14 CpGs was optimized. Markers *TRIM59* and cg13327545 were not amplified in multiplex due to non-specific hybridizations leading to the final optimized multiplex of *RUNX1*, *RIN2*, *HUNK* tissue-specific markers plus *OTUD7A*, *FHL2*, cg10501210, *LHFPL4*, *ELOVL2*, *HOXC4*, *PDE4C*, *EDARADD*, *ASPA* age-correlated CpG sites. An example SNaPshot electropherogram of the optimized multiplex is shown in Fig. 1.

To check tissue-specificity and age correlation of the optimized marker set, a preliminary analysis using both saliva and buccal cells from two individuals of age extremes (23 and 86 years old) was completed (Supplementary Table S3). Tissue specificity was not detected for *RIN2*, since the same absence of methylation pattern was observed for both saliva and buccal cells. To check that the absence of methylation was not a technical problem, and since *RIN2* was selected for detecting differences between blood and saliva, blood samples from the same individuals were tested and detected DNA methylation levels of 0.39 and 0.4, respectively.

In the case of *RUNX1*, some dispersion was detected in the patterns displayed by age or tissue, preventing objective interpretation with this marker. However, *HUNK* had differences in average DNA methylation levels between both tissues (0.31 and 0.17 for saliva and buccal cells, respectively).

For age-correlation, six of the nine CpG sites (cg10501210, *LHFPL4*, *ELOVL2*, *PDE4C*, *ASPA* and *EDARADD*) gave average DNA methylation difference between extreme ages equal to or higher than 0.19. Differences were displayed by *HOXC4*, *OTUD7A* and *FHL2* at lower levels (0.1, 0.05 and 0.12, respectively).

Table 2
Summary of the ten selected CpG sites correlated with age in saliva, based on the statistical assessment of GSE92767, as well as the 3 selected age correlated CpG sites in somatic tissues based on bibliographic review.

Gene	CpG_ID	GRCh38 chromosome position	Correlation with age (r_s)	Methylation differences at extreme ages
GSE92767 assessment				
OTUD7A	cg04875128	chr15:31483692	0.860	0.312
FHL2	cg06639320	chr2:105399282	0.824	0.322
TRIM59	cg07553761	chr3:160450189	0.803	0.305
RHBDL2	cg10500653	chr1:38941979	0.814	0.334
none	cg10501210	chr1:207823675	-0.864	0.674
none	cg10804656	chr10:22334531	0.828	0.317
LHFPL4	cg11084334	chr3:9552580	0.846	0.345
none	cg13327545	chr10:22334619	0.818	0.302
ELOVL2	cg16867657	chr6:11044644	0.898	0.385
HOXC4	cg18473521	chr12:54054481	0.824	0.389
Bibliographic review				
PDE4C	none	chr19:18233131	na	na
EDARADD	cg09809672	chr1:236394382	na	na
ASPA	cg02228185	chr17:3476273	na	na

3.3. A statistical tissue prediction model

The training set comprising 91 saliva samples and 93 buccal swabs was analyzed with the optimized multiplex to develop a tissue prediction model for saliva and buccal cells. The corresponding dispersion diagrams for *HUNK* and *RUNX1* markers is shown in Fig. 2. Dispersion correlated with the tissue-of-origin is observed, with higher methylation levels for *HUNK* in saliva samples, and for *RUNX1* in buccal cells.

In order to predict tissue of origin, logistic regression was applied exploring three different models: model 1 (*HUNK* plus *RUNX1*), model 2 (*HUNK*) and model 3 (*RUNX1*). The corresponding performance metrics are described in Table 3. Comparable AUC values of 0.95, 0.95 and 0.92 for model 1, 2 and 3, respectively were obtained. Similar percentage of correct classifications was also recorded, with model 1 having the highest value at 88.6 %. However, some differences were found with sensitivity and specificity values, considering buccal cell samples as the control (i.e., a high specificity indicates good classification of buccal cell samples and a high sensitivity good classification of saliva samples). Model 1 gave a higher sensitivity (0.96) compared to specificity (0.82). Therefore, model 1 results show that saliva samples classify better than swab samples. In contrast, model 2 gave a sensitivity of 0.78 and a specificity of 0.96; model 3 gave a sensitivity of 0.81 and a specificity of 0.9. Therefore, single marker models classify saliva samples less efficiently than buccal swab samples.

Considering the highest rate of correct classifications obtained (88.59 %), model 1 was selected for validation with a testing set of 184 samples (N = 93 saliva and N = 91 buccal cells. A correct tissue-of-origin prediction rate of 83.7 % for test set samples was obtained.

3.4. A statistical age prediction model for saliva and buccal swab samples

Tissue-independent as well as tissue-combined models were explored for age prediction. For the saliva-specific and buccal swab-specific age prediction models, 184 saliva and 184 buccal swab samples were used as training sets, respectively. The training set of 184 volunteers (N = 91 saliva and N = 93 buccal swabs) was used to develop the combined age prediction model for saliva and buccal cell samples. Dispersion plots in Fig. 3 indicate the patterns obtained for the cg10501210, *LHFPL4*, *ELOVL2*, *PDE4C*, *HOXC4*, *OTUD7A*, *FHL2*, *ASPA* and *EDARADD* markers adopted. Six markers showed hypermethylation with increased age (*LHFPL4*, *ELOVL2*, *PDE4C*, *HOXC4*, *OTUD7A* and *FHL2*); while cg10501210, *ASPA* and *EDARADD* had decreasing methylation levels with increasing age. If considering both tissues combined (saliva and buccal swabs), the highest correlation with age was found in *PDE4C* (r_s =0.806) and *LHFPL4* (r_s =0.805), followed by *ELOVL2* (r_s =0.659), *OTUD7A* (r_s =0.642), *EDARADD* (r_s =−0.572) and *HOXC4* (r_s =0.569). However, low levels of correlation were detected in cg10501210, *FHL2* and *ASPA* (r_s =−0.313, 0.198 and −0.332, respectively). At the same time, these three markers showed the highest levels of dispersion between saliva and buccal cells, (SD> 0.1). If taking into account both tissues independently, correlations followed a similar trend (Supplementary Fig. S1-S2). Whereas the highest age correlation was displayed by *LHFPL4* and *PDE4C* (r_s =0.815 and 0.832 in saliva and buccal swabs, respectively), the lowest levels of correlation were observed in cg10501210 (r_s =−0.429, −0.422), *FHL2* (r_s =0.392, 0.231) and *ASPA* (r_s =−0.521, −0.44).

Taking into account these observations, multivariate quantile regression was tested on several age prediction models consisted of different combinations of CpG sites: model 1 (9 CpGs: cg10501210,

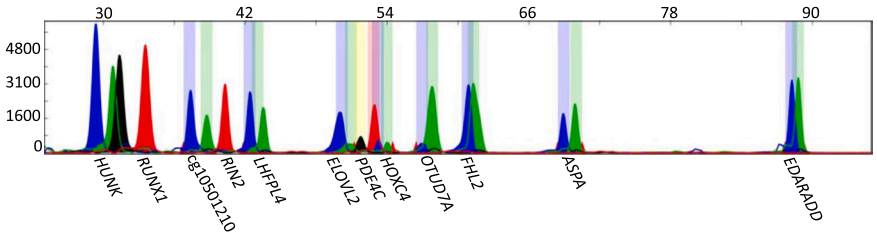

Fig. 1. Example electropherogram of the optimized SNaPshot™ multiplex assay containing 3 tissue-specific and nine age correlated CpG sites using 100 ng of genomic DNA.

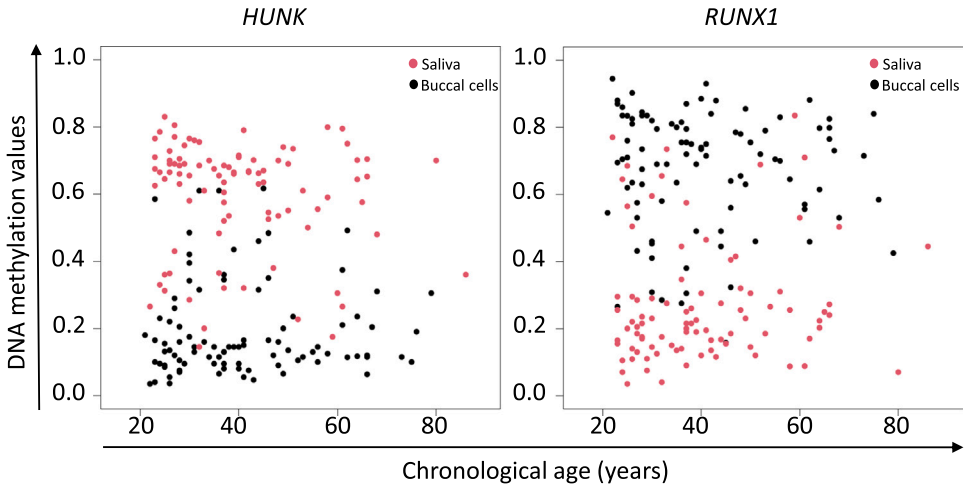

Fig. 2. Dispersion diagrams (DNA methylation values in front of chronological age) for *HUNK* and *RUNX1* (tissue-specific CpG sites) for 184 individuals from 21 to 86 years old (N = 91 saliva and N = 93 buccal swabs).

Table 3
Summary of the predictive performance metrics for the three logistic models tested on the training set (N = 91 saliva and N = 93 buccal swabs, 21–86 years old). AUC: Area under the curve.

Model	CpG_ID	Gen	AUC	Sensitivity	Specificity	Correct classifications
Model 1	cg03044684 & cg04915566	*HUNK & RUNX1*	0.95	0.96	0.82	88.59 %
Model 2	cg03044684	*HUNK*	0.95	0.78	0.96	86.87 %
Model 3	cg04915566	*RUNX1*	0.92	0.81	0.90	85.87 %

LHFPL4, *ELOVL2*, *PDE4C*, *HOXC4*, *OTUD7A*, *FHL2*, *ASPA* and *EDARADD*), model 2 (8 CpGs with *ASPA* excluded), model 3 (8 CpGs with *FHL2* excluded), model 4 (8 CpGs with cg10501210 excluded), model 5 (7 CpGs with cg10501210 and *FHL2* excluded), model 6 (7 CpGs with cg10501210 and *ASPA* excluded), model 7 (7 CpGs with *FHL2* and *ASPA* excluded) and model 8 (6 CpGs with cg10501210, *FHL2* and *ASPA* excluded). To evaluate the accuracy of the models, a k-fold cross-validation was carried out. The "k-fold" divides the total number of individuals into groups of similar sizes, in this case, 10 groups were created, each containing 10 % of the subjects. Each model was tested for each of the clusters, therefore, each time one of the clusters was selected as a test set, it faced the remaining nine that make up the training set. The corresponding performance metrics for the training sets are described in Table 4.

Inter-training set comparisons show that the correct classification rates are similar among them (%CP±PI: 76.66 %, 75.37 % and 76.23 %, for saliva, buccal swab and the combined model, respectively). However, more remarkable differences were found when comparing prediction errors, especially between the buccal swab-specific and the combined model (average MAE: ± 3.89 and ± 4.35, respectively). Nevertheless, the saliva-specific model showed a better prediction error than the combined model (average MAE: ± 3.55). Based on these results, and due to the fact that many forensic specimens will comprise a mixture of saliva and buccal cells with different cell proportions, e.g., cigarette butts, the corresponding age prediction model to be developed was selected to cover both tissues simultaneously (combined model).

Intra-training set comparisons of the combined model showed that the highest error and lowest correct classification rate were obtained with model 8 (MAE: ± 5.23, RMSE: 7.54 and %CP±PI: 74.06 %), which lacks the 3 CpG sites with the lowest levels of correlation with age and highest dispersion between saliva and buccal cells (cg10501210, *FHL2* and *ASPA*). When including these CpG sites (model 1), error decreased (MAE: ± 3.31) but the correct classification rate is only marginally improved (74.38 %). Among all models tested, the best balance between error and correct classification was obtained with model 7, which excludes *FHL2* and *ASPA* (MAE: ± 3.54, RMSE: 6.23 and %CP±PI: 76.08 %). Subsequently, we selected the age prediction model for saliva and buccal cells based on CpGs cg10501210, *LHFPL4*, *ELOVL2*, *PDE4C*, *HOXC4*, *OTUD7A* and *EDARADD*. Predicted versus chronological age is plotted for the final 7-CpG age prediction model in Fig. 4. The quantiles 0.5, 0.1 and 0.9 are represented by a black line and dashed dark red lines, respectively and the gray line represents perfect correlation. The Fig. 4 plot shows that the 0.5 quantile line is more separated in older ages, possibly due to the low number of samples available for this age range. The non-parallel prediction intervals also show the reduced precision in the highest age ranges.

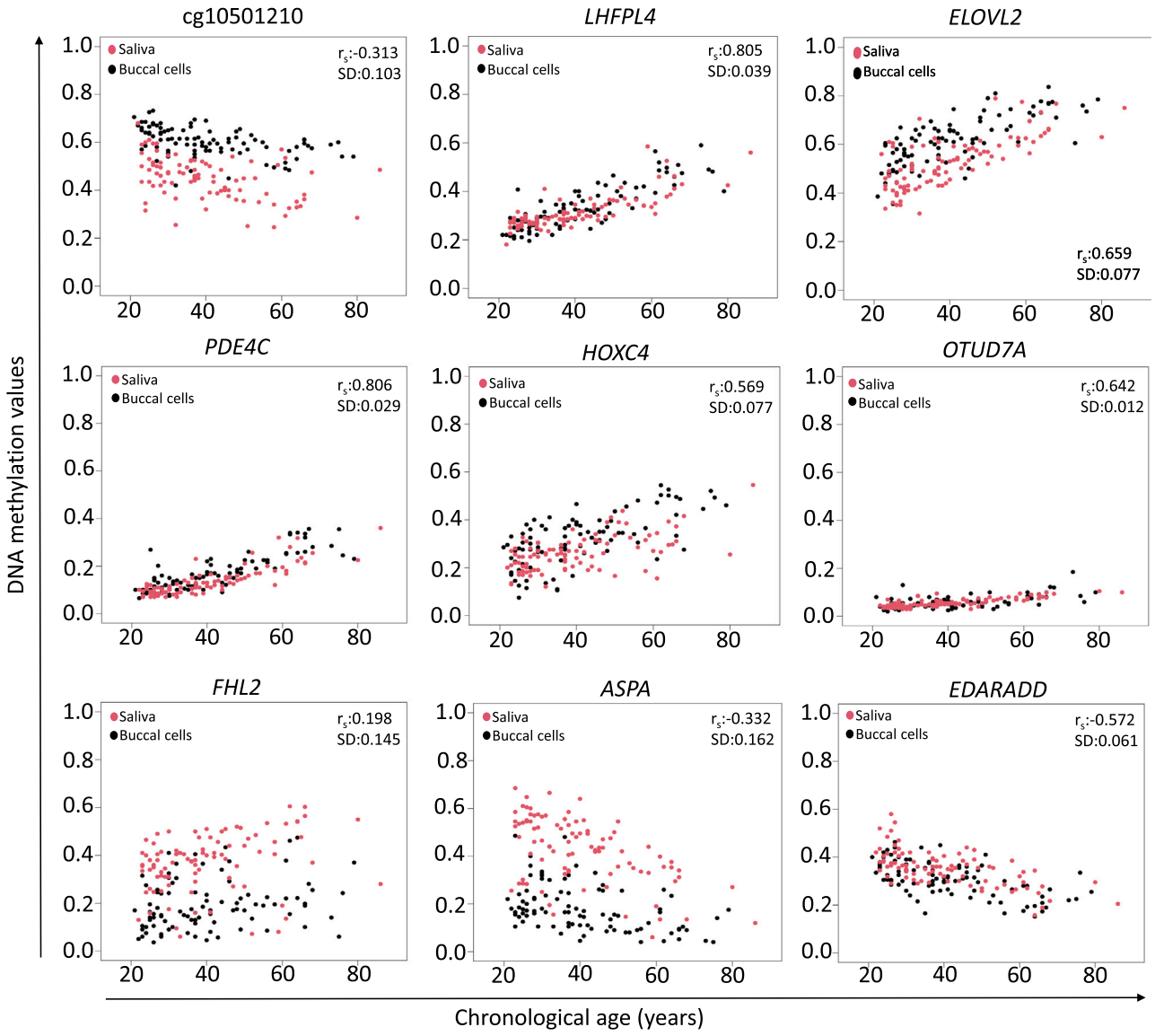

Fig. 3. Dispersion diagrams (DNA methylation values in front of chronological age) for cg10501210, *LHFPL4*, *ELOVL2*, *PDE4C*, *HOXC4*, *OTUD7A*, *FHL2*, *ASPA* and *EDARADD* (age correlated CpG sites) for 184 individuals from 21–86 years old (N = 91 saliva and N = 93 buccal swabs).

As well as cross-validation, an additional validation step consisted of a testing set of 184 samples (N = 93 saliva and N = 91 buccal swabs) ranging from 21–86 years old that were analyzed using the final age prediction model, providing an MAE of ± 3.66 % and 71.2 % correct classifications. The final online age prediction model developed in our study has now been placed in the open-access *Snipper* forensic classification website and is freely available at: http://mathgene.usc.es/cgi-bin/snps/age_tools/processmethylation-saliva-buccalswab.cgi. The underlying model equations for predicted age and prediction intervals are the following:

Predicted age in years = 29·33 - (50·52 x cg10501210) + (9·23 x *LHFPL4*) + (36·46 x *ELOVL2*) + (74·32 x *PDE4C*) + (11·23 x *HOXC4*) + (84·74 x *OTUD7A*) - (15·03 x *EDARADD*)·

Minimum Prediction (MinPred – q10) = 29·36 - (42·87 x cg10501210) + (15·41 x *LHFPL4*) + (11·09 x *ELOVL2*) + (74·17 x *PDE4C*) + (32·51 x *HOXC4*) + (29·13 x *OTUD7A*) - (20·54 x *EDARADD*)·

Maximum Prediction (MaxPred – q90) = 11·3 - (43·57 x cg10501210) + (20·74 x *LHFPL4*) + (54·72 x *ELOVL2*) + (78·25 x *PDE4C*) - (7·06 x *HOXC4*) + (179·95 x *OTUD7A*) + (4·16 x *EDARADD*)·

Once the age prediction model was generated, the possibility that the tissue could be considered as an additional variable was evaluated. To assess this, the prediction model was generated again by adding the tissue-of-origin of each of the samples in the training set as a co-variable. For this extended model, an MAE of ± 3.84 years, RMSE of 6.31 and % CP±PI of 78.22 % was obtained after cross-validation. Next, in order to evaluate the test set, the 2-CpG prediction model was used to predict the tissue-of-origin of the test samples. Adding the inferred tissue to the test set, produced an MAE of ± 3.78 years, RMSE of 6.6 and %CP±PI of 70.11 %. Comparing these results with those obtained when using the model without tissue source prediction, indicates the tissue as a co-variable does not improve the model.

3.5. Forensic validation of the age prediction model

To evaluate the predictive tests developed for the analysis of typical forensic samples with degradation and/or low-level DNA, the robustness and sensitivity of the final model were assessed.

A chain of models was generated by deleting one of the CpGs included in the final model, simulating random loss of one of the markers. For each of the six CpGs models generated, the training set was

Table 4
Summary of predictive performance metrics for the eight multivariate quantile regression models tested, based on three training sets: the saliva training set (N = 184 saliva, 21–86 years old), the buccal swab training set (N = 184 buccal swabs, 21–86 years old) and the combined training set (N = 91 saliva and N = 93 buccal swabs, 21–86 years old). All data represent the k-fold cross-validation. The selected model, based on the best balance between error and correct classification, is marked in bold. MAE: median absolute error, MAE_{mean}: mean absolute error, RMSE: root-mean-square error and %CP±PI: percentage of correct classifications within the prediction intervals.

Tissue	Model	CpG number	MAE	MAE_{mean}	RMSE	%CP±PI
Saliva	Model 1	9 CpGs	±3.17	±4.79	6.46	76.55 %
	Model 2	8 CpGs with *ASPA* excluded	±2.98	±4.66	6.4	77.11 %
	Model 3	8 CpGs with *FHL2* excluded	±3.29	±4.76	6.47	75.49 %
	Model 4	8 CpGs with cg10501210 excluded	±3.79	±5.04	6.76	75.59 %
	Model 5	7 CpGs with cg10501210 and *FHL2* excluded	±3.85	±5.17	6.93	76.61 %
	Model 6	7 CpGs with cg10501210 and *ASPA* excluded	±3.96	±4.97	6.69	74.45 %
	Model 7	7 CpGs with *FHL2* and *ASPA* excluded	±3.31	±4.69	6.37	78.74 %
	Model 8	6 CpGs with cg10501210, *FHL2* and *ASPA* excluded	±4.02	±5.10	6.91	78.74 %
Buccal swab	Model 1	9 CpGs	±3.85	±5.01	6.35	75.47 %
	Model 2	8 CpGs with *ASPA* excluded	±4.41	±5.09	6.45	74.91 %
	Model 3	8 CpGs with *FHL2* excluded	±4.13	±4.90	6.24	75.53 %
	Model 4	8 CpGs with cg10501210 excluded	±4.45	±5.27	6.66	76.64 %
	Model 5	7 CpGs with cg10501210 and *FHL2* excluded	±4.89	±5.52	6.94	74.42 %
	Model 6	7 CpGs with cg10501210 and *ASPA* excluded	±4.22	±5.15	6.63	73.86 %
	Model 7	7 CpGs with *FHL2* and *ASPA* excluded	±4.16	±4.99	6.36	75.47 %
	Model 8	6 CpGs with cg10501210, *FHL2* and *ASPA* excluded	±4.72	±5.43	6.86	76.64 %
Combined (saliva and buccal swabs)	Model 1	9 CpGs	±3.31	±4.57	6.06	74.38 %
	Model 2	8 CpGs with *ASPA* excluded	±3.66	±4.75	6.20	74.99 %
	Model 3	8 CpGs with *FHL2* excluded	±3.67	±4.78	6.32	74.36 %
	Model 4	8 CpGs with cg10501210 excluded	±3.78	±5.05	6.52	77.72 %
	Model 5	7 CpGs with cg10501210 and *FHL2* excluded	±4.18	±5.43	6.96	77.28 %
	Model 6	7 CpGs with cg10501210 and *ASPA* excluded	±3.77	±4.93	6.39	80.96 %
	Model 7	**7 CpGs with *FHL2* and *ASPA* excluded**	**±3.54**	**±4.79**	**6.23**	**76.08 %**
	Model 8	6 CpGs with cg10501210, *FHL2* and *ASPA* excluded	±5.23	±5.93	7.54	74.06 %

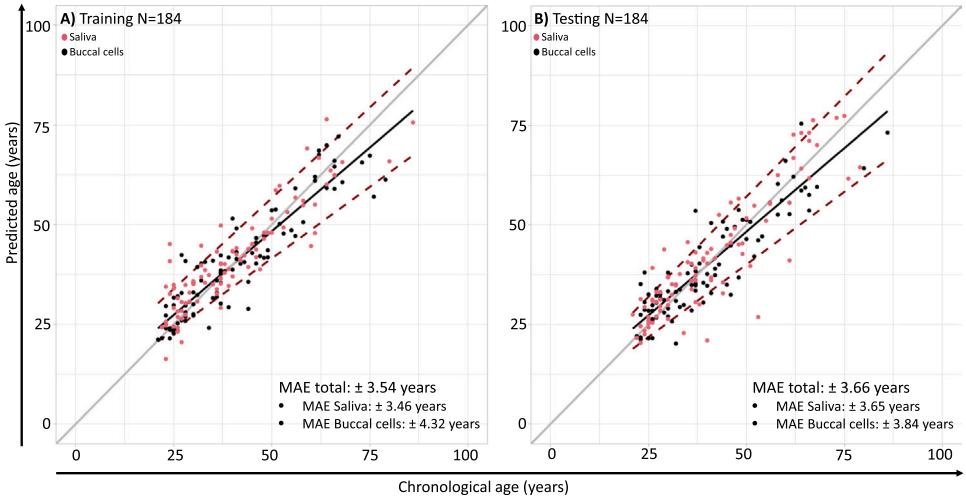

Fig. 4. Predicted versus chronological age for the final age prediction model for saliva and buccal cells for A) the training set composed of 184 individuals from 21–86 years old (N = 91 saliva and N = 93 buccal swabs) and for B) the testing set composed of 184 samples from 21–86 years old (N = 93 saliva and N = 91 buccal swabs). Predictions were performed under multivariate quantile regression using seven markers: cg10501210, *LHFPL4*, *ELOVL2*, *PDE4C*, *HOXC4*, *OTUD7A* and *EDARADD*. The black diagonal line represents the 0.5 quantile and the discontinuous dark red lines the corresponding 0.1 and 0.9 quantiles. The gray line represents perfect correlation. The data represent the k-fold cross-validation.

evaluated by cross-validation and the test set. Results are outlined in Supplementary Table S4. This analysis identified those markers with the strongest contribution to the final age prediction model. The exclusion of cg10501210 increased the MAE to ± 5.23 years in the cross-validation of the training set, and exclusion of *PDE4C* increased the classification error to ± 4.03 years in the testing set. Excluding the four other markers did not greatly affect the errors obtained compared to the full 7-CpG model, so the impact of their loss is minimal.

Lastly, bisulfite conversion was performed using 100 ng of genomic DNA. To evaluate if lower quantities of input DNA could produce results of comparable quality, serial dilutions were tested on two individuals (23 and 79 years old) for both saliva and buccal cells, using input DNA quantities for bisulfite conversion of 100 ng, 75 ng, 50 ng, 25 ng, 10 ng and 1 ng. The corresponding DNA methylation values and predicted ages are listed in Supplementary Table S5. To evaluate the differences detected in DNA methylation values between input DNAs, the standard deviation (SD) was used for comparisons (Supplementary Table S6). No standard deviations higher than 0.1 were observed in any of the markers up to 10 ng. For 1 ng only 4 markers presented a higher deviation than 0.1: *ELOVL2* (SD=0.19 and SD=0.20) in two of the four samples analyzed, *RUNX1* (SD=0.20) in one sample, cg10501210 (SD=0.13) and *HOXC4* also in single samples (SD=0.16).

4. Discussion

Individual age estimation has been a topic of great interest in forensic genetics for the last years. DNA methylation has become the biomarker of choice for inferring this characteristic [45], with prediction models published using several techniques [19,20,24,38,46,47] and different tissues [18–29], although most of them have focused on blood samples. Other tissues of relevance for forensic DNA analysis, and for which age prediction models are beginning to be developed are saliva

and buccal cells [22–25]. Cellular composition of saliva and buccal swab samples has been shown to be different, with saliva composed of a majority of leukocytes and buccal swab samples of epithelial cells [31]. However, it has also been observed in previous studies that cellular proportions can vary greatly between individuals, with the saliva samples containing a variable quantity of leukocytes in the range 16–95 %, and buccal swab samples between 5 %–65 % [22,48]. Taking this into consideration, an initial step of the present study was to develop a prediction model in order to infer the tissue of origin.

The selected tissue prediction markers have not been previously reported. Each marker was selected considering differences between pairs of tissues: saliva versus buccal cells (*HUNK*), blood versus buccal cells (*RUNX1*) and blood versus saliva (*RIN2*). From these three candidate markers, *RIN2* showed no variation in the DNA methylation patterns for the tissues of interest (saliva and buccal swabs) and therefore, was discarded from subsequent analyses. In contrast, differences were distinct for *RUNX1* and particularly *HUNK* (Fig. 2), with this pair showing opposite trends in DNA methylation levels (average DNA methylation: 0.576 for saliva versus 0.199 for buccal cells in *HUNK*, and 0.297 for saliva versus 0.670 for buccal cells in *RUNX1*). Likely due to the possible variations in the composition of the tissues collected (higher percentages of leukocytes or epithelial cells) [22], in some samples, differences were also observed within the same tissue, for example *RUNX1* gave differences up to 0.34 between some saliva samples. Additional cell-specific markers such as *CD6, SERPINB5* [22] and *PTPN7* [25] have been reported in other studies. The selection of these different markers could be due to screens made of alternative datasets. For the selection of *CD6* and *SERPINB5,* Eipel et. al used datasets GSE50586 [35] and GSE39981 [49], the former with data from buccal swab samples and the latter from blood samples. Using these data in combination, they selected CpGs that showed differences according to the tissue of origin. It should be noted that in our case we only used GSE50586 to evaluate whether tissue-specific markers related to buccal cells were correlated with age. On the other hand, the selection of *PTPN7* came from the Hong et al. study evaluating DNA methylation differences between blood and buccal cells. In our case, we selected a dataset containing samples of different tissues for each individual [34], trying to limit the possible differences between individuals related to the varied cellular proportions in saliva and buccal swab samples.

HUNK (hormonally up-regulated Neu-associated kinase) is a gene predicted to be involved in intracellular signal transduction and protein phosphorylation, while the protein encoded by *RUNX1* (*RUNX* family transcription factor 1) is involved in the development of normal hematopoiesis. Once both genes were selected as candidate markers for the inference of the tissue-of-origin, logistic regression was an informative system to explore the most accurate combination of markers, i.e., model 1 (*HUNK* and *RUNK1*), model 2 (*HUNK* only) and model 3 (*RUNX1* only). The main difference between double CpG-sites and each of the single CpG-site models was the detected imbalance between the sensitivity and specificity. While the 2-CpG-site model had a higher sensitivity than specificity (0.96 versus 0.82, respectively), the opposite was observed for the single-site models (0.78 versus 0.96 for model 2, and 0.81 versus 0.9 for model 3). Selection of the most accurate tissue prediction model was subsequently based on the additional metric of correct classification rate, with model 1 giving the best predictive performance of 88.59. Nevertheless, classifying these types of samples is complicated by the wide range of cellular proportions discussed above, as well as the admixed nature of some forensic specimens, e.g., cigarette butts. Therefore, for the second stage of the reported study, the generation of an age prediction model for oral cavity fluids which covered both saliva and buccal cells, was considered a better strategy than the development of different models for independent tissues. Even so, independent age prediction models for saliva and buccal cells were explored. Although the saliva-specific model showed the most accurate prediction (average MAE: ± 3.55), we decided to focus on the combined model since it will cover the maximum cell proportion variability in most forensic scenarios covering these samples.

To identify the most accurate age prediction model amongst nine saliva/buccal cell age correlated CpGs, different combinations of CpG sites were explored under multivariate quantile regression analysis testing up to eight different combined models (Table 4). Different age prediction models have been published based on different statistical tools, including linear regression [19,23–25,36,46,50], quadratic regression [28], machine learning [46] and quantile regression [20]. Although linear regression is the most commonly applied statistical analysis for age prediction, in this study quantile regression was selected, as its main advantage is the ability to provide age-specific prediction intervals, in addition to the predicted age.

From the CpG combinations tested, the selection of the most accurate age prediction model 7 was based on the best balance between error and the correct classification rate, with an MAE of ± 3.54, RMSE: 6.23 and % CP±PI: 76.08 %. Model 7 comprised CpG sites cg10501210, *LHFPL4*, *ELOVL2*, *PDE4C*, *HOXC4*, *OTUD7A* and *EDARADD*, and discarding *FHL2* and *ASPA*. Their contribution to the tested models is insufficient to improve predictive performance. This was not unexpected given the low age-correlations displayed (0.198 and −0.332, respectively) plus high levels of tissue dispersion (SD= 0.145 and 0.162, respectively).

Previous age predictors targeting the oral cavity have been developed as tissue-independent models, obtaining prediction errors close to ± 5 years; including, Bocklandt et. al [9], Eipel et. al [22] and Schwender et. al [24] with reported MAEs of ± 5.2 (saliva), ± 4.3 years (buccal cells) and ± 5.11 years (buccal cells), respectively. Common to all three studies is the use of just three CpG sites compared to the seven of the present study, which could explain the higher prediction errors observed. Additional tissue-independent models presenting prediction errors similar to the present study such as Hong et. al [25], Jung et. al [23] and Wozniak et. al [50] with MAEs of ± 3.13 (saliva), ± 3.55 years (buccal cells) and ± 2.5 years (buccal cells), respectively, were based on 5–7 CpG sites. While the models presented by Hong et al. and Wozniak et al. are uniquely focused on saliva and buccal swab samples, respectively, the combined model developed by our study covers both tissues, being more reliable in forensic scenarios where a mixture of saliva and buccal cells is under study, such as cigarette butts. A similar strategy to the present study was developed by Jung et al. [23], building a 5-CpG tissue-combined age prediction model, including saliva, buccal swabs and blood samples. The prediction error obtained was MAE: ± 3.55, practically identical to the present study. Considering all these results and the fact that models of other tissues also systematically present errors close to ± 3 years, it is reasonable to conclude that the lowest error obtainable with current technologies has been reached. Independently of the tissues covered, the main improvement provided by the prediction model proposed in the present study in comparison to the previous ones is the underlying statistical method used – quantile regression – providing not only the predicted age but the age-specific prediction intervals as well. Since errors are usually narrower at younger samples rather than at older individuals, to provide a specific interval of ages could improve the accuracy of results.

Considering the models discussed above, it is evident that certain markers appear recurrently in multiple age predictors for saliva and buccal swabs, namely *ELOVL2*, *PDE4C*, *EDARADD* and *KLF14*. Genes *ELOVL2*, *PDE4C* and *EDARADD* are present in our model but with different CpG positions (except cg09809672 in *EDARADD*, shared with Schwender's [24] model). Comparing the markers in these four genes in the other studies shows that only the CpG of *PDE4C* is shared between Eipel's [22] and Schwender's [24] models. In *KLF14,* not used in our study, only cg14361627 is shared between Hong's [25] and Jung's [23] models. This CpG is in the list of 49 CpGs of the preselected markers (Supplementary Table S1) but did not meet the selection criteria for our model. Our marker selection was based on the GSE92767 dataset [25], the same dataset used for Hong's marker selection but different markers were selected by each study using the same dataset. Different approaches were used for marker selection by Hong, with linear regression

and stepwise regression used to identify markers with an R^2 greater than 0.65 and a difference between maximum and minimum β-scores greater than 0.1. This compares to our use of Spearman's correlation to select markers with a correlation greater than |0.8| and a difference between extreme age donors greater than |0.3|. The motivation to change the selection criteria for marker selection when assessing the GSE92767 dataset was based on the lack of normality found for 15 % of the residuals of the models (independent linear regression models for each CpG on the dataset). Therefore, a non-parametric method such as the Spearman coefficient was found to be more suitable for this analysis.

Regarding markers included in our prediction model, cg10501210 was reported as a marker related with aging in blood monocytes [51], showing a similar DNA methylation trend when analyzing saliva and buccal cells samples in our study. Although less evident than for *FHL2* and *ASPA*, the correlation with age and tissue dispersion detected for this marker ($r_s = -0.313$ and SD= 0.103) suggested exclusion form the final age prediction model. However, its removal from the final model has the greatest effect, as shown in the robustness analysis. The gene *LHFPL4* (LHFPL tetraspan subfamily member 4), is a member of the superfamily of tetraspan transmembrane protein encoding genes. Mutations in one LHFP-like gene result in deafness in humans and mice, and a second LHFP-like gene is fused to a high-mobility group gene in a translocation-associated lipoma. To the best of our knowledge, our study detected this marker to be correlated with age in saliva and buccal cells for the first time. The cg11084334 CpG analyzed in *LHFPL4* presented amongst the highest age correlation values ($r_s = 0.805$), as well as showing minimal dispersion between tissues (SD= 0.039). Correlation with age in blood has been observed in other CpG positions of *LHFPL4* (cg24866418 and cg12841266) [52]. The gene *ELOVL2* (ELOVL fatty acid elongase 2) has been widely reported as a key age correlated marker [12,53,54] and has been incorporated in most of the age prediction models developed so far. This marker has been reported to correlate with age in multiple forensic tissues such as blood [19,20,23,28,50], saliva [23], buccal cells [23,24,50], teeth [28] and bones [50]. More specifically, it is noteworthy that the cg16867657 CpG analyzed in our study has been reported in other studies to be correlated with age either in blood [19,20,28], buccal cells [24] or teeth [28]. Gene *PDE4C* (phosphodiesterase 4 C) had the strongest correlation with age in saliva and buccal cells was ($r_s = 0.806$), and has been published in age prediction models for different tissues including saliva, buccal cells and blood [18,20,22,28,37]. In gene *HOXC4* (homeobox C4), the cg18473521 CpG analyzed in this work has shown correlation with age in blood samples [55]. Gene *OTUD7A* (OTU deubiquitinase 7A), which encodes a protein acting on TNF receptor associated factor 6 (TRAF6) to control nuclear factor kappa B expression, is used for the first time in an age prediction model in our study. Although *OTUD7A* has previously shown correlation with age in blood [20] and saliva [25], it was not included in published any model. Finally, *EDARADD* has been reported to show age correlated CpG positions, with cg09809672 used in this study also reported in previous blood, saliva, buccal cell and bone models [9,20,24,28,50].

Our studies showed the age predictive performance of the saliva and buccal cell model was not improved by adding tissue-of-origin information. A similar analysis was performed by Eipel et. al for buccal swab samples [22]. In Eipel's study, combined age and cell-type prediction models reported age prediction errors with this model (training MAD ± 4.66; testing MAD ± 5.09) that improved on age correlated markers only (training MAD: ± 4.3; testing MAD: ± 7.03). This suggests that introducing the cellular composition as a co-variable has more effect than the tissue of origin. Therefore, assessment of the cellular proportions may be the most effective way to introduce tissue-of-origin information as a co-variable in an age prediction model – certainly for the buccal cavity.

Finally, considering that in forensic DNA analysis degraded and low-level DNA concentrations are commonly encountered, our evaluations of the robustness of the model with missing data and amounts of input DNA for bisulfite conversion were particularly relevant.

Similar predictive performance was obtained for all step-wise exclusions of markers with the exception of cg10501210 (MAE: ± 5.23, and %CP±PI=74.06 %). The absence of this CpG produced the greatest increase in error. However, it should be noted that if missing data are present, incorrect DNA methylation measurement could be also occurring at the detected methylated and unmethylated peaks. In this case, to run duplicates or even triplicates of the sample is recommended in order to double-check the methylation values obtained.

An important factor for forensic sensitivity of methylation tests is the bisulfite conversion step, representing an aggressive reduction of the input DNA. Since use of 100 ng is not common practice in casework, the serial dilutions that were evaluated up to 10 ng showed no standard deviations greater than 0.1. For 1 ng input, some markers showed deviation values above the established limit for 3 of 4 samples. Thus, it is a viable strategy to start with a minimum of 10 ng of genomic DNA. Very similar results have been obtained by Aliferi et.al [21] and Wózniak et.al [50], indicating analyses with less than 10 ng of DNA caused significant variations in DNA methylation values. When comparing these studies to data reported here, it is worth noting that different technologies have been used, massive parallel sequencing versus SNaPshot, suggesting the limitation is not the detection methodology, but the DNA degradation or loss during bisulfite conversion process itself, or the stochastic variability of the analyzed molecules.

Acknowledgements

This project was funded by the Consellería de Cultura, Educación e Ordenación Universitaria e da Consellería de Economía, Emprego e Industria from Xunta de Galicia, Spain (Modalidade B, ED481B 2018/010) by a postdoctorate grant awarded to AFA. MVL is supported by the Ministerio de Educación, Cultura y Ciencia, Spain (PID2019-107876RB-I00).M.d.l.P. is supported by a post-doctorate grant funded by the Consellería de Cultura, Educación e Ordenación Universitaria e da Consellería de Economía, Emprego e Industria from Xunta de Galicia, Spain (ED481D-2021-008). J.R. is supported by the "Programa de axudas á etapa predoutoral" funded by the Consellería de Cultura, Educación e Ordenación Universitaria e da Consellería de Economía, Emprego e Industria from Xunta de Galicia, Spain (ED481A-2020/039).

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.fsigen.2022.102770.

References

[1] A. Freire-Aradas, C. Phillips, M. Lareu, Forensic individual age estimation with DNA: from initial approaches to methylation tests, Forensic Sci. Rev. 29 (2) (2017) 121–144.
[2] W. Parson, Age estimation with DNA: From forensic DNA fingerprinting to forensic (Epi) genomics: a mini-review, Gerontology 64 (4) (2018) 326–332.
[3] S. Walsh, F. Liu, A. Wollstein, L. Kovatsi, A. Ralf, A. Kosiniak-kamysz, et al., The HIrisPlex system for simultaneous prediction of hair and eye colour from DNA, Forensic Sci. Int Genet 7 (1) (2013) 98–115.
[4] M. Marcińska, E. Pośpiech, S. Abidi, J.D. Andersen, M. van den Berge, Á. Carracedo, et al., Evaluation of DNA variants associated with androgenetic alopecia and their potential to predict male pattern baldness, PLoS One 10 (5) (2015), e0127852.
[5] A. Abbott, DNA clock may aid refugee age check, Nature 561 (2018) 15.
[6] R. Noroozi, S. Ghafouri-Fard, A. Pisarek, J. Rudnicka, M. Spólnicka, W. Branicki, et al., DNA methylation-based age clocks: From age prediction to age reversion, Ageing Res Rev. 68 (2021), 101314.
[7] Z.D. Smith, A. Meissner, DNA methylation: roles in mammalian development, Nat. Rev. Genet 14 (3) (2013) 204–220.
[8] V.K. Rakyan, T.A. Down, S. Maslau, T. Andrew, T.P. Yang, H. Beyan, et al., Human aging-associated DNA hypermethylation occurs preferentially at bivalent chromatin domains, Genome Res 20 (4) (2010) 434–439.
[9] S. Bocklandt, W. Lin, M.E. Sehl, F.J. Sa, J.S. Sinsheimer, S. Horvath, et al., Epigenetic predictor of age, PLoS One 6 (6) (2011), e14821.

[10] J.T. Bell, P.-C. Tsai, T.-P. Yang, R. Pidsley, J. Nisbet, D. Glass, et al., Epigenome-wide scans identify differentially methylated regions for age and age-related phenotypes in a healthy ageing population, PLoS Genet 8 (4) (2012), e1002629.
[11] H. Heyn, M. Esteller, DNA methylation profiling in the clinic: applications and challenges, Nat. Rev. Genet 13 (10) (2012) 679–692.
[12] G. Hannum, J. Guinney, L. Zhao, L. Zhang, G. Hughes, Genome-wide methylation profiles reveal quantitative views of human aging rates, Mol. Cell 49 (2) (2013) 359–367.
[13] S. Horvath, DNA methylation age of human tissues and cell types, Genome Biol. 14 (10) (2013) R115.
[14] A. Johansson, S. Enroth, U. Gyllensten, Continuous aging of the human DNA methylome throughout the human lifespan, PLoS One 8 (6) (2013), e67378.
[15] C.A. Reynolds, Q. Tan, E. Munoz, J. Jylhävä, J. Hjelmborg, L. Christiansen, et al., A decade of epigenetic change in aging twins: genetic and environmental contributions to longitudinal DNA methylation, Aging Cell 19 (8) (2020), e13197.
[16] Y. Wang, R. Karlsson, E. Lampa, Q. Zhang, Å.K. Hedman, M. Almgren, Epigenetic influences on aging: a longitudinal genome-wide methylation study in old Swedish twins, Epigenetics 13 (9) (2018) 975–987.
[17] L.D. Moore, T. Le, G. Fan, DNA methylation and its basic function, Neuropsychopharmacology 38 (1) (2013) 23–38.
[18] C.I. Weidner, Q. Lin, C.M. Koch, L. Eisele, F. Beier, P. Ziegler, et al., Aging of blood can be tracked by DNA methylation changes at just three CpG sites, Genome Biol. 15 (2) (2014) R24.
[19] R. Zbieć-Piekarska, M. Sólnicka, T. Kupiec, A. Parys-proszek, Z. Makowska, A. Paleczka, et al., Development of a forensically useful age prediction method based on DNA methylation analysis, Forensic Sci. Int Genet 17 (2015) 173–179.
[20] A. Freire-Aradas, C. Phillips, A. Mosquera-Miguel, L. Girón-Santamaría, A. Gómez-Tato, M. Casares De Cal, 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.
[21] A. Aliferi, D. Ballard, M.D. Gallidabino, H. Thurtle, L. Barron, D. Syndercombe-Court, DNA methylation-based age prediction using massively parallel sequencing data and multiple machine learning models, Forensic Sci. Int Genet 37 (2018) 215–226.
[22] M. Eipel, F. Mayer, T. Arent, M.R.P. Ferreira, C. Birkhofer, U. Gerstenmaier, et al., Epigenetic age predictions based on buccal swabs are more precise in combination with cell type-specific DNA methylation signatures, Aging 8 (5) (2016) 1034–1048.
[23] S.-E. Jung, S. Min, S. Rom, E. Hee, K. Shin, H. Young, 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.
[24] K. Schwender, O. Holländer, S. Klopfleisch, M. Eveslage, M.F. Danzer, H. Pfeiffer, et al., Development of two age estimation models for buccal swab samples based on 3 CpG sites analyzed with pyrosequencing and minisequencing, Forensic Sci. Int Genet 53 (2021), 102521.
[25] S.R. Hong, S.E. Jung, E.H. Lee, K.J. Shin, W.I. Yang, H.Y. Lee, DNA methylation-based age prediction from saliva: high age predictability by combination of 7 CpG markers, Forensic Sci. Int Genet 29 (2017) 118–125.
[26] W.J. Lee, C.M. Choung, Y.J. Jung, H.Y. Lee, S.-K. Lim, A validation study of DNA methylation-based age prediction using semen in forensic casework samples, Leg. Med 31 (2018) 74–77.
[27] T.G. Jenkins, K.I. Aston, B. Cairns, A. Smith, D.T. Carrell, Paternal germ line aging: DNA methylation age prediction from human sperm, BMC Genom. 19 (1) (2018) 763.
[28] B. Bekaert, A. Kamalandua, S.C. Zapico, W.Van De Voorde, B. Bekaert, Improved age determination of blood and teeth samples using a selected set of DNA methylation markers, Epigenetics 10 (10) (2015) 922–930.
[29] H.Y. Lee, S.R. Hong, J.E. Lee, I.K. Hwang, N.Y. Kim, J.M. Lee, et al., Epigenetic age signatures in bones, Forensic Sci. Int Genet 46 (2020), 102261.
[30] L.E. Reinius, N. Acevedo, M. Joerink, G. Pershagen, S.-E. Dahlén, D. Greco, et al., Differential DNA methylation in purified human blood cells: Implications for cell lineage and studies on disease susceptibility, PLoS One 7 (7) (2012), e41361.
[31] C. Theda, S.H. Hwang, A. Czajko, Y.J. Loke, P. Leong, J.M. Craig, Quantitation of the cellular content of saliva and buccal swab samples, Sci. Rep. 8 (1) (2018) 6944.
[32] S. Horvath, J. Oshima, G.M. Martin, A.T. Lu, A. Quach, S. Felton, et al., Epigenetic clock for skin and blood cells applied to Hutchinson Gilford progeria syndrome and ex vivo studies, Aging (Albany NY) 10 (7) (2018) 1758–1775.
[33] S. Köchl, H. Niederstätter, W. Parson, DNA extraction and quantitation of forensic samples using the phenol-chloroform method and real-time PCR, Methods Mol. Biol. 297 (2005) 13–30.
[34] R.C. Slieker, S.D. Bos, J.J. Goeman, J.V.M.G. Bovée, R.P. Talens, R. Breggen, Van Der, et al., Identification and systematic annotation of tissue-specific differentially methylated regions using the Illumina 450k array, Epigenetics Chromatin 6 (1) (2013) 26.
[35] M.J. Jones, P. Farré, L.M. Mcewen, J.L. Macisaac, K. Watt, S.M. Neumann, et al., Distinct DNA methylation patterns of cognitive impairment and trisomy 21 in down syndrome, BMC Med Genom. 6 (2013) 58.
[36] Y. Huang, J. Yan, J. Hou, X. Fu, L. Li, Y. Hou, Developing a DNA methylation assay for human age prediction in blood and bloodstain, Forensic Sci. Int Genet 17 (2015) 129–136.
[37] J.J. Marqueta-Gracia, M. Álvarez-Álvarez, M. Baeta, L. Palencia-Madrid, E. Prieto-Fernández, R.J. Ordoñana, et al., Genetics differentially methylated CpG regions analyzed by PCR-high resolution melting for monozygotic twin pair discrimination, Forensic Sci. Int Genet 37 (2018) e1–e5.
[38] Y. Hamano, S. Manabe, C. Morimoto, S. Fujimoto, K. Tamaki, Forensic age prediction for saliva samples using methylation-sensitive high resolution melting: exploratory application for cigarette butts, Sci. Rep. 7 (5) (2017) 10444.
[39] F.M. You, N. Huo, Y.Q. Gu, M. Luo, Y. Ma, D. Hane, et al., BatchPrimer3: a high throughput web application for PCR and sequencing primer design, BMC Bioinforma. 9 (2008) 253.
[40] X. Robin, N. Turck, A. Hainard, N. Tiberti, F. Lisacek, J. Sanchez, et al., pROC: an open-source package for R and S+ to analyze and compare ROC curves, BMC Bioinforma. 12 (2011) 77.
[41] Koenker R., Portnoy S., Ng P., Zeileis A., Grosjean P., Ripley B. Package quantreg: Quantile Regression. 2015.
[42] Alfons A. Package cvTools: Cross-validation tools for regression models. 2015.
[43] Wickham H., Chang W. Package ggplot2: An implementation of the grammar of graphics. 2015.
[44] Team R Core. R, A Language and Environment for Statistical Computing, R Foundation for Statistical Computing,, Vienna, Austria, 2020 (Available from), ⟨https://www.r-project.org/⟩.
[45] A. Vidaki, B. Daniel, D.S. Court, Forensic DNA methylation profiling — potential opportunities and challenges, Forensic Sci. Int Genet 7 (5) (2013) 499–507.
[46] S.R. Hong, K. Shin, S. Jung, E.H. Lee, H.Y. Lee, Platform-independent models for age prediction using DNA methylation data, Forensic Sci. Int Genet 38 (2019) 39–47.
[47] H. Alghanim, K. Balamurugan, B. Mccord, Development of DNA methylation markers for sperm, saliva and blood indentification using pyrosequencing and qPCR/HRM, Anal. Biochem 611 (2020), 113933.
[48] C. Thiede, G. Prange-Krex, J. Freiberg-Richter, M. Bornhäuser, G. Ehninger, Buccal swabs but not mouthwash samples can be used to obtain pretransplant DNA fingerprints from recipients of allogeneic bone marrow transplants, Bone Marrow Transpl. 25 (5) (2000) 575–577.
[49] W.P. Accomando, J.K. Wiencke, E.A. Houseman, H.H. Nelson, K.T. Kelsey, Quantitative reconstruction of leukocyte subsets using DNA methylation, Genome Biol. 15 (3) (2014) R50.
[50] A. Wózniak, A. Heidegger, D. Piniewska-Róg, E. Pośpiech, C. Xavier, A. Pisarek, et al., Development of the VISAGE enhanced tool and statistical models for epigenetic age estimation in blood, buccal cells and bones, Aging 13 (5) (2021) 6459–6484.
[51] L. Tserel, M. Limbach, M. Saare, K. Kisand, A. Metspalu, L. Milani, et al., CpG sites associated with NRP1, NRXN2 and miR-29b-2 are hypomethylated in monocytes during ageing, Immun. Ageing 11 (1) (2014) 1.
[52] H. Alsaleh, P.R. Haddrill, Identifying blood-specific age-related DNA methylation markers on the Illumina methylationEPIC BeadChip, Forensic Sci. Int 303 (2019), 109944.
[53] P. Garagnani, M.G. Bacalini, C. Pirazzini, D. Gori, C. Giuliani, D. Mari, et al., Methylation of ELOVL2 gene as a new epigenetic marker of age Aging Cell, Aging 11 (6) (2012) 1132–1134.
[54] H. Heyn, N. Li, H.J. Ferreira, S. Moran, D.G. Pisano, A. Gomez, et al., Distinct DNA methylomes of newborns and centenarians, Proc. Natl. Acad. Sci. USA 109 (26) (2012) 10522–10527.
[55] J. Naue, H.C.J. Hoefsloot, O.R.F. Mook, L. Rijlaarsdam-hoekstra, M.C.H.Van Der Zwalm, P. Henneman, et al., Chronological age prediction based on DNA methylation: massive parallel sequencing and random forest regression, Forensic Sci. Int Genet 31 (2017) 19–28.