

Northumbria Research Link

Citation: Aliferi, Anastasia, Ballard, David, Gallidabino, Matteo, Thurtle, Helen, Barron, Leon and Syndercombe Court, Denise (2018) DNA methylation-based age prediction using massively parallel sequencing data and multiple machine learning models. *Forensic Science International: Genetics*, 37. pp. 215-226. ISSN 1872-4973

Published by: Elsevier

URL: <http://dx.doi.org/10.1016/j.fsigen.2018.09.003>
<<http://dx.doi.org/10.1016/j.fsigen.2018.09.003>>

This version was downloaded from Northumbria Research Link:
<http://nrl.northumbria.ac.uk/id/eprint/35880/>

Northumbria University has developed Northumbria Research Link (NRL) to enable users to access the University's research output. Copyright © and moral rights for items on NRL are retained by the individual author(s) and/or other copyright owners. Single copies of full items can be reproduced, displayed or performed, and given to third parties in any format or medium for personal research or study, educational, or not-for-profit purposes without prior permission or charge, provided the authors, title and full bibliographic details are given, as well as a hyperlink and/or URL to the original metadata page. The content must not be changed in any way. Full items must not be sold commercially in any format or medium without formal permission of the copyright holder. The full policy is available online: <http://nrl.northumbria.ac.uk/policies.html>

This document may differ from the final, published version of the research and has been made available online in accordance with publisher policies. To read and/or cite from the published version of the research, please visit the publisher's website (a subscription may be required.)

DNA methylation-based age prediction using massively parallel sequencing data and multiple machine learning models

Anastasia Aliferi^a, David Ballard^{a,*}, Matteo D. Gallidabino^{a,b}, Helen Thurtle^a, Leon P. Barron^a, Denise Syndercombe Court^a

^a *King's Forensics, Department of Analytical, Environmental and Forensic Sciences, Faculty of Life Sciences and Medicine, King's College London, 150 Stamford Street, London SE1 9NH, United Kingdom*

^b *Centre for Forensic Science, Department of Applied Sciences, Faculty of Health and Life Sciences, Northumbria University Newcastle, Ellison Building, NE1 8ST Newcastle Upon Tyne, United Kingdom.*

Abstract

The field of DNA intelligence focuses on retrieving information from DNA evidence that can help narrow down large groups of suspects or define target groups of interest. With recent breakthroughs on the estimation of geographical ancestry and physical appearance, the estimation of chronological age comes to complete this circle of information. Recent studies have identified methylation sites in the human genome that correlate strongly with age and can be used for the development of age-estimation algorithms. In this study, 110 whole blood samples from individuals aged 11-93 years were analysed using a DNA methylation quantification assay based on bisulphite conversion and massively parallel sequencing (Illumina MiSeq) of 12 CpG sites. Using this data, 17 different statistical modelling approaches were compared based on root mean square error (RMSE) and a Support Vector Machine with polynomial function (SVMp) model was selected for further testing. For the selected model (RMSE= 4.9 years) the mean average error (MAE) of the blind test (n=33) was calculated at 4.1 years, with 52% of the samples predicting with less than 4 years of error and 86% with less than 7 years. Furthermore, the sensitivity of the method was assessed both in terms of methylation quantification accuracy and prediction accuracy in the first validation of this kind. The described method retained its accuracy down to 10 ng of initial DNA input or ~2ng bisulphite PCR input. Finally, 34 saliva samples were analysed and following basic normalisation, the chronological age of the donors was predicted with less than 4 years of error for 50% of the samples and with less than 7 years of error for 70%.

* Corresponding author: david.ballard@kcl.ac.com

1. Introduction

In a forensic era where intelligence information regarding an individual's physical appearance can be retrieved from DNA material [1-3], the accurate determination of chronological age from crime scene samples has the potential to significantly aid forensic investigations towards identifying and finding unknown individuals. In the majority of the forensic cases where intact skeletal remains are available, age determination can be conducted successfully by anthropological measurements and calculations as well as cross-referencing with medical records. In the quest of identifying the perpetrator of a crime, however, it is highly unlikely for the biological evidence to consist of something other than body fluids, shed hairs and/or fingerprints. In those scenarios, especially when any direct comparisons with DNA and/or fingerprint databases are unable to provide a definitive match, the need for an age prediction method based on biological material emerges.

Over recent years, several approaches employing biomarkers for age prediction have been investigated. Extensive research has focused on the correlation of age and telomere length [3-9], while biomarkers including the quantification of a 4977bp deletion in the mitochondrial genome [10-12], measurement of aspartic acid racemisation [13-15], detection of somatic gene rearrangement in T-cells via its products (sjTREC) [16], measurement of advanced glycation end product accumulation (AGEs) [17] and, finally, analysis of mRNA profiles [18] have also been investigated over the years. However, several restrictions including significant effects of variables other than age [9, 19-25], lack of precision and reproducibility [26, 27], as well as restrictions in age range [18, 25, 28] and tissue applicability [27, 29] are overshadowing the proposed age prediction methods.

Differentiation in gene expression governs to a significant extent most physiological processes, including ageing. As epigenetic factors are known for their key role in modulating gene expression, it comes as no surprise that current research on age estimation has navigated heavily towards this area [30-32]. Epigenetic factors comprise of post-translational histone modifications, nucleosomal remodelling, chromatin looping, certain non-coding RNAs and DNA methylation.

DNA methylation is a chemical modification that primarily affects cytosines when these are followed by guanines in a 5'-3' direction in the DNA double helix and, in mammalian cells, results in the addition of a methyl group (-CH₃) to their 5' carbon (C5). These 5'-3' CG methylation sites in the DNA are called 'CpG' dinucleotides and are mostly methylated in the human genome (70-80%) [33]. The unmethylated CpGs are predominantly encountered in groups of high CG density known as 'CpG islands' most commonly located at the 5' end of the regulatory region of genes [34]. Their position around the regulatory region of genes, as well as the fact that approximately 60% of human genes [34] can be linked to specific CpG islands, suggests methylation's key role in modulating gene

expression [31]. Although a general tendency for hypo-methylation with age has been observed in the human genome [35-37], recent studies have brought to light evidence suggesting that methylation can also occur de novo in regions related to key developmental genes [38-40], highlighting a role in the physiological process of human ageing.

Since the first evidence of correlation between human ageing and DNA methylation, a significant amount of research has been conducted in this direction. A methylation-based age prediction method incorporating data from various CpG sites was proposed as early as 2010 by Teschendorff et al. in their study [40], followed by Horvath's epigenetic 'ageing clock' formed by 353 age-correlated CpG sites in 2013 [41]. Furthermore, in addition to models using large numbers of methylation sites, a number of age estimation models have been successfully developed based on less than 10 CpG sites [30, 42, 43] suggesting the high potential of DNA methylation in age determination.

Recent studies on DNA methylation have also identified a number of characteristics that highlight the potential of this approach for forensic applications where the number of unknown variables in regard to the DNA source increases drastically. The majority of studies on DNA methylation and age correlation have failed to identify any sex-specific bias for age-related CpG sites [37, 42, 44-47], while a recent study suggests that ethnicity can also be excluded as a factor of bias [48]. Furthermore studies have shown no significant changes in the methylation status when comparing samples of living and deceased individuals [45, 49] or fresh and long stored samples [43, 46, 50-52] suggesting a high stability for DNA methylation that can contribute significantly to the robustness of a DNA methylation based assay.

On the other hand, there are points raised by the literature that require further investigation if a method is to be applied in forensics. While a number of studies have been able to identify CpG sites exhibiting similar correlation with age between two or more distinct tissues [31, 39-41, 49, 53, 54], several studies report tissue specificity of their age-related CpG sites and subsequent failure to reproduce predictions in multiple tissues [32, 41, 46, 55-60]. These results suggest that differential methylation with age might be similar or significantly different between different tissues depending on the specific CpG site and, therefore, when designing an age-prediction model for forensic applications the multi-tissue applicability of the method should be investigated thoroughly. Furthermore, it is possible that the ideal DNA methylation-based age estimation tool that would be applicable to all tissues might not be a feasible target and tissue specific models should be designed instead. A second issue highlighted by current research is that of the effect of various health and environmental factors on the methylation status of age related CpG sites [32, 41, 61-64] which introduces a new parameter that needs to be investigated when validating an age prediction model based on DNA methylation. Finally, two other

important factors to consider when designing a method for application in forensics is target size and sensitivity, as DNA evidence recovered from crime scenes is often of poor quality and low quantity. Methods based on amplicons over 300bp long [43, 44], or DNA material requirements higher than 100ng [30, 44, 46, 49, 53] might not be applicable to degraded DNA samples. A limited number of reports have been recently published using more forensically relevant quantities of bisulphite-converted DNA in the PCR stage. In their studies, Naue et al. as well as Zbieć-Piekarksa et al. have reportedly achieved high prediction accuracy using as little as 10ng in the PCR stage [43, 65], while similar studies have reported promising results using 20ng of DNA in the same stage [45, 66].

The quantification of DNA methylation for age estimation is currently conducted with four main methods: (i) pyrosequencing, (ii) the EpiTYPER system based on MALDI-TOF mass spectrometry, (iii) methylation SNaPshot and (iv) massively parallel sequencing (MPS). Although all these techniques have been successfully applied in age estimation, massively parallel sequencing appears to be the best candidate for forensic applications as it is characterised by high sensitivity, single base resolution and large multiplexing capabilities. Furthermore, the MPS technology has been successfully applied to multiple aspects of forensic analysis [67-71] and DNA methylation-based age prediction methods have been recently attempted on the MPS platform providing promising preliminary results, with one method reporting a mean average deviation (MAD) of 4.4 years between the true and the predicted age of the donors using 16 markers [72] and a second reporting a MAD of 3.2 years using both 13 or a sub-selection of 4 markers [65].

Finally, when it comes to correlating DNA methylation data and age in a prediction model, a variety of different statistical approaches have been adopted by the various research groups. Most prediction models so far have been developed using linear univariate [43, 46, 48, 53] or multivariate least-squares regression analysis [30, 44, 64]. However, the success of recent machine learning approaches [65, 72, 73] has begun to raise doubts on the ability of linear models to truly grasp the complexity of the relationship between the DNA methylation state and human ageing [44, 49, 61]. However, even with recent publications focusing on models applying complicated algorithms [65, 72, 73], there has been no detailed comparison, to this day, on the efficiency and complementarity of the different modelling approaches when it comes to methylation-based age prediction.

The aim of this study is to systematically optimise all the parameters of a methylation-based age estimation assay in order to make this type of analysis realistic in a forensic scenario. From assessing the robustness of the method through reproducibility experiments, to selecting the optimal statistical approach for the predictive modelling and investigating into parameters such as sensitivity and multi-

tissue applicability, this study attempts to break down a 12 CpG assay based on the marker selection by Vidaki et al. in 2017 [72] and place it under the forensic microscope.

2. Materials and Methods

2.1 Sample Collection

This study operated under ethical approval granted by the Biomedical Sciences, Dentistry, Medicine and Natural & Mathematical Sciences Research Ethics Subcommittee (BDM/13/14-30) in regards to sample collection from various tissues for DNA methylation analysis. Whole blood samples were collected from a total of 110 unrelated donors aged 11 to 92.9 years through venepuncture. Additionally, 34 saliva samples and 11 semen samples were collected from unrelated donors aged between 16-90.5 and 23-50 respectively by deposition in 15mL universal receptacles. Full informed consent regarding the analysis was acquired prior to sampling from the donors or their parents or legal guardians for the cases of under-aged individuals. Samples were stored at 4oC. All semen donors are also part of the saliva sample set, while blood donors originated from a different group of participants. All samples were unconnected to any disease study, and information was not collected regarding medical history of the donors in the effort to create an inclusive, unbiased dataset that would be representative of the general population.

2.2 DNA Standards of known methylation

In order to assess the sensitivity of the method two pre-mixed methylation standards (EpigenDx, Massachusetts, USA) corresponding to 5% and 25% methylation were used in inputs of 50, 25, 10 and 1ng.

2.3 DNA Extraction and Quantification

Genomic DNA extractions were carried out using a BioRobot®EZ1 automated purification instrument (Qiagen, Hilden, Germany) in combination with the EZ1 Blood kit for whole blood samples and the EZ1 DNA Investigator kit for all other samples. For semen samples, differential extraction was performed in combination with the BioRobot®EZ1 and EZ1 DNA Investigator kit, in order to separate the sperm and epithelial fractions and only the sperm fraction was analysed further. Samples were subsequently stored at -20°C. Quantification of DNA extracts was conducted using the Quantifiler® Trio DNA Quantification kit in combination with the ABI PRISM® 7500 Sequence Detection System, both provided by Thermo-Fisher Scientific (Massachusetts, USA). The

manufacturer's guidelines [74] were followed throughout the protocol in half volumes and all samples were quantified in duplicates.

2.4 Sodium Bisulphite Conversion

Treatment with sodium bisulphite was employed for the conversion of unmethylated cytosines to uracils in the DNA samples. A total of 50ng of DNA from each sample or calibration standard was treated using the MethylEdge® Bisulphite Conversion System (Promega Corporation, Wisconsin USA) and the converted DNA was eluted in 10 μ L of the elution buffer provided. Eluates were stored at 4°C for up to a week and at -20°C up to one month according to the manufacturer's guidelines [75]. The approximate recovery of DNA following bisulphite conversion using this chemistry has been calculated as 52% [76] and therefore the final concentration of the eluate is estimated at approximately 2.6ng/ μ L.

2.5 Age-associated CpG sites

This study is based on 12 out of the 16 CpG sites previously described by Vidaki et al. [72] in 2017. Removal of the remaining 4 markers was deemed necessary due to their poor amplification efficiency and low overall contribution to the prediction accuracy. Information on the location and gene association of the CpGs is displayed in Table 1. Primers used in this study are of the same design as those originally described by Vidaki et al. [72] with the exception of cg17274064 for which primers were redesigned to improve the amplification efficiency (Forward primer sequence: GGGAGGGAATAAGTATTTTTTTAA, Reverse primer sequence: ACAACTAAAATAACTCCACTTTC).

2.6 Amplification of the bisulphite-converted DNA

Amplification of DNA following bisulphite treatment was performed using two multiplex reactions, the first containing primers for the amplification of amplicons 1-7 (Table 1) and the second targeting amplicons 8-12 (Table 1). The Qiagen Multiplex PCR kit was used for both reactions in half volume (25 μ L). Each reaction comprised of 12.5 μ L of 2x Qiagen Multiplex PCR Master Mix (providing a concentration of 3mM MgCl₂), an additional 1 μ L of 25mM MgCl₂ solution for a final concentration of 4mM, 2 μ L (~5ng) of bisulphite treated DNA or calibration standard and 9.5 μ L of primer mix. The final concentration of primers in the two multiplex reactions ranged from 0.2 to 0.5 μ M depending on the efficiency of the primers (Table 2). The reaction conditions were: (1) 95°C for 15min, (2) 32

cycles consisting of 94°C for 30s, T_m (see Table 2) for 30s and 72°C for 30s, (3) 72°C for 4min followed by a hold at 4°C (Table 2).

2.7 Post-PCR Purification and Quantification

Following amplification, samples were purified using the MinElute® PCR Purification kit by Qiagen in order to remove unincorporated primer residues [77]. Elution was performed in 11 µL PCR-grade water. Prior to library preparation all samples were quantified using the Qubit® dsDNA HS Assay kit (Thermo Fisher Scientific) according to the manufacturer's guidelines [78] and in combination with the Qubit 2.0 Fluorometer instrument and clear thin-walled 0.5mL PCR tubes.

2.8 Library Preparation and Quantification

The preparation of sequencing libraries was performed with the Kapa Hyper Prep® kit for Illumina (Roche, Basel, Switzerland) starting with 50ng of purified PCR product per sample. Library preparation was performed according to the manufacturer's specifications [79] in half volumes. For the size selection stages, AMPure® XP Beads (Beckman Coulter Genomics, California USA) and Illumina Resuspension Buffer were used (Illumina, California USA). Finally, library amplification was performed for 8 cycles.

Quantification of the libraries was conducted with the KAPA Library Quantification Kit for Illumina platforms (Roche) [80]. Libraries were diluted 1:4000 in PCR-grade water prior to quantification and analysed in duplicate. Following quantification, DNA libraries were normalised to 20nM using Tris-HCL 10mM/ pH 8.5 with 0.1% Tween (EBT buffer) and were pooled together in equal amounts to a final volume of 240µL (24 samples per run). Following denaturation and dilution to 10pM, 500µL of library was mixed with 100µL of denatured 20pM PhiX control (Illumina) and loaded in the MiSeqFGx instrument (Illumina) using the MiSeq version 2 cartridge and reagents.

2.9 Sequencing

Sequencing of the libraries was performed using the Illumina MiSeqFGx benchtop instrument (Illumina). Sample sheets and sample plates were created in the Illumina Experiment Manager software and the instrument was set to perform paired-end sequencing of 151bp in each direction while the analysis workflow was set to 'FASTQ only'. The online platform Basespace® (Illumina) was used for monitoring the performance of the runs as well as retrieve the sequencing files.

2.10 Data Analysis and Normalisation

Analysis of the FASTQ files was conducted with the Burrows-Wheeler Aligner (BWA) [81], Sequence Alignment/Map (SAMtools) [82], and Genome Analysis Toolkit (GATK, Broad Institute, Massachusetts USA) [83] software. Sequences were aligned to a custom genome where all non-CpG cytosines were replaced by thymines. For CpG positions information was collected for the presence of both cytosines and thymines. Files were extracted in variant call format (VCF) using the R® Project for Statistical Computing software in combination with R Studio® platform and were subsequently processed with Microsoft Office Excel software. The methylation percentage (β -values) for the 12 targeted CpGs was calculated by comparing the number of cytosine reads (suggesting the presence of methylation) to the combined total of cytosine and thymine (suggesting the absence of methylation) reads at each locus. A similar analysis was carried out for all non-CpG cytosine sites in each locus in order to establish the conversion efficiency of the bisulphite treatment. Non-CpG cytosines are expected to be free of methylation [84, 85] and therefore should be converted to uracils and subsequently to thymines following bisulphite treatment and amplification. Any cytosines therefore detected in those positions are indicative of incomplete conversion and the methylation percentages for the relevant CpGs were corrected accordingly. Average methylation values between duplicates was calculated based on the number of sequencing reads for each duplicate and each marker, where the methylation value of the duplicate with the higher number of sequencing reads contributed accordingly high to the final methylation score for the relevant marker following the equation:

$$\begin{aligned}
 & \text{Average methylation value for CpGi} \\
 &= (\text{CpGi Methylation Value } a) * \left(\frac{\text{CpGi Reads } a}{\text{CpGi Reads } a + \text{CpGi Reads } b} \right) \\
 &+ (\text{CpGi Methylation Value } b) * \left(\frac{\text{CpGi Reads } b}{\text{CpGi Reads } a + \text{CpGi Reads } b} \right)
 \end{aligned}$$

Where CpGi corresponds to a specific marker and a and b correspond to the two replicates of the specific sample. Prior to statistical analysis and modelling methylation β -values were converted to M-values following the equation:

$$M_i = \log_2 \left(\frac{\beta_i}{1 - \beta_i} \right)$$

where M_i represents the M-value for a certain marker in a specific sample and β_i represents the equivalent β -value, and were normalised by centring around the median value for the dataset. Datasets corresponding to different tissues were normalised separately.

2.11 Age Prediction

For model fitting, the whole blood dataset was randomly split into training (70%, n=76) and validation (i.e., blind) subsets (30%, n=33). Using the training dataset, different software was then used to assess different kinds of models.

The Trajan® Neural Network Simulation package (Trajan Software Ltd., Durham UK) was employed in order to perform generalised regression neural network modelling (GRNN), an artificial neural network modelling approach that uses radial basis and linear functions together to rapidly learn from existing knowledge and produce a prediction output (i.e. age), and provide a direct comparison to previously reported results [72]. Parameter tuning was performed by holdout cross-validation using an internal verification subset composed by 10 out the 76 samples used for training. During each training round the software was set to develop 106 networks and display the best 50. Those networks were subsequently assessed on the degree and consistency in prediction accuracy across the training and verification subsets and the best networks were put through a new round of training until the point when no further improvement was observed.

Additionally, R project for statistical computing software version 3.3.3 in combination with the caret package [86] was employed in order to test fourteen regression methods: ordinary linear (LM), partial least squares (PLS), ridge regression (Ridge), elastic net (Enet), lasso regression (LASSO), bagging multivariate adaptive regression splines (BagMARS), k nearest neighbours (KNN), extreme learning machines (ELM), single-layer feedforward perceptron neural network using a single hidden layer (NNet.SLP) and two hidden layers (NNET.2MLP), support vector machines with radial (SVMr) and polynomial function (SVMp) as kernels, random forest exploiting classification trees algorithms as base learners (RFclass) and boosted trees (BT). Parameter tuning was performed by leave-one-out cross-validation.

All models were finally validated using the validation (blind) subset. Both the GRNN networks and R models were trained and blind tested using the same sample subsets.

2.12 Statistical analysis

The comparison of the different models was based on root-mean-square error (RMSE). Mean absolute error and median absolute error values were also used for the purpose of comparison with previous studies. In order to assess the similarity of the developed statistical models, residuals for the different samples were compared between the models using analysis of variance (ANOVA) and paired t-test with Bonferroni post-hoc correction. Additionally, a general classification and regression tree model,

pruning on error, was employed to group the different models based on similarity of the residuals. The attempt to combine the different approaches was carried out by selecting the best performing model of each group (based on RMSE) and averaging the predicted ages between those models.

3. Results and Discussion

3.1 Reproducibility assessment

The first step in assessing the validity of this method was to establish its reproducibility in terms of the quantification of DNA methylation. In order to investigate this, 110 blood samples were put through the process in duplicate starting from the bisulphite conversion stage. The average absolute difference in methylation between duplicates was calculated to be less than 3% for all markers (Fig. 1). Taking into account the range in DNA methylation observed in the 12 markers used in this study (between 20-40%) a maximum of 3% average difference in methylation quantification between replicates was considered a satisfactory result.

3.2 Development of an age prediction model using Generalised Neural Networks

In the prediction model described by Vidaki et al. the mean absolute prediction error (MAE) calculated during the development of the model using publicly available methylation data increased from 3.3 to 7.1 years of age when the model was applied to data collected in-house via an MPS method even after normalisation had been applied [72]. Variation in the technical aspects of a DNA methylation quantification method has been shown to affect the final methylation values obtained and, thus, would be expected to affect prediction accuracy between different datasets [42]. For this reason, this increase in the prediction error was believed to occur partly due to variations in the processing of samples as well as the sequencing techniques between the BeadChip array methodology, used to generate the data included in the publicly available datasets, and the forensically orientated method developed in this study. In order to investigate the magnitude of this platform-based effect, the age prediction model was retrained using the same statistical modelling approach but restricting the training dataset to data solely generated with the developed MPS method. Blood samples from 110 individuals of known age were analysed in duplicate and the results were subsequently used to train (66 samples), validate (10 samples) and blind test (33 samples) a generalised regression neural network for age prediction using the same software as the previous publication [72]. The distribution of different ages in the dataset is shown in Figure 2.

The mean absolute prediction error of the model was calculated at 0.8 years of age for the training set (n=66), 2.8 years of age for the validation set (n=10) and at 4.7 years for the external blind test set

(n=33) (Fig.3) placing the developed model amongst the most accurate ones using a limited number of markers published to this day [30, 44, 47, 65]. This prediction error is very similar to the expected mean absolute prediction error calculated in the BeadChip array training data of the original model (3.3 years) [72], suggesting that the previous loss of accuracy was indeed a result of inter-method variability. Furthermore, the high similarity in the accuracy of the two models is impressive given the fact that the model developed in this study was trained with approximately 13 times fewer samples than the original model developed by Vidaki et al. [72]. However, this apparently high performance is perhaps offset somewhat when comparing the training and verification set prediction accuracies to that of the external blind test set accuracy. It was apparent that this artificial neural network approach may have suffered from shortcomings with generalisability using fewer training cases (e.g. over-training, etc.).

3.3 Evaluating the efficiency of different modelling approaches

While statistical modelling using generalised regression neural networks (GRNN) generated an age prediction model with a relatively low mean absolute prediction error and root square mean error (RMSE=5.8 years), several factors, including its susceptibility to overfitting and loss of generalizability when the training dataset is small (n<1000), suggested that this modelling approach is less than ideal for this set of data. A GRNN ensemble model consisting of 6 individual GRNNs was also developed in an attempt to further stabilize the algorithm and increase its ability to generalize but the prediction error obtained in the blind test set for this model was very similar to that of the single GRNN (RMSE=6.1 years, MAE=4.9 years). In order to determine the optimal approach, 15 additional statistical models were trained using an identical training subset. While MAE is the most popular statistics for reporting prediction errors in this type of application, RMSE is a more appropriate statistic for comparing the performance of different prediction algorithms developed on the same dataset and thus it was chosen for this study. Based on this comparison and while linear models did not outperform the GRNN model, several non-linear approaches showed increased accuracy, with the support vector machine with polynomial function (SVMp) giving an RMSE of 4.9 years (MAE=4.1 years) in the validation subset (Fig.4).

Comparison of the individual sample residuals between models (i.e. the difference between the actual and predicted age for each sample, with every model) revealed no significant difference in the prediction error for 15 out of 17 models, with the exception of the Bagging Multivariate Adaptive Regression Splines (BagMARS) and the Neural Network 2 Layer Perceptron (NNet.2MLP). Furthermore, analysis of variance showed a significant difference in the error between the

NNet.2MLP and the rest of the models ($p < 0.0017$) on post hoc analysis. Finally, using a general classification tree model, pruning on the error, the different models were separated in 4 groups based on the similarity of the residuals for the different samples. Based on these observations, an attempt to combine the predictions across the different groups, by averaging the predicted ages for the different samples for the best performing model of each group with each other, was made but this did not result in any increase in accuracy compared to the best performing individual model (SVMp). This result suggests that a considerable proportion of the prediction error in this dataset is sample specific rather than depended on the statistics behind the prediction model used, meaning that different samples predict with a similar level of accuracy across all models tested. It is possible, however, that this is a result of the limited training dataset and a prediction method developed on a larger dataset could benefit from the combination of multiple independent statistical approaches that would introduce orthogonality into the statistics enabling confidence interval estimations. In this case, taking the previous results into account, the SVMp model was chosen for further applications of this age prediction method (Fig.5).

3.4 Sensitivity assessment

Having optimised this age prediction model and assessed the reproducibility of this analysis at standard amounts of DNA input (50ng), for the assay to be forensically relevant it should also be applicable to forensically relevant levels of DNA. The next steps of this study were designed to test this prediction method in a more realistic set of conditions, starting with a lower DNA input. Two methylation standards corresponding to 5% and 25% methylation were analysed starting with an initial input (before bisulphite conversion) of 50 (optimum), 25, 10 and 1ng. It is important to note that while these amounts correspond to the original DNA quantity used for the analysis, the final input in the PCR, following bisulphite conversion (approximately 52% recovery [76]) with elution at 10 μ L and use of 2 μ L for each of the 2 multiplex reactions, was calculated to be approximately 10, 5, 2 and 0.2ng respectively. The analysis was performed in duplicate and the methylation values obtained were compared with the average value for the 50ng input which is the optimum input upon which the method was developed (Fig.6). The accuracy in the quantification of DNA methylation was retained down to 10ng of initial input (~2ng in the PCR stage) and, while certain markers (cg07158339, cg0693994, cg20692569) retained their accuracy down to 1ng of initial input (~200pg in the PCR stage), for most markers an increase in the quantification error was observed when 1ng was used as starting material. These results correspond with the findings reported by Naue et al. in 2018, where through simulation experiments the authors suggest that an input in the order of 5ng can

be used to detect differences in methylation of approximately 10% but higher inputs are required in order to achieve resolution any higher than this [87].

Furthermore, whole blood samples from 6 donors of known age (21.3-79.7 years old) were also analysed in duplicate starting with different DNA quantities. The same initial DNA inputs of 50 (optimum), 25, 10 and 1ng were used and the methylation values obtained were used for the generation of age estimates from the age prediction model. Comparison of the error in age prediction (MAE) obtained for the different DNA inputs suggests that, while the accuracy of the predictions is retained down to 10ng of original DNA input (~2ng in the PCR stage), the error in age prediction increases significantly ($p < 0.05$), at approximately 5-fold, when 1ng is used as starting material (Fig.7). These results are a direct match to the results obtained from the previous experiment where the accuracy in methylation quantification is shown to be compromised for PCR inputs below ~2ng for most markers (Fig.6) and once again correspond with the observations made by Naue et al. in 2018 [87]. Although a limit of ~2ng is not comparable to that of highly sensitive forensic methods and thus further improvement is required for the method to be universally applicable to forensic investigations, it still has the best sensitivity reported to this day for DNA methylation based age prediction [30, 43-47], providing encouraging results for the future of this study. While previous publications have reported successful DNA methylation-based age prediction using DNA amounts as low as 10-20ng [43, 45, 65, 66], this value refers to the PCR input rather than the original DNA amount used for bisulphite conversion and thus corresponds to the highest input used in this study. Furthermore, given the fact that the proposed method is currently targeted at blood samples rather than contact traces, a higher DNA yield is expected. Finally, given the rapid increase observed in both quantification error and prediction accuracy between 10 and 1ng this range should be investigated further in the future in order to determine the true 'tipping point' in the sensitivity.

3.5 Application in different tissues

Finally, while this DNA methylation quantification method and age prediction model were developed on whole blood, their applicability to saliva and semen, which together with blood form the three most commonly encountered body fluids in forensic investigations, was investigated. A total of 34 saliva and the sperm fractions of 11 semen samples were analysed with the developed method. The reason behind choosing to analyse the sperm fraction instead of the whole semen lies with the increased complexity of a tissue consisting of both sperm and epithelial cells as well as the fact that in real life cases any potential age prediction assay would be performed following traditional DNA analysis which, in the majority of cases involving semen stains, would require differential extraction

for the isolation of the sperm fraction. Methylation values corresponding to the samples of the two tissues were put through the prediction model following normalisation and resulting in a successful age prediction for the saliva samples with a mean absolute prediction error of 7.3 years (Fig.8). The relatively low increase in the prediction error in saliva, especially given the fact that this method was targeted on whole blood and the prediction model was trained solely on data deriving from whole blood samples, suggests that a common model for DNA methylation-based age prediction can be developed for both whole blood and saliva. More importantly, these results correspond with previous findings suggesting that certain age-correlated DNA methylation markers can be applicable to more than one tissues [31, 39-41, 49, 53, 54]. The next step from this would be to re-train a model solely on saliva samples while keeping the same marker set. However, this was not possible in this study due to the limited number of available samples.

On the other hand, in the case of sperm tissue, no methylation (0%) was detected for any marker in any of the samples, deeming any attempt to predict chronological age with this marker set meaningless. Similar results, with age prediction methods applied successfully to both blood and saliva but failing to produce good results in semen, have been previously reported [60] and a possible explanation for these findings could be the methylation reprogramming that is known to occur during gamete formation [88]. Even though there are published studies suggesting that DNA methylation in semen can be used for age prediction it is not clear if these results represent the sperm or the epithelial fraction of the semen, since the analysis is taking place in whole semen samples [56, 60]. Even when the epithelial fraction is markedly low compared to the sperm fraction in a semen sample, large differences in the methylation values between the two could result in a notable difference in the DNA methylation quantification values obtained from whole semen when compared with sperm for the same sample. It is however, possible that through mechanisms of genetic imprinting, certain methylation sites in the gamete DNA would represent the chronological age of the donor and therefore could potentially be used in age prediction.

4. Conclusions

The results obtained from this study provide strong evidential support to recent publications suggesting that a DNA methylation-based age prediction method can be developed in a way applicable to forensic casework. The small amplicon size (<200bp) and the relatively high sensitivity (~10ng of initial DNA extracted from a crime scene stain or ~2ng of bisulphite treated DNA) suggest that there is potential for such a method to be applied to forensic samples of poor quality and/or low quantity. At the same time, the relatively large number of predictors, incorporated in the two

multiplex reactions described in this method, allows for the inclusion of multiple genomic locations and thus enhances the robustness [46] without compromising the sensitivity. Furthermore, this study addresses the issue of statistical modelling for methylation analysis of relatively small datasets and comes to the same conclusion as Xu et al. [73], suggesting that support vector machines offer a potentially more robust, accurate and generalizable modelling approach. With a training/validation set consisting of 76 whole blood samples, the developed model was able to successfully predict the chronological age of 33 new samples with a mean absolute error (MAE) of 4.1 years and a root mean square error (RMSE) of 4.9 years. While the MAE statistic is used throughout this study in order for the results to be comparable with relevant publications [30, 43, 45, 46], RMSE was used both to compare between the different models tested and to describe the accuracy of the selected model. This measure, previously adopted by other studies [30, 31, 46, 49, 60, 65, 89], was selected due to its ability to describe both the mean and the spread of the deviation/error within a specific dataset. In order to simplify any comparisons this study also reports the median absolute error (3.8 years) and the percentage of samples within a certain error ranges (52% of the samples predicting with less than 4 years of error and 86% with less than 7 years) for the final prediction model, following the layout chosen by previous publications [43, 44]. An important note on the reported accuracy of the model is that unlike most studies on forensically orientated DNA methylation-based age prediction that focus on adults (over the age of 18 years), in this study the dataset also includes younger individuals starting at 11 years of age. Removing these samples from the dataset could potentially improve the prediction accuracy of the proposed method even further as it has been shown that methylation patterns can differ between adulthood and childhood for certain markers [90] most likely due to the high activation of the immune system and development during the first years of life [91]. This is the first method, to our knowledge, to achieve a combination of age prediction accuracy and sensitivity of this magnitude and it provides strong evidence to suggest that a DNA methylation-based age prediction method applicable to forensic casework samples can be successfully developed. Furthermore, this prediction method was successfully applied to saliva samples with 50% of the samples predicting with less than 4 years of error and 70% with less than 7 years (MAE=7.3 years, RMSE=11.1 years), suggesting that an age estimation method applicable to multiple tissues is a realistic target for forensically orientated DNA methylation-based age prediction methods employing a limited number of predictors. Overall, this is the first study of its kind to take a DNA methylation-based age prediction method designed for forensic analysis further than the proof-of-concept stage, testing its sensitivity, statistical modelling and multi-tissue applicability, all at the same time. However, while this is a step forward towards the implementation of this type of analysis in the forensic field, it is only one of many required, with the

investigation of larger datasets as well as the use of extensive cross-validation being the first ones to follow.

References

- [1] W. Branicki, U. Brudnik, T. Kupiec, P. Wolańska-Nowak, A. Wojas-Pelc, Determination of phenotype associated SNPs in the MC1R Gene, *J. Forensic Sci.* 52 (2), 2007, 349-354.
- [2] O. Maroñas, C. Phillips, J. Söchtig, A. Gomez-Tato, R. Cruz, J. Alvarez-Dios, M.C. de Cal, Y. Ruiz, M. Fondevila, Á. Carracedo, M.V. Lareu, Development of a forensic skin colour predictive test, *Forensic Sci. Int. Genet.* 13, 2014, 34-44.
- [3] S. Walsh, F. Liu, K.N. Ballantyne, M. van Oven, O. Lao, M. Kayser, IrisPlex: A sensitive DNA tool for accurate prediction of blue and brown eye colour in the absence of ancestry information, *Forensic Sci. Int. Genet.* 5, 2011, 170-180.
- [4] A. Tsuji, A. Ishiko, T. Takasaki, N. Ikeda, Estimating age of humans based on telomere shortening, *Forensic Sci. Int.* 126 (3), 2002, 197-199.
- [5] F. Ren, C. Li, H. Xi, Y. Wen, K. Huang, Estimation of human age according to telomere shortening in peripheral blood leukocytes of Tibetan, *Am. J. Forensic Med. Pathol.* 30 (3), 2009, 252-255.
- [6] A.O. Karlsson, A. Svensson, A. Marklund, G. Holmlund, Estimating human age in forensic samples by analysis of telomere repeats, *Forensic Sci. Int. Genet. Suppl. Ser.* 1 (1), 2008, 569-571.
- [7] S. Hewakapuge, R.A.H. van Oorschot, P. Lewandowski, S. Baindur-Hudson, Investigation of telomere lengths measurement by quantitative real-time PCR to predict age, *Leg. Med.* 10 (5), 2008, 236-242.
- [8] E.L.B. Barrett, T.A. Burke, M. Hammers, J. Komdeur, D.S. Richardson, Telomere length and dynamics predict mortality in a wild longitudinal study, *Mol. Ecol.* 22 (1), 2013, 249-259.
- [9] P. Slijepcevic, DNA damage response, telomere maintenance and ageing in light of the integrative model, *Mech. Ageing Dev.* 129 (1-2), 2008, 11-16.
- [10] C. Meissner, P. Bruse, S.A. Mohamed, A. Schulz, H. Warnk, T. Storm, M. Oehmichen, The 4977 bp deletion of mitochondrial DNA in human skeletal muscle, heart and different areas of the brain: a useful biomarker or more?, *Exp. Gerontol.* 43 (7), 2008, 645-652.
- [11] C. Meissner, N. von Wurmb, B. Schimansky, M. Oehmichen, Estimation of age at death based on quantitation of the 4977-bp deletion of human mitochondrial DNA in skeletal muscle, *Forensic Sci. Int.* 105 (2), 1999, 115-124.

- [12] C. Meissner, N. von Wurmb, M. Oehmichen, Detection of the age-dependent 4977 bp deletion of mitochondrial DNA. A pilot study, *Int. J. Legal Med.* 110 (5), 1997, 288-291.
- [13] S. Ohtani, I. Abe, T. Yamamoto, An application of D- and L-aspartic acid mixtures as standard specimens for the chronological age estimation, *J. Forensic Sci.* 50 (6), 2005, 1298-1302.
- [14] R.C. Dobberstein, J. Huppertz, N. von Wurmb-Schwark, S. Ritz-Timme, Degradation of biomolecules in artificially and naturally aged teeth: implications for age estimation based on aspartic acid racemization and DNA analysis, *Forensic Sci. Int.* 179 (2-3), 2008, 181-191.
- [15] J. Lowenson, S. Clarke, Does the chemical instability of aspartyl and asparaginyl residues in proteins contribute to erythrocyte aging? The role of protein carboxyl methylation reactions, *Blood Cells* 14 (1), 1988, 103-118.
- [16] D. Zubakov, F. Liu, M.C. van Zelm, J. Vermeulen, B.A. Oostra, C.M. van Duijn, G.J. Driessen, J.J.M. van Dongen, M. Kayser, A.W. Langerak, Estimating human age from T-cell DNA rearrangements, *Curr. Biol.* 20 (22), 2010, R970-R971.
- [17] Y. Sato, T. Kondo, T. Ohshima, Estimation of age of human cadavers by immunohistochemical assessment of advanced glycation end products in the hippocampus, *Histopathology* 38 (3), 2001, 217-220.
- [18] M. Alvarez, J. Ballantyne, The identification of newborns using messenger RNA profiling analysis, *Anal. Biochem.* 357 (1), 2006, 21-34.
- [19] T. von Zglinicki, G. Saretzki, W. Döcke, C. Lotze, Mild Hyperoxia Shortens Telomeres and Inhibits Proliferation of Fibroblasts: A Model for Senescence?, *Exp. Cell Res.* 220, 1995, 186-193.
- [20] S. Oikawa, S. Tada-Oikawa, S. Kawanishi, Site-specific DNA damage at the GGG sequence by UVA involves acceleration of telomere shortening, *Biochemistry* 40 (15), 2001, 4763-4768.
- [21] T. von Zglinicki, Role of oxidative stress in telomere length regulation and replicative senescence, *Ann. N. Y. Acad. Sci.* 908 (1), 2000, 99-110.
- [22] S.W. Brouillette, J.S. Moore, A.D. McMahon, J.R. Thompson, I. Ford, J. Shepherd, C.J. Packard, N.J. Samani, Telomere length, risk of coronary heart disease, and statin treatment in the West of Scotland Primary Prevention Study: a nested case-control study, *Lancet* 369 (9556), 2007, 107-114.
- [23] K.D. Salpea, P.J. Talmud, J.A. Cooper, C.G. Maubaret, J.W. Stephens, K. Abelak, S.E. Humphries, Association of telomere length with type 2 diabetes, oxidative stress and UCP2 gene variation, *Atherosclerosis* 209 (1), 2010, 42-50.
- [24] N.M. Simon, J.W. Smoller, K.L. McNamara, R.S. Maser, A.K. Zalta, M.H. Pollack, A.A. Nierenberg, M. Fava, K.-K. Wong, Telomere shortening and mood disorders: preliminary support for a chronic stress model of accelerated aging, *Biol. Psychiatry* 60 (5), 2006, 432-435.

- [25] A. Aviv, W. Chen, J.P. Gardner, M. Kimura, M. Brimacombe, X. Cao, S.R. Srinivasan, G.S. Berenson, Leukocyte telomere dynamics: longitudinal findings among young adults in the Bogalusa Heart Study, *Am. J. Epidemiol.* 169 (3), 2009, 323-329.
- [26] L. Tomaska, J. Nosek, Telomere heterogeneity: taking advantage of stochastic events, *FEBS Lett.* 583 (7), 2009, 1067-1071.
- [27] P.M. Lansdorp, N.P. Verwoerd, F.M. van de Rijke, V. Dragowska, M.-T. Little, R.W. Dirks, A.K. Raap, H.J. Tanke, Heterogeneity in telomere length of human chromosomes, *Hum. Mol. Genet.* 5 (5), 1996, 685-691.
- [28] A. Pilin, F. Pudil, V. Bencko, Changes in colour of different human tissues as a marker of age, *Int. J. Legal Med.* 121 (2), 2007, 158-162.
- [29] S. Melov, J.M. Shoffner, A. Kaufman, D.C. Wallace, Marked increase in the number and variety of mitochondrial DNA rearrangements in aging human skeletal muscle, *Nucleic Acids Res.* 23 (20), 1995, 4122-4126.
- [30] C.I. Weidner, Q. Lin, C.M. Koch, L. Eisele, F. Beier, P. Ziegler, D.O. Bauerschlag, K. Jöckel, R. Erbel, T.W. Mühleisen, M. Zenke, T.H. Brümmendorf, W. Wagner, Aging of blood can be tracked by DNA methylation changes at just three CpG sites, *Genome Biol.* 15 (2), 2014, 24.
- [31] G. Hannum, J. Guinney, L. Zhao, L. Zhang, G. Hughes, S. Sada, B. Klotzle, M. Bibikova, J.-B. Fan, Y. Gao, R. Deconde, M. Chen, I. Rajapakse, S. Friend, T. Ideker, K. Zhang, Genome-wide methylation profiles reveal quantitative views of human aging rates, *Mol. Cell* 49 (2), 2013, 359-367.
- [32] S. Horvath, Y. Zhang, P. Langfelder, R.S. Kahn, M.P. Boks, K. van Eijk, L.H. van den Berg, R.A. Ophoff, Aging effects on DNA methylation modules in human brain and blood tissue, *Genome Biol.* 13 (10), 2012, 97.
- [33] M. Ehrlich, M.A. Gama-Sosa, L.H. Huang, R.M. Midgett, K.C. Kuo, R.A. McCune, C. Gehrke, Amount and distribution of 5-methylcytosine in human DNA from different types of tissues of cells, *Nucleic Acids Res.* 10 (8), 1982, 2709-2721.
- [34] F. Antequera, A. Bird, Number of CpG islands and genes in human and mouse, *Proc. Natl. Acad. Sci. U. S. A.* 90 (24), 1993, 11995-11999.
- [35] V.L. Wilson, P.A. Jones, DNA methylation decreases in aging but not in immortal cells, *Science* 220 (4601), 1983, 1055-1057.
- [36] E.G. Hoal-van Helden, P.D. van Helden, Age-related methylation changes in DNA may reflect the proliferative potential of organs, *Mutat. Res.* 219, 1989, 263-266.
- [37] Å. Johansson, S. Enroth, U. Gyllenstein, Continuous aging of the human dna methylome throughout the human lifespan, *PLoS One* 8 (6), 2013.

- [38] V.K. Rakyán, T.A. Down, S. Maslau, T. Andrew, T. Yang, H. Beyan, P. Whittaker, O.T. McCann, S. Finer, A.M. Valdes, R.D. Leslie, P. Deloukas, T.D. Spector, Human aging-associated DNA hypermethylation occurs preferentially at bivalent chromatin domains, *Genome Res.* 20 (4), 2010, 434-439.
- [39] C.M. Koch, W. Wagner, Epigenetic-aging-signature to determine age in different tissues, *Aging* 3 (10), 2011, 1018-1027.
- [40] A.E. Teschendorff, U. Menon, A. Gentry-Maharaj, S.J. Ramus, D.J. Weisenberger, H. Shen, M. Campan, H. Nouchmehr, C.G. Bell, A.P. Maxwell, D.A. Savage, E. Mueller-Holzner, C. Marth, G. Kocjan, S.A. Gayther, A. Jones, S. Beck, W. Wagner, P.W. Laird, I.J. Jacobs, M. Widschwendter, Age-dependent DNA methylation of genes that are suppressed in stem cells is a hallmark of cancer, *Genome Res.* 20 (4) (2010) 440-446.
- [41] S. Horvath, DNA methylation age of human tissues and cell types, *Genome Biol.* 14 (10), 2013, 115.
- [42] S. Bocklandt, W. Lin, M.E. Sehl, F.J. Sánchez, J.S. Sinsheimer, S. Horvath, E. Vilain, Epigenetic predictor of age, *PLoS One* 6 (6), 2011.
- [43] R. Zbieć-Piekarska, M. Spólnicka, T. Kupiec, Ż. Makowska, A. Spas, A. Parys-Proszek, K. Kucharczyk, R. Płoski, W. Branicki, Examination of DNA methylation status of the ELOVL2 marker may be useful for human age prediction in forensic science, *Forensic Sci. Int. Genet.* 14, 2015, 161-167.
- [44] A. Freire-Aradas, C. Phillips, A. Mosquera-Miguel, L. Girón-Santamaría, A. Gómez-Tato, M. Casares de Cal, J. Álvarez-Dios, J. Ansedo-Bermejo, M. Torres-Español, P.M. Schneider, E. Pośpiech, W. Branicki, Á. Carracedo, M.V. Lareu, 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.
- [45] Y. Hamano, S. Manabe, C. Morimoto, S. Fujimoto, M. Ozeki, K. Tamaki, Forensic age prediction for dead or living samples by use of methylation-sensitive high resolution melting, *Leg. Med.* 21, 2016, 5-10.
- [46] 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.
- [47] R. Zbieć-Piekarska, M. Spólnicka, T. Kupiec, A. Parys-Proszek, Ż. Makowska, A. Pałeczka, K. Kucharczyk, R. Płoski, W. Branicki, Development of a forensically useful age prediction method based on DNA methylation analysis, *Forensic Sci. Int. Genet.* 17, 2015, 173-179.

- [48] S.B. Zaghlool, M. Al-Shafai, W.A. Al Muftah, P. Kumar, M. Falchi, K. Suhre, Association of DNA methylation with age, gender, and smoking in an Arab population, *Clin. Epigenet.* 7 (1), 2015, 6.
- [49] B. Bekaert, A. Kamalandua, S.C. Zapico, W. Van de Voorde, R. Decorte, Improved age determination of blood and teeth samples using a selected set of DNA methylation markers, *Epigenetics* 10 (10), 2015, 922-930.
- [50] M.V. Hollegaard, J. Grauholm, B. Nørgaard-Pedersen, D.M. Hougaard, DNA methylome profiling using neonatal dried blood spot samples: A proof-of-principle study, *Mol. Genet. Metab.* 108 (4), 2013, 225-231.
- [51] N. Wong, R. Morley, R. Saffery, J. Craig, Archived Guthrie blood spots as a novel source for quantitative DNA methylation analysis, *Biotechniques* 45 (4), 2008, 423-424.
- [52] J.P. Antunes, T. Madi, K. Balamurugan, R. Bombardi, G. Duncan, B. McCord, DNA methylation markers as a powerful technique to discriminate body fluids present in crime scenes, *Proceedings of the 24th International Symposium on Human Identification*, 2014.
- [53] I. Florath, K. Butterbach, H. Müller, M. Bewerunge-Hudler, H. Brenner, Cross-sectional and longitudinal changes in DNA methylation with age: an epigenome-wide analysis revealing over 60 novel age-associated CpG sites, *Hum. Mol. Genet.* 23 (5), 2014, 1186-1201.
- [54] J.T. Bell, P.C. Tsai, T.P. Yang, R. Pidsley, J. Nisbet, D. Glass, M. Mangino, G. Zhai, F. Zhang, A. Valdes, S.Y. Shin, E.L. Dempster, R.M. Murray, E. Grundberg, A.K. Hedman, A. Nica, K.S. Small, C. MuTher, E.T. Dermitzakis, M.I. McCarthy, J. Mill, T.D. Spector, P. Deloukas, Epigenome-wide scans identify differentially methylated regions for age and age-related phenotypes in a healthy ageing population, *PLoS Genet.* 8 (4), 2012, e1002629.
- [55] F. Song, S. Mahmood, S. Ghosh, P. Liang, D.J. Smiraglia, H. Nagase, W.A. Held, Tissue specific differentially methylated regions (TDMR): changes in DNA methylation during development, *Genomics*, 93 (2), 2009, 130-139.
- [56] L.L. Ma, S.H. Yi, D.X. Huang, K. Mei, R.Z. Yang, Screening and identification of tissue-specific methylation for body fluid identification, *Forensic Sci. Int. Genet. Suppl. Ser.* 4 (1), 2013, e37-e38.
- [57] B. Kwabi-Addo, W. Chung, L. Shen, M. Ittmann, T. Wheeler, J. Jelinek, J.P. Issa, Age-related DNA methylation changes in normal human prostate tissues, *Clin. Cancer Res.* 13 (13), 2007, 3796-3802.
- [58] A. Zykovich, A. Hubbard, J.M. Flynn, M. Tarnopolsky, M.F. Fraga, C. Kerksick, D. Ogborn, L. MacNeil, S.D. Mooney, S. Melov, Genome-wide DNA methylation changes with age in disease-free human skeletal muscle, *Aging Cell* 13 (2), 2014, 360-366.

- [59] D. Soares Bispo Santos Silva, J. Antunes, K. Balamurugan, G. Duncan, C. Sampaio Alho, B. McCord, Evaluation of DNA methylation markers and their potential to predict human aging, *Electrophoresis* 36 (15), 2015, 1775-1780.
- [60] H.Y. Lee, S.-E. Jung, Y.N. Oh, A. Choi, W.I. Yang, K.-J. Shin, Epigenetic age signatures in the forensically relevant body fluid of semen: a preliminary study, *Forensic Sci. Int. Genet.* 19, 2015, 28-34.
- [61] A.E. Teschendorff, J. West, S. Beck, Age-associated epigenetic drift: implications, and a case of epigenetic thrift?, *Hum. Mol. Genet.* 22 (R1), 2013, R7-R15.
- [62] M.S. Almén, E.K. Nilsson, J.A. Jacobsson, I. Kalnina, J. Klovins, R. Fredriksson, H.B. Schiöth, Genome-wide analysis reveals DNA methylation markers that vary with both age and obesity, *Gene* 548 (1), 2014, 61-67.
- [63] M.L. Mansego, F.I. Milagro, M.Á. Zulet, M.J. Moreno-Aliaga, J.A. Martínez, Differential DNA methylation in relation to age and health risks of obesity, *Int. J. Mol. Sci.* 16 (8), 2015, 16816-16832.
- [64] S.H. Yi, L.C. Xu, K. Mei, R.Z. Yang, D.X. Huang, Isolation and identification of age-related DNA methylation markers for forensic age-prediction, *Forensic Sci. Int. Genet.* 11, 2014, 117-125.
- [65] J. Naue, H.C.J. Hoefsloot, O.R.F. Mook, L. Rijlaarsdam-Hoekstra, M.C.H. van der Zwalm, P. Henneman, A.D. Kloosterman, P.J. Verschure, Chronological age prediction based on DNA methylation: Massive parallel sequencing and random forest regression, *Forensic Sci. Int. Genet.* 31, 2017, 19-28.
- [66] S. Cho, S.-E. Jung, S.R. Hong, E.H. Lee, J.H. Lee, S.D. Lee, H.Y. Lee, Independent validation of DNA-based approaches for age prediction in blood, *Forensic Sci. Int. Genet.* 29, 2017, 250-256.
- [67] D.H. Warshauer, D. Lin, K. Hari, R. Jain, C. Davis, B. LaRue, J.L. King, B. Budowle, STRait Razor: a length-based forensic STR allele-calling tool for use with second generation sequencing data, *Forensic Sci. Int. Genet.* 7 (4), 2013, 409-417.
- [68] W. Parson, C. Strobl, G. Huber, B. Zimmermann, S.M. Gomes, L. Souto, L. Fendt, R. Delport, R. Langit, S. Wootton, R. Lagacé, J. Irwin, Evaluation of next generation mtGenome sequencing using the Ion Torrent Personal Genome Machine (PGM), *Forensic Sci. Int. Genet.* 7 (5), 2013, 543-549.
- [69] Y. Xue, Q. Wang, Q. Long, B.L. Ng, H. Swerdlow, J. Burton, C. Skuce, R. Taylor, Z. Abdellah, Y. Zhao, D.G. MacArthur, M.A. Quail, N.P. Carter, H. Yang, C. Tyler-Smith, Human Y chromosome base-substitution mutation rate measured by direct sequencing in a deep-rooting pedigree, *Curr. Biol.* 19 (17), 2009, 1453-1457.

- [70] J. Weber-Lehmann, E. Schilling, G. Gradl, D.C. Richter, J. Wiehler, B. Rolf, Finding the needle in the haystack: differentiating “identical” twins in paternity testing and forensics by ultra-deep next generation sequencing, *Forensic Sci. Int. Genet.* 9, 2014, 42-46.
- [71] Y. Yang, B. Xie, J. Yan, Application of next-generation sequencing technology in forensic science, *Genomics, Proteomics and Bioinformatics* 12 (5), 2014, 190-197.
- [72] A. Vidaki, D. Ballard, A. Aliferi, T.H. Miller, L.P. Barron, D. Syndercombe Court, DNA methylation-based forensic age prediction using artificial neural networks and next generation sequencing, *Forensic Sci. Int. Genet.* 28, 2017, 225-236.
- [73] C. Xu, H. Qu, G. Wang, B. Xie, Y. Shi, Y. Yang, Z. Zhao, L. Hu, X. Fang, J. Yan, L. Feng, A novel strategy for forensic age prediction by DNA methylation and support vector regression model, *Sci. Rep.* 5, 2015, 17788.
- [74] Quantifiler™ HP and Trio DNA Quantification Kits User Guide, 2017, Thermo Fisher Scientific.
- [75] MethylEdge™ Bisulfite Conversion System Instructions for use of product N1301, 2013, Promega Corporation.
- [76] C.A. Leontiou, M.D. Hadjidaniel, P. Mina, P. Antoniou, M. Ioannides, P.C. Patsalis, Bisulfite Conversion of DNA: Performance comparison of different kits and methylation quantitation of epigenetic biomarkers that have the potential to be used in non-invasive prenatal testing, *PLoS One* 10 (8), 2015, e0135058.
- [77] MinElute® PCR Purification Kit Quick Start Protocol, 2011, Qiagen Sample and Assay Technologies.
- [78] User Guide: Qubit® dsDNA HS Assay Kits for use with the Qubit® Fluorometer (all models), 2015, Life Technologies Molecular Probes.
- [79] KAPA Hyper Prep Kit Technical Data Sheet KR0961 – v5.16, 2016, KAPA Biosystems.
- [80] KAPA Library Quantification Kit Technical Data Sheet KR0405 – v8.17, 2017, KAPA Biosystems.
- [81] H. Li, R. Durbin, Fast and accurate short read alignment with Burrows–Wheeler transform, *Bioinformatics* 25 (14), 2009, 1754-1760.
- [82] H. Li, B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, R. Durbin, S. Genome Project Data Processing, The Sequence Alignment/Map format and SAMtools, *Bioinformatics* 25 (16), 2009, 2078-2079.
- [83] A. McKenna, M. Hanna, E. Banks, A. Sivachenko, K. Cibulskis, A. Kernytsky, K. Garimella, D. Altshuler, S. Gabriel, M. Daly, M.A. DePristo, The genome analysis toolkit: a MapReduce

- framework for analyzing next-generation DNA sequencing data, *Genome Res.* 20 (9), 2010, 1297-1303.
- [84] E. Li, Y. Zhang, DNA methylation in mammals, *Cold Spring Harb. Perspect. Biol.* 6 (5), 2014, a019133.
- [85] Y. Gruenbaum, R. Stein, H. Cedar, A. Razin, Methylation of CpG sequences in eukaryotic DNA, *FEBS Lett.* 124 (1), 1981, 67-71.
- [86] M. Kuhn, K. Johnson, *Applied predictive modeling*, 2013, Springer; New York (USA).
- [87] J. Naue, H.C.J. Hoefsloot, A.D. Kloosterman, P.J. Verschure, Forensic DNA methylation profiling from minimal traces: how low can we go?, *Forensic Sci. Int. Genet.* 33, 2018, 17-23.
- [88] W. Reik, W. Dean, J. Walter, Epigenetic reprogramming in mammalian development, *Science* 293 (5532), 2001, 1089-1093.
- [89] 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.
- [90] A. Freire-Aradas, C. Phillips, L. Girón-Santamaría, A. Mosquera-Miguel, A. Gómez-Tato, M.Á. Casares de Cal, J. Álvarez-Dios, M.V. Lareu, Tracking age-correlated DNA methylation markers in the young, *Forensic Sci. Int. Genet.* 36, 2018, 50-59.
- [91] N. Acevedo, L.E. Reinius, M. Vitezic, V. Fortino, C. Söderhäll, H. Honkanen, R. Veijola, O. Simell, J. Toppari, J. Ilonen, M. Knip, A. Scheynius, H. Hyöty, D. Greco, J. Kere, Age-associated DNA methylation changes in immune genes, histone modifiers and chromatin remodeling factors within 5 years after birth in human blood leukocytes, *Clin. Epigenet.* 7 (1), 2015, 34.

<i>Marker</i>	<i>CpG</i>	<i>Chromosomal location</i>	<i>Gene</i>
1	cg04084157	7: 100,809,049	VGF – nerve growth factor inducible precursor
2	cg02085507	19: 6,739,192	TRIP10 – thyroid hormone receptor interactor 10
3	cg04528819	7: 130,418,315	KLF14 – Kruppel-like factor 14
4	cg19761273	17: 80,232,096	CSNK1D – casein kinase 1; delta isoform 1
5	cg20692569	7: 72,848,481	FZD9 – frizzled 9
6	cg27544190	21: 33,785,434	C21orf63 – chromosome 21 open reading frame 63
7	cg01511567	11: 57,103,631	SSRP1 – structure specific recognition protein 1
8	cg22736354	6: 18,122,719	NHLRC1 – malin
9	cg17274064	21: 40,033,892	ERG – v-etserythroblastosis virus E26 oncogene like isoform 2
10	cg07158339	9: 71,650,237	FXN – frataxin, mitochondrial isoform 1 preproprotein
11	cg05442902	22: 21,369,010	P2RXL1 – purinergic receptor P2X-like 1; orphan receptor
12	cg06493994	6: 25,652,602	SCGN – secretagogin precursor

Table 1 - Chromosomal location and genetic information on the 12 CpG sites employed in this study.

Marker	CpG	Primer concentration in multiplex PCR (μ M)	Annealing temperature
1	cg04084157	0.3	50°C for the first 7 cycles and 48°C for the next 25 cycles
2	cg02085507	0.4	
3	cg04528819	0.3	
4	cg19761273	0.5	
5	cg20692569	0.5	
6	cg27544190	0.4	
7	cg01511567	0.2	
8	cg22736354	0.3	52°C
9	cg17274064	0.4	
10	cg07158339	0.4	
11	cg05442902	0.3	
12	cg06493994	0.2	

Table 2 - Details of the multiplex reactions employed in this study.

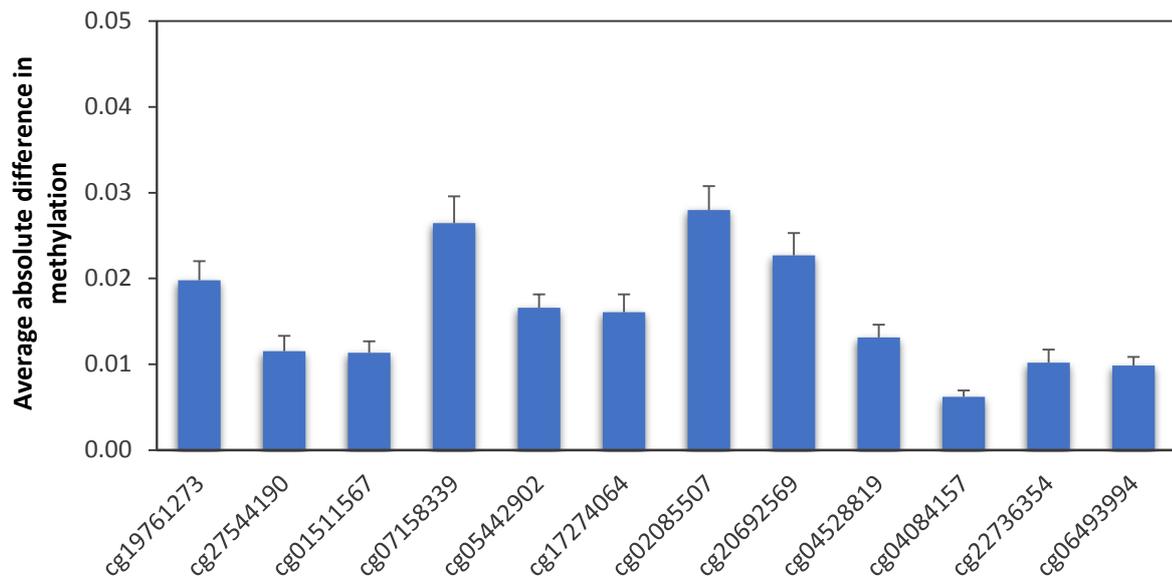


Figure 1 - Average absolute difference between duplicates (n=110) for the 12 different markers. The error bars represent the standard error.

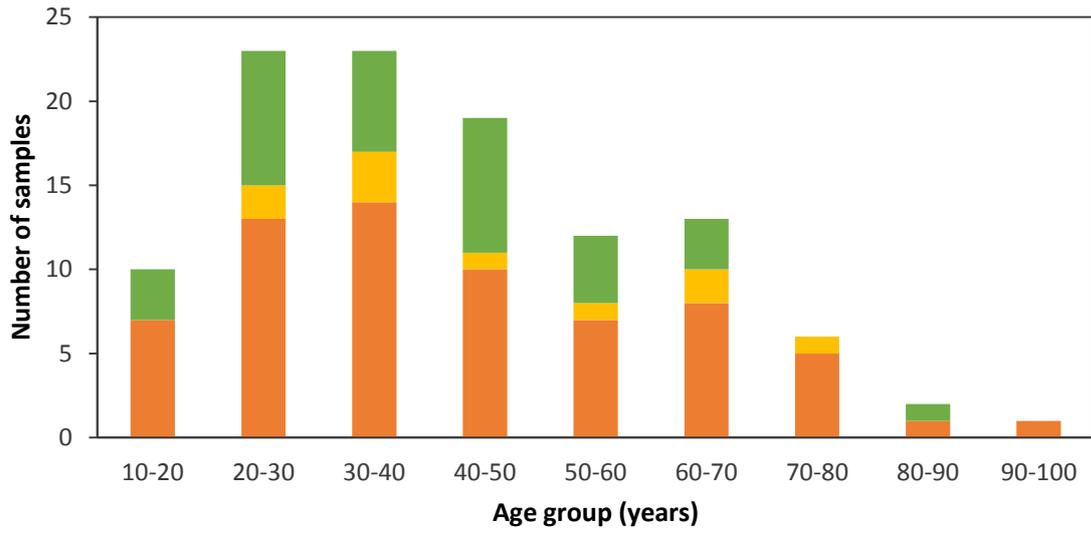


Figure 2 - Distribution of blood samples used in the model between the different age groups in the training (orange, n=66), validation (yellow, n=10) and test set (green, n=33).

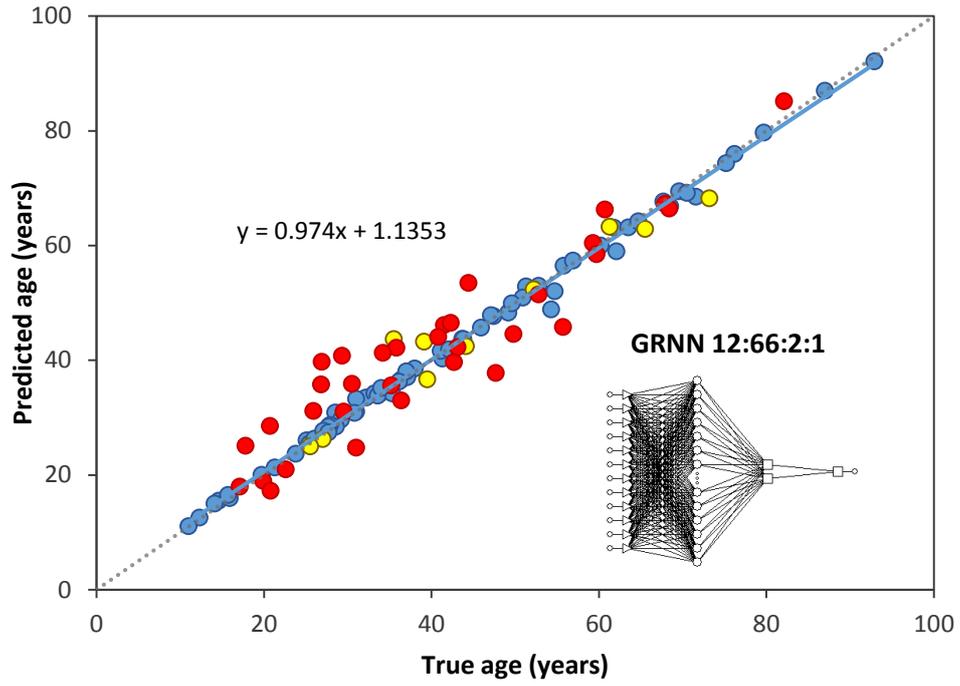


Figure 3 - Comparison between the predicted and the true age for the training (blue, n=66), validation (orange, n=10) and blind test set (red, n=33). The mean absolute prediction error was calculated at 0.8, 2.8 and 4.7 years respectively.

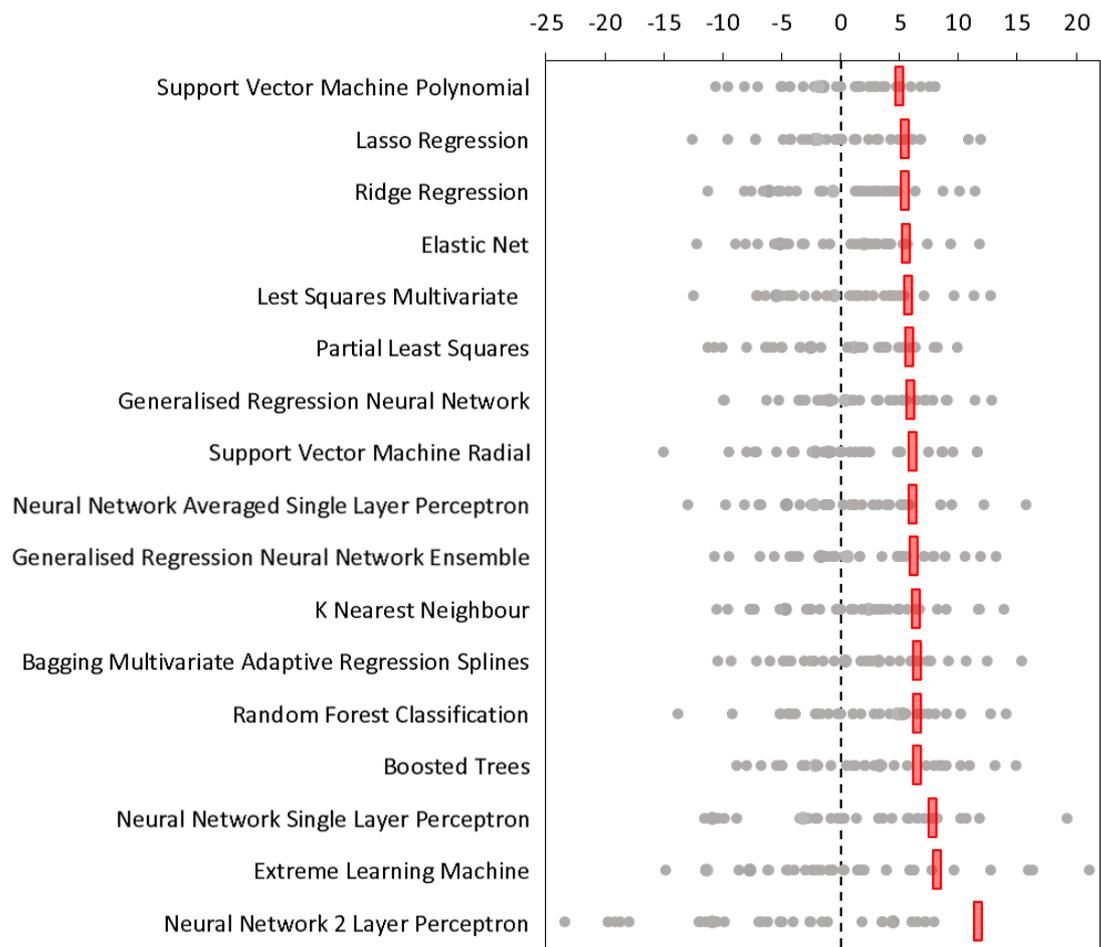


Figure 4 - Combined graph of the residuals (grey) and RMSEs (red) for the blind test set (n=33) for the different statistical models.

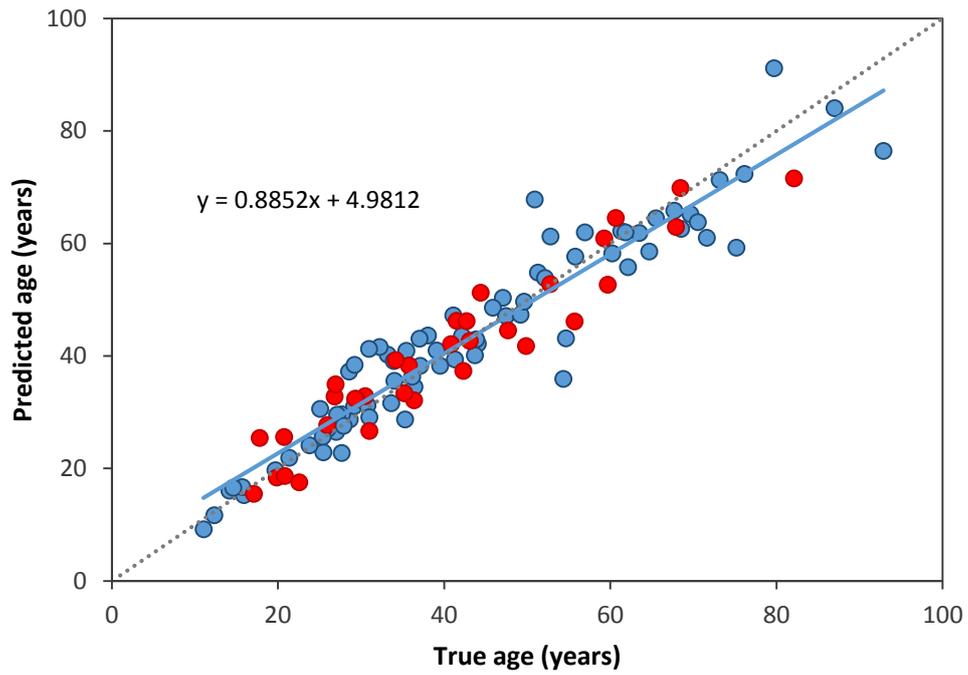


Figure 5 - Comparison between the predicted and the true age for the training (blue, n=76) and blind test set (red, n=33) in the SVMp model. The mean absolute prediction error was calculated at 4.0 and 4.1 years respectively.

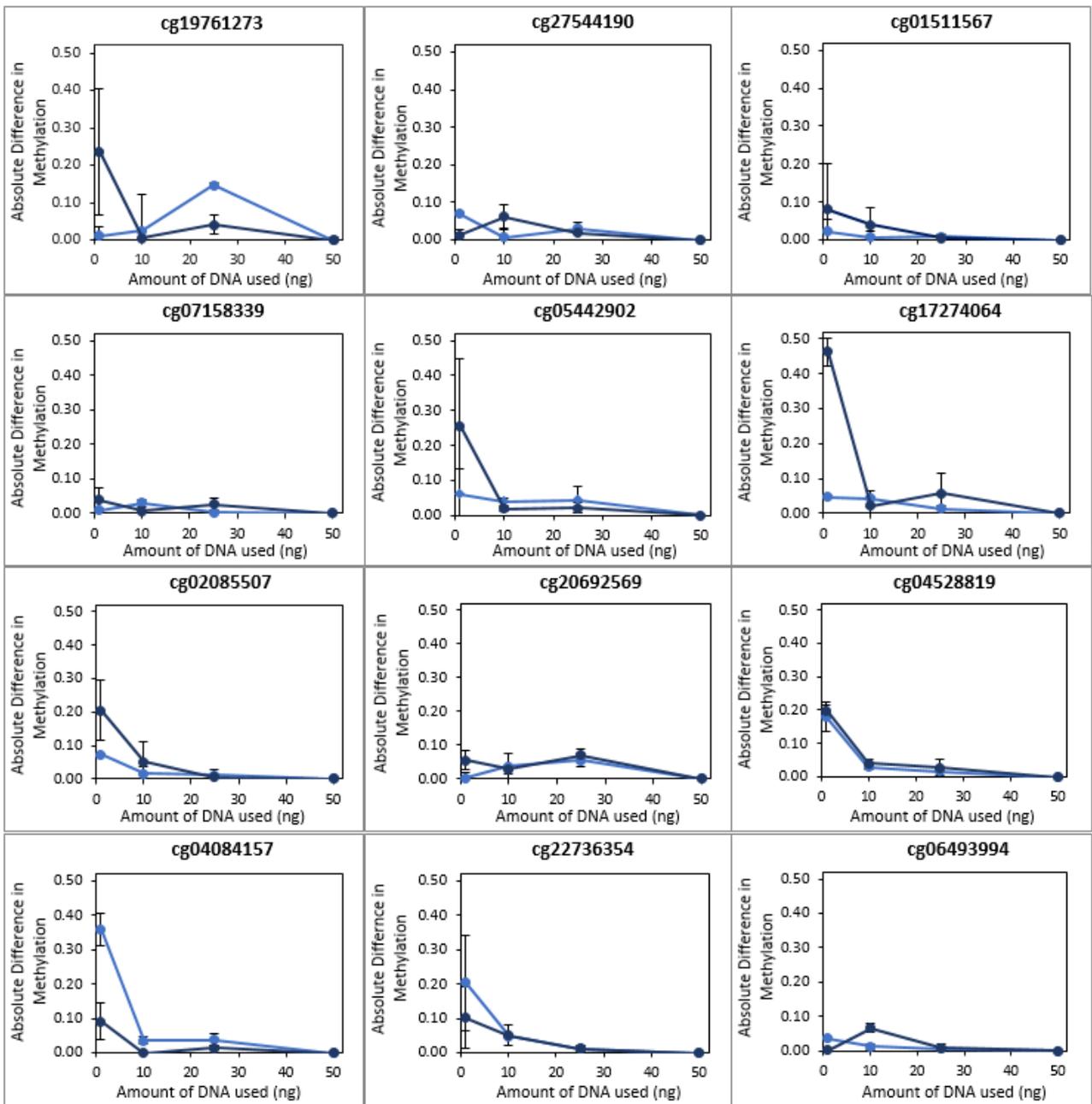


Figure 6 - Amount of initial DNA input (ng) and the absolute difference in the quantification of DNA methylation from the 50ng input (optimum), for all 12 CpG sites. Two different pre-mixed methylation standards were used in this assessment, the first corresponding to 5% (light blue) and the second to 25% methylation (dark blue).

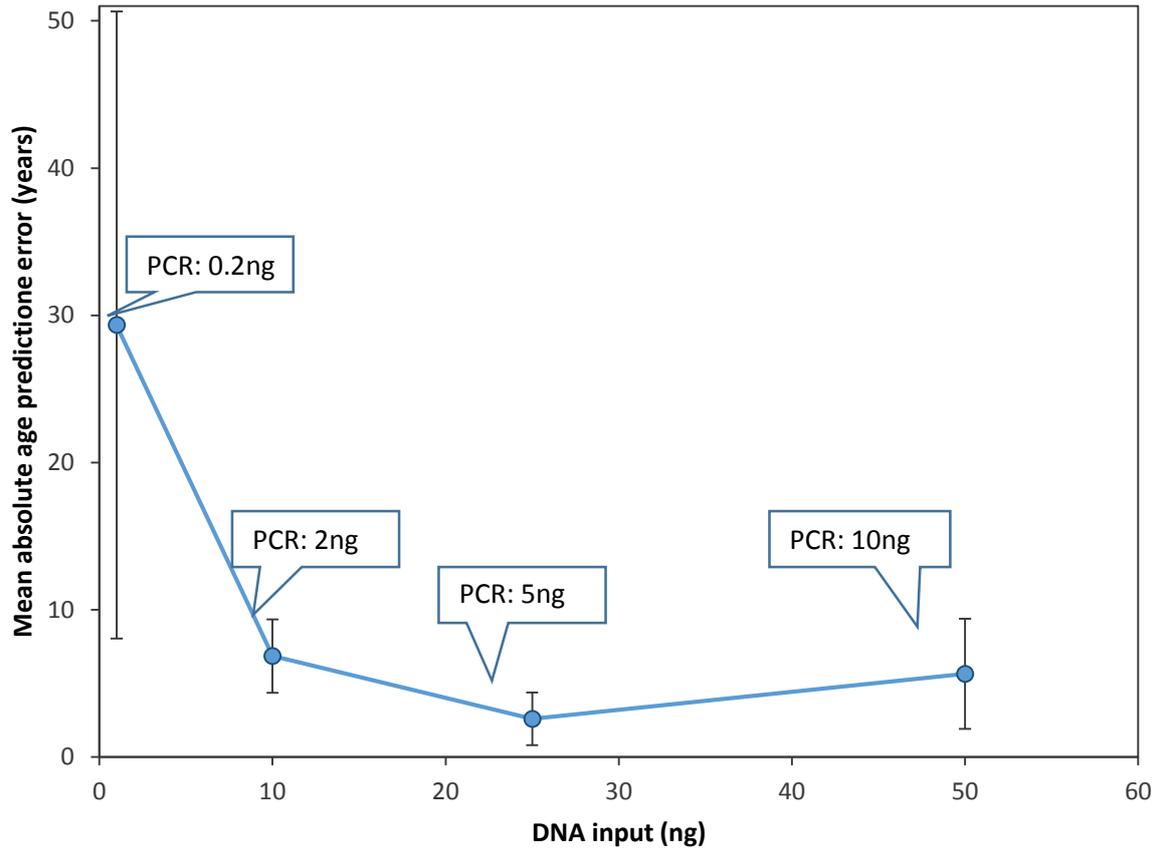


Figure 7 - Amount of initial DNA input (ng) (x axis) and MAE for age prediction for 6 blood samples analysed in duplicate. The estimated DNA input (ng) in the PCR stage (post-bisulphite treatment) is also depicted in the graph.

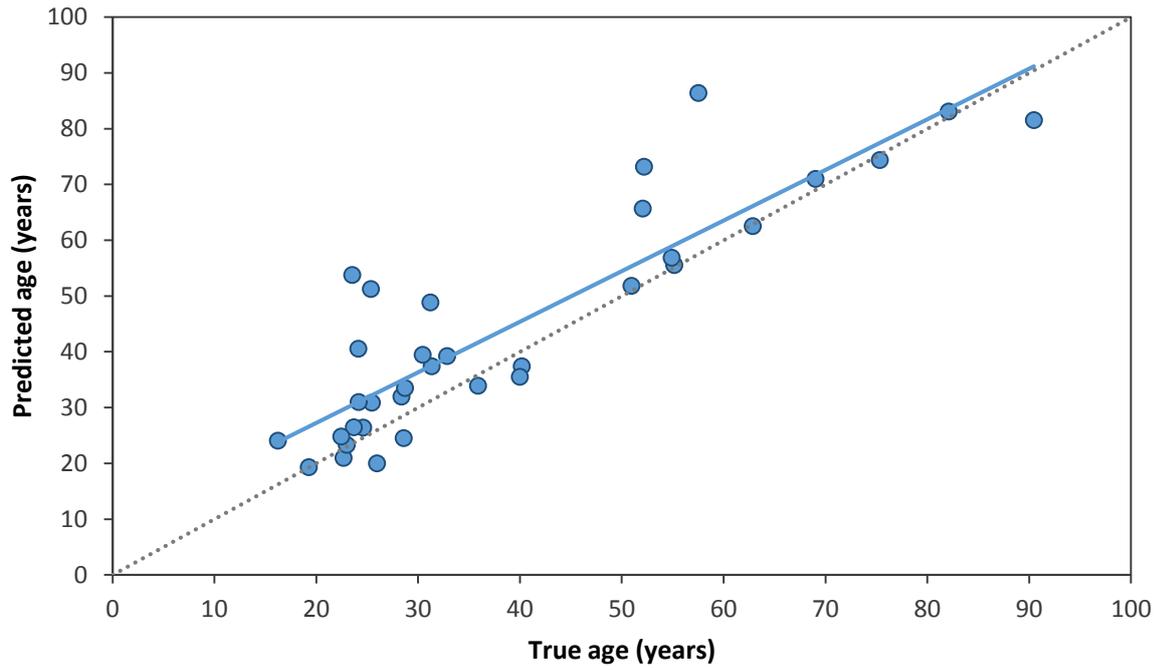


Figure 8 - Comparison between the predicted and the true age for 34 saliva samples using the age prediction model developed in whole blood using Support Vector Machines with polynomial kernel function (SVMp). The mean absolute prediction error was calculated at 7.3 years and the root mean square error at 11.1 years.