Forensic Science International: Genetics 31 (2017) 19–28

Contents lists available at ScienceDirect

Forensic Science International: Genetics

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

Research paper

Chronological age prediction based on DNA methylation: Massive parallel sequencing and random forest regression

Jana Naue[a,*], Huub C.J. Hoefsloot[a], Olaf R.F. Mook[b], Laura Rijlaarsdam-Hoekstra[a], Marloes C.H. van der Zwalm[a], Peter Henneman[b], Ate D. Kloosterman[c,d,1], Pernette J. Verschure[a,*,1]

[a] University of Amsterdam, Swammerdam Institute for Life Sciences, Science Park 904, 1098XH Amsterdam, The Netherlands
[b] Amsterdam Medical Center, Clinical Genetics, Meibergdreef 9, 1105AZ, Amsterdam, The Netherlands
[c] Netherlands Forensic Institute, Biological Traces, Laan van Ypenburg 6, 2497GB Den Haag, The Netherlands
[d] University of Amsterdam, Institute for Biodiversity and Dynamics, Science Park 904, 1098XH Amsterdam, The Netherlands

ARTICLE INFO

Article history:
Received 10 July 2017
Received in revised form 26 July 2017
Accepted 30 July 2017
Available online 1 August 2017

Keywords:
Age prediction
DNA methylation
Massive parallel sequencing
Machine learning

ABSTRACT

The use of DNA methylation (DNAm) to obtain additional information in forensic investigations showed to be a promising and increasing field of interest. Prediction of the chronological age based on age-dependent changes in the DNAm of specific CpG sites within the genome is one such potential application. Here we present an age-prediction tool for whole blood based on massive parallel sequencing (MPS) and a random forest machine learning algorithm. MPS allows accurate DNAm determination of pre-selected markers and neighboring CpG-sites to identify the best age-predictive markers for the age-prediction tool. 15 age-dependent markers of different loci were initially chosen based on publicly available 450K microarray data, and 13 finally selected for the age tool based on MPS (DDO, ELOVL2, F5, GRM2, HOXC4, KLF14, LDB2, MEIS1-AS3, NKIRAS2, RPA2, SAMD10, TRIM59, ZYG11A). Whole blood samples of 208 individuals were used for training of the algorithm and a further 104 individuals were used for model evaluation (age 18–69). In the case of KLF14, LDB2, SAMD10, and GRM2, neighboring CpG sites and not the initial 450K sites were chosen for the final model. Cross-validation of the training set leads to a mean absolute deviation (MAD) of 3.21 years and a root-mean square error (RMSE) of 3.97 years. Evaluation of model performance using the test set showed a comparable result (MAD 3.16 years, RMSE 3.93 years). A reduced model based on only the top 4 markers (ELOVL2, F5, KLF14, and TRIM59) resulted in a RMSE of 4.19 years and MAD of 3.24 years for the test set (cross validation training set: RMSE 4.63 years, MAD 3.64 years). The amplified region was additionally investigated for occurrence of SNPs in case of an aberrant DNAm result, which in some cases can be an indication for a deviation in DNAm.

Our approach uncovered well-known DNAm age-dependent markers, as well as additional new age-dependent sites for improvement of the model, and allowed the creation of a reliable and accurate epigenetic tool for age-prediction without restriction to a linear change in DNAm with age.

1. Introduction

For a long time, DNA methylation (DNAm) was a “black spot” for forensic scientists. Over the years it is recognized that DNAm analysis can give additional forensic relevant information in parallel to the DNA profile [1,2]. DNAm, which mostly occurs in a CpG sequence context represents an epigenetic mark and plays an important role in cell regulation to establish and maintain cell identity [3–5]. These functions explain why DNAm at specific sites in the genome shows a cell-type specific pattern and how it can be used for tissue/body-fluid discrimination [6–8]. Furthermore, it has been revealed that DNAm alters with age within each individual [9–12]. This alteration in DNAm has been shown to occur at specific CpG sites in all individuals however with individual differences in “speed”, showing more DNAm differences in aged twins compared to young ones [13,14].

Within the last years, epigenetic age-dependent sites were identified in multiple studies [9–11,15] and assays for forensic

* Corresponding authors.
E-mail addresses: j.naue@uva.nl (J. Naue), p.j.verschure@uva.nl (P.J. Verschure).
[1] These authors contributed equally to the work.

http://dx.doi.org/10.1016/j.fsigen.2017.07.015

purposes were developed to predict the age of an individual as an investigative lead [16–24]. Most of these studies concentrated on marker selection based on previously published markers or used datasets containing DNAm values from defined CpG sites on the 27K or 450K microarrays for marker selection. Different models for age prediction such as multivariate linear regression models [24] or machine learning algorithms, such as support vector machine (SMV) [23] and artificial neural networks (ANN) [18], were applied. Model performance, is mostly described using the mean absolute deviation (MAD) or the root mean square error (RMSE). Previous forensic studies using blood samples gave MAD values of around 3.5–4.5 years for independent test sets [17,18,20,22,25]. The number of markers varied between 3 and 20, and the covered age range was not identical. Determination of DNAm was in most studies done either by pyrosequencing [20,26] or SNaPshot [24,27] analysis, representing common methods in a forensic laboratory. Massive parallel sequencing (MPS) is another promising method, allowing multiplexed sequencing and accurate determination of DNAm. Vidaki et al. also applied MPS analyzing blood samples, resulting in a MAD of 7.45 years [18].

Within our study, we present a complete MPS-based assay consisting of a training set composed of 208 individuals and an independent test set of 104 individuals, investigating the age range of 18–69 years. Furthermore, marker selection was done from "scratch" using only publicly available 450K datasets that included the pure raw data and not only pre-processed data based on different analysis packages. Our uniform analysis pipeline for all datasets allowed for a better quality, normalization and therefore batch control. The machine learning algorithm Random Forest regression (RFR) was used for marker selection and to build the model. RFR is an ensemble tool based on decision trees [28]. A defined number of trees using selected features (in our case the CpG marker) leads to an age prediction based on the average estimation of each tree. Random sub-setting of the samples within the training set for construction of each tree avoids overfitting to the training set samples. Special attention was also payed to neighboring CpG sites not covered by the 450K microarray and aberrant DNAm results, and age-prediction outliers, were carefully analyzed.

2. Material and methods

A workflow over the different parts is given in Suppl. Fig. S1.

2.1. Marker selection using public data

2.1.1. Datasets and quality control

Datasets containing DNAm levels measured with Infinium 450 K Human Methylation Beadchip (Illumina) (so called beta values) were selected from the Gene Expression Omnibus Database (GEO) (Suppl. Table S1). The search was restricted to datasets with available raw data files *(.idat)* and additional information on age and gender. In some cases, the age was obtained after author contact. Only control samples of healthy individuals were considered in case of disease versus control studies. Analysis of the 450 K data was done using R v3.2.3. All samples from the datasets went through a quality control pipeline using the *MethylAid* package v1.4.0 and standard settings [29]. All quality checks, sample-dependent (e.g. bisulfite conversion efficiency) as well as sample-independent (e.g. low background noise), needed to be passed for inclusion of each sample for further analysis.

2.1.2. Pre-processing

Subsequently, the microarray results of all samples were normalized to remove technical variation between measurements. This was done with the *funnorm* normalization approach incorporated in the *minfi* R package v1.16.0 [30] which is based on control probes analyzed for each sample avoiding normalization on biological effects [31]. Prior to further analysis, a high amount of the 450 K probes was removed due to different reasons: 1. Probes containing SNPs according to the SNPdb 137 within the probe sequence or the extension site could lead to a deviation of the measured methylation level and removal was done using the *minfi* package. 2. Probes hybridizing to the X or Y chromosome were not considered as the aim was to find non-gender dependent markers, 3. Probes reported to cross-hybridize were removed as they could lead to unspecific methylation levels [32], 4. Jaffe and Irizarry showed that some CpG sites can show a blood sub-cell type specific methylation pattern [33]. Sites reported with statistically significant differences ($p < 0.05$) between cell types that could have an effect were removed to avoid measurement of the change in cell composition in blood due to age rather than a direct correlation of DNAm and age, 5. Only probes with a p-detection value (<0.01) were kept to rely further analysis only on very high quality signals. The beta value of a single probe within a sample was otherwise removed. Probes failing in over 20% of samples were excluded, including only CpG sites for the analysis whose DNAm is represented in enough samples.

2.1.3. Marker selection

Due to the large number of probes on the 450K array compared to the restricted number of samples available, a pre-selection of features was done using 10% of the public data to limit the CpG sites combining the top features obtained using RFR, linear coefficient analysis, and mutual information, respectively. A validation scheme based on Wessels et al. (2005) was used for further marker selection using the RFR approach and based on the other 90% of the training set resulting in a list of the most promising markers for each dataset included [34]. The use of this validation scheme reduces the risk of over-fitting by repeated random splitting of the samples for training and testing. RFR is implemented in the *RandomForest* v.4.6 R package.

2.2. Experimental set-up for analysis of DNAm with MPS

2.2.1. Samples

5–10 ml whole blood from 324 individuals from 18 to 69 years were obtained from Sanquin (Dutch blood bank) paying attention to an equal gender and age distribution. Individuals gave written consent to Sanquin, allowing the use of remaining blood routinely collected for disease screening purposes for the use in research. Twelve samples were used for preliminary experiments. The other 312 samples were randomized (to exclude collection effects) and divided into a training set (208 samples) and test set (104 samples). DNA from 200 to 400 μl of blood was extracted using the QiaAmp Blood mini kit (Qiagen, Hilden, Germany) according to manufacturer's recommendation. DNA amount was measured using the Nanodrop 2000 (Thermo Fisher Scientific (TFS), MA, USA).

2.2.2. Bisulfite conversion, PCR and MPS

300 ng of DNA was bisulfite converted using the Gold Kit (Zymo Research, CA, USA). SsDNA amount was measured using the ssDNA Qubit Quantification Kit (TFS). Each marker was amplified in a 6.5 μl reaction volume. The reaction mix consisted of 2.5 μl HotStarTaq Master Mix Kit (Qiagen), 2.5 pmol forward and reverse primer each (Suppl. Table S2), 0.5 μg BSA (TFS), and DNA-free H_2O ad 4 μl. 2.5 μl (10 ng) DNA was added to each reaction. A touch-up PCR was run under the following conditions: 95 °C 10 min; 15 cycles: 98 °C 45 s, 54 °C 30 s, 72 °C 30 s; 25 cycles: 98 °C 45 s, 62 °C 30 s, 72 °C 30 s; final elongation at 72 °C for 10 min. Technical bias within the PCR due to a CpG site in the primer was reduced using a wobble site in the primer. Furthermore, the primer region was

checked for known SNPs using SNPcheck v3 (www.snpcheck.net). Amplicons per individual were pooled depending on PCR efficiency. PCR products were cleaned with 1.9 x magnetic beads (GE Healthcare, Little Chalfont, UK) prepared according to [35]. 6.5 μl of pooled sample was added to 27 μl of Platinum® PCR SuperMix High Fidelity (TFS) and 5 pmol of each index-primer (dual NexteraXT indexing). The run consisted of 95 °C 1 min, 72 °C 5 min; 6 cycles: 95 °C 30 s, 62 °C 2 min, 72 °C 2 min; final elongation at 72 °C for 3 min. PCR products were cleaned twice with 1.6 x beads (GE healthcare). The 2100 Bioanalyzer using the DNA 1000 kit (both Agilent Genomics, CA, USA) was used to check the quality of the amplicon pool of each individual and to calculate the mean base pair value for equimolar pooling. PCR products were measured using the dsDNA high sensitivity Qubit Quantification Kit (TFS), and equimolar pooled. The final DNA pool was diluted to 4–7 nM and 600 μl of a final 8–14 pM dilution (dependent on machine and run) was sequenced on a MiSeq using the 2 × 150 bp v2 Kit (Illumina, CA, USA).

2.2.3. MPS data analysis

The obtained FastQ files were 5′ and 3′ trimmed, and quality checked using *TrimGalore* v0.4.3 (having the FastQC package included) [36]. Paired-end reads were joined using *PEAR* v0.9.10 [37] (removing overlapping read ends). Reads were aligned to the human reference hg19 via *samtools* implemented in the *biscuit* v0.2.0 package pipeline and CpG DNAm values extracted using *MethylDackel* v0.2.1 [38–40] (quality threshold set to 20 for phred and read quality). For the final DNAm analysis, the obtained (un-) methylated read counts were used to calculate the DNAm level allowing the adding of further reads from another run. Initially, we aimed for a coverage of 800 x. DNAm results with lower coverage were not directly excluded, but the binomial formula was used to calculate the 95% confidence interval for 50% of DNAm (highest variation) as done by Masser et al. [41] to evaluate the accuracy of the obtained DNAm for further decision.

The Non-CpG (i.e. CHH and CHG) methylation levels were extracted and the mean non-CpG DNAm per marker calculated. Afterwards, the conversion efficiency of one sample was calculated by taking 100% minus the mean non-CpG DNAm.

Furthermore, SNPs were extracted from the MPS results using the SNP-extraction function of biscuit v 2.2.

2.2.4. Sanger sequencing for additional SNP analysis

Bisulfite converted DNA gives only a limited possibility to check for SNPs within the amplified region due to the conversion of C > T. Non-converted DNA was analyzed with another set of primers surrounding the PCR fragment of the MPS approach in case of unexpected DNAm results. For primers and detailed run conditions see Suppl. Table S2. PCR products were cleaned with 10U Exonuclease I and 1U thermo-sensitive Alkaline Phosphatase (FastAP) (both Thermo Fisher Scientific) for 30 min at 37 °C and enzyme-inactivated for 15 min at 85 °C. Cleaned PCR products were sent for sequencing (Eurofins, Germany).

2.3. RFR model construction using the MPS data of the training set

The obtained DNAm levels of the training set and the corresponding chronological age of the individuals were used for optimization and model training.

2.3.1. Final model optimization

The Spearman correlation coefficient (SciPy v0.19.0, python3 v3.5) was calculated for all 95 age-dependent CpGs sites. The CpG site showing the highest correlation with age per locus was chosen for the final model. For optimization of the RFR model, two important parameters were tuned within R. The number of features available to be considered at each split (*mtry*) and the minimal *nodesize* (limiting how often a tree can be splitted). A *mtry* range from 3 to 6 and a *nodesize* from 1 to 8 was considered. A final *mtry* of 4 and minimal *nodesize* of 2 was chosen and the number of trees was set to 1000.

2.3.2. Cross-validation of the training set for model performance evaluation

To measure the initial performance of the model, the RMSE +/− 95% CI and MAD +/− 95% CI of ten times repeated 5-fold cross validation of the training set data were calculated. Using repeated cross-validation allows shuffling and therefore excludes fold-specific differences.

2.4. Prediction of the test set

After finalizing the model, the ages of the test set individuals (104 samples) were predicted using a trained RFR model based on all 208 samples of the training set. To determine the performance, the RMSE and MAD from the chronological age were calculated for the test set.

Grouping into age groups was done either on the chronological age (three equal large age groups) or the predicted age to evaluate performance within groups. Further analysis to investigate the difference between the gender was done using the Wilcoxon test for the DNAm and prediction (scipy.stats, python).

3. Results and discussion

3.1. Analysis of public data

Eight GSE datasets containing raw data were selected from the GEO database after an initial screening and quality filtering. However, these datasets show a difference in blood cell-type composition (4 times whole blood (394), 4 times buffy coat (852)) as well as a very wide range of age composition (with an overrepresentation of older people). An overview of the GSE characteristics is given in Suppl. Table S1. To exclude a bias caused by an unequal age distribution or a mixture of different blood cell type amounts, different datasets (only whole blood, only buffy coat, without cord blood) were created and run with the marker selection pipeline. As whole blood represents the most important tissue for forensics, we also created a dataset with a maximum number of samples per age to compensate for unequal age distributions within the original datasets. The different datasets with a subset of the whole blood dataset were generated to allow for a better understanding how different blood cell types or an overrepresentation of old individuals could influence the set of selected markers. The most promising 15 loci per dataset with the highest feature importance obtained from the marker selection pipeline using Random Forest regression are shown in Suppl. Table S3. The top six age-dependent loci per dataset, which were at least also once identified in one of the other datasets, were selected. However only one 450K probe per locus was considered. As MPS allows to combine a lot of markers, EIF1 and RPA2 were also included since they appeared promising although not within the top six markers per dataset. Finally, 15 loci were chosen for analysis using massive parallel sequencing (MPS): cg16867657 (ELOVL2), cg12934382 (GRM2), cg11807280 (MEIS1-AS3), cg02872426 (DDO), cg06874016 (NKIRAS2), cg08097417 (KLF14), cg06784991 (ZYG11A), cg07553761 (TRIM59), cg16054275 (F5), cg13959344, cg03224418 (SAMD10), cg08262002 (LDB2), cg25410668 (RPA2), and cg22156456 (EIF1). Furthermore, an age-independent marker from the 450 K data, that showed a low interquantile range over the whole age range, was considered for control purposes: cg10007452 (PLAGL1).

Table 1
Final MPS marker of the age-prediction assay. Overview about location (hg19/GRCh37.p13), gene name and age tendency. In most cases, multiple CpG sites are covered in a MPS analysis. Choosing the best CpG site per marker led to an overlap with the 450 K sites in nine of the 13 markers. Chr.: chromosome.

MPS marker	Gene	Full gene name	Chr.	Position CpG (GRCh37.p13)	Selected 450 K marker	Total CpG sites (CpG of interest)[b]	DNAm change with age
DDO_1	DDO	D-aspartate oxidase	6	110736772	cg02872426	1 CpG (1)	down
ELOVL2_6	ELOVL2	ELOVL fatty acid elongase 2	6	11044877	cg16867657	16 CpGs (6)	up
F5_2	F5	Homo sapiens coagulation factor V	1	169556022	cg16054275	3 CpGs (2)	down
GRM2_9	GRM2	Glutamate receptor, metabotropic 2	3	51741152	–	17 CpGs (9)	up
HOXC4_1	HOXC4	Homeobox C4	12	54448265	cg18473521	3 CpGs (1)	up
KLF14_2	KLF14	Kruppel-like factor 14	7	130419118	–	6 CpGs (2)	up
LDB2_3	LDB2	LIM domain binding 2	4	16575420	–	3 CpGs (3)	down
NKIRAS2_2	NKIRAS2	NFKB inhibitor interacting Ras-like 2	17	40177415	cg06874016	3 CpGs (2)	down
RPA2_3	RPA2	Replication protein A2	1	28241577	cg25410668	4 CpGs (3)	up
SAMD10_2	SAMD10	Sterile alpha motif domain containing 10	20	62611844	–	6 CpGs (2)	down
TRIM59_5	TRIM59	Tripartite motif containing 59	3	160167977	cg07553761	11 CpGs (5)	up
MEIS1_1	MEIS1-AS3[a]	MEIS1 antisense RNA 3	2	66654644	cg11807280	1 CpG (1)	down
ZYG11A_4	ZYG11A	Zyg-11 family member A	1	53308768	cg06784991	21 CpGs (4)	up

[a] according to EPIC Infinum array, gene association not yet in the 450 K manifest.
[b] CpGs with C inside the primer region not counted; counting started in positive strand orientation.

3.2. Massive parallel sequencing results of the training set

3.2.1. Final marker and CpG site choice

The 15 age-dependent markers and an age-independent marker were analyzed in the 208 samples of the MPS analyzed training set. In the final assay, two age-dependent markers were excluded due to inefficient amplification (cg22156456 (EIF1)) and a too low correlation with age (Spearman's rho: 0.57) for the investigated age range of 18–69 years (cg13959344), respectively. Analysis of PLAGL1 (cg10007452 = CpG site 14) as a constant marker did not help to improve the model or detection of outliers and this marker was not further considered to create the final model.

We calculated the Spearman correlation coefficient between age and each CpG site for the training set samples to select the most age-dependent CpG site for each amplicon of the 13 loci. No alternative sites existed in case of DDO and MEIS1-AS3 as only one site is covered by the amplicon (DDO_1 and MEIS1_1). For nine loci, the most age-dependent CpG site overlaps with the original selected 450 K site (Suppl. Fig. S2). In case of four loci a neighboring CpG site led to a slightly higher Spearman correlation (SAMD10_2: −0.68 vs. SAMD10_3 (cg03224418): −0.67, GRM2_9: 0.75 vs. GRM2_7 (cg12934382): 0.74, LDB2_3: −0.72 vs. LDB2_1 (cg08262002): −0.69, and KLF14_2: 0.83 vs. KLF14_3: 0.81 (cg08097417)). The final marker list is given in Table 1 (cf. Suppl. Table S2 for additional information on amplicon length and analyzed strand) and the DNAm results of the training set are shown in Fig. 1 and Suppl. Table S4a. Some CpG sites show a stronger correlation with age than others, however RFR is using also weak markers to create a strong model. Furthermore, addition of more markers could be advantageous especially in the case of DNAm outliers.

3.2.2. Role of the age markers

The 13 annotated genes code for proteins or an antisense RNA involved in multiple pathways (analyzed using www.reactome.org,www.uniprot.org and the gene database of www.ncbi.nlm.nih.gov): Developmental biology (HOXC4, MEIS1), metabolism (ELOVL2: fatty acids, DDO: amino acids and derivates), hemostasis

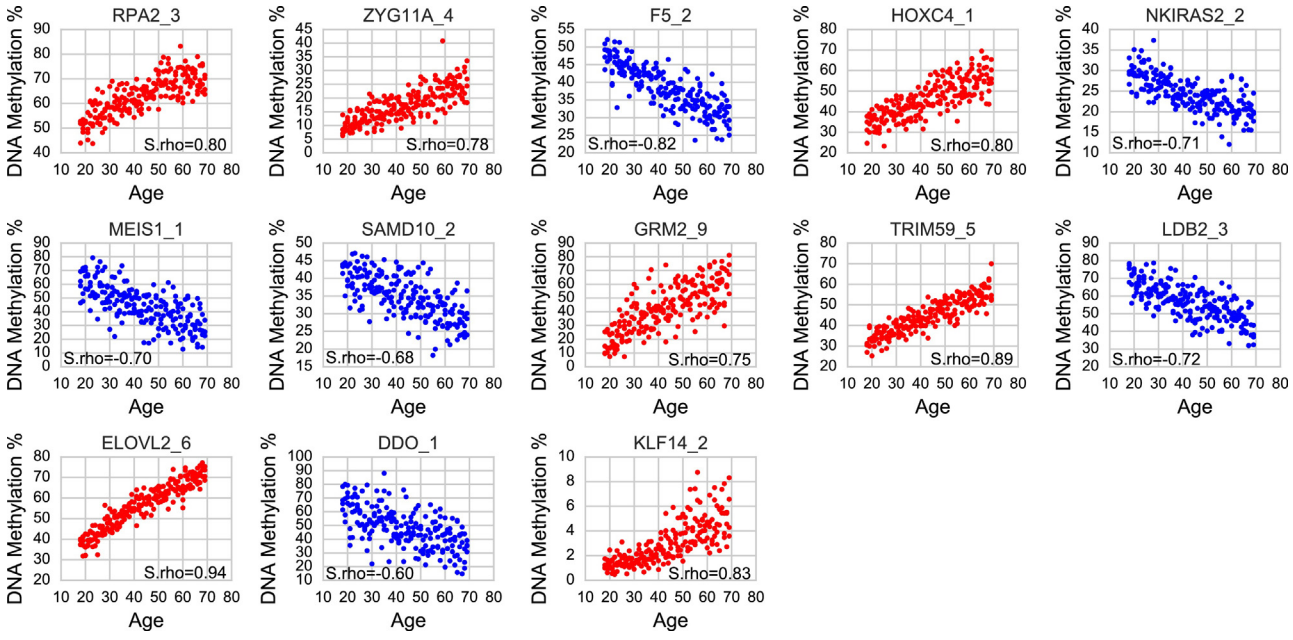

Fig. 1. Final age-dependent markers. DNAm results for the 208 samples of the training set are plotted. The DNAm of seven markers show an increase with age (red) and a decrease for six markers (blue). SAMD10_2, GRM2_9, LDB2_3, and KLF14_2 are neighboring CpG sites of the original 450K site. Spearman coefficient is provided (S.rho: Spearman's rho). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

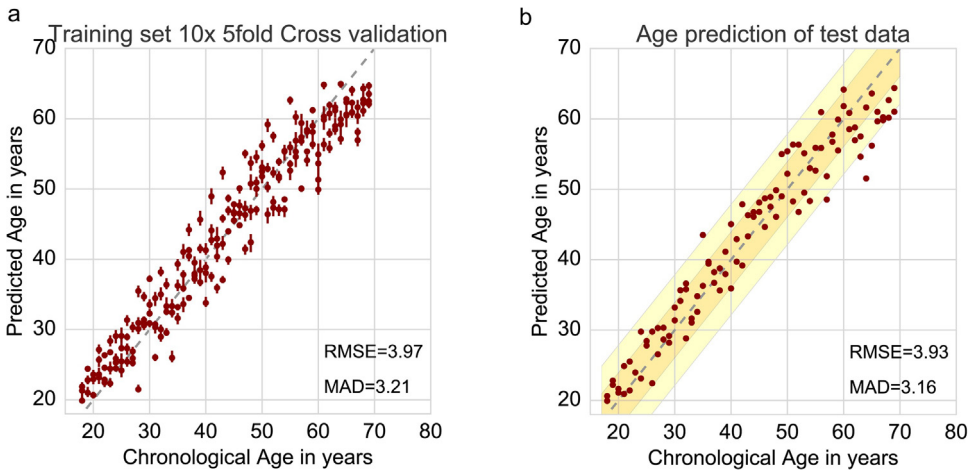

Fig. 2. Age-prediction using the final 13 markers. a) Cross-Validation of the training set. b) Test dataset predicted with the model trained on all 208 training set samples. The orange and yellow range represent the RMSE (70.19% of data) and 2xRMSE (94.2%) of the test set. The maximum deviation from the chronological age observed was 12.5 years. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

and the metabolism of proteins (F5), the immune system (NKIRAS2, TRIM59), transcription (RPA2, KLF14, LDB2) and DNA repair, replication, cell cycle and cellular response to external stimuli (RPA2). No specific function was found for SAMD10. The functions of the markers do not play a direct role when applying the assay, but should be considered to explain outliers, especially as the health background of unknown individuals is expected to be unknown or undiagnosed when analyzing living individuals.

3.2.3. Random forest model creation and evaluation of the training set

To obtain a first generalized indication of RFR performance of the training set, a 10-times repeated 5 x cross validation (CV) was performed leading to a mean RMSE of 3.97 years (95% CI: 3.93–4.06) and MAD of 3.21 years (95% CI: 3.17–3.27) (Fig. 2a). The model was in each fold of the 5×CV build on 80% of the whole training set, and the other 20% of the samples were held back for testing. The predictions within each repetition are quite close to each other, showing the stability of the RFR. The final model for age prediction of the test set was constructed including all 208 samples in one RFR model.

3.3. Model evaluation using an independent test set

3.3.1. MPS analysis of test set

104 samples were independently investigated to test model performance. The same range as for the training set was used (18–69) with two individuals per year (1 female, 1 male). All DNAm results are given in Suppl. Table S4a.

3.3.2. Age prediction using the test set

The chronological age of the 104 individuals was predicted using the final model based on all 208 train samples resulting in a RMSE of 3.93 years and MAD of 3.16 years, confirming the obtained result of the cross-validation (Fig. 2b). Comparing the results of the CV of the training set with the test shows that the model evaluation of the training set is robust and that the independent analysis of test data fell within the expected range of variation. The predicted results and prediction error of the test set are provided in Suppl. Table 4b.

The RMSE is an appropriate measurement for model performance and can be used for the interval forecast. Hong et al. used the RMSE and 2xRMSE for a prediction model of saliva [24]. In case of a normal distribution, around 68% and 95% of samples are expected to be predicted within a RMSE and 2 x RMSE range. From our test set 70.19% of the samples resulted in an error within the RMSE range of the test set (+/− 3.93 years) and 94.2% within 2 x RMSE range (+/− 7.86 years).

3.3.3. Deviation per age groups

We grouped the samples into three age categories, to have a closer look how the accuracy of the prediction depends on the actual chronological age. Older individuals showed an increased deviation, and their age is rather underpredicted, whereas the age prediction of young individuals resulted in slight overestimation (Fig. 3). The high deviation in older individuals is most likely caused by the general known increased inter-individual variation between older people compared to young individuals possibly due to epigenetic drift [11]. Furthermore, it needs to be considered that the model was built covering the specific age range 18–69 years, and that no information about marker behavior outside this range is available, therefore no prediction outside this range will occur and this can also lead to less accurate estimates for samples at the age limit border.

We calculated how many samples per age group fall within the RMSE and 2 x RMSE obtained with the test set (Table 2, upper 3

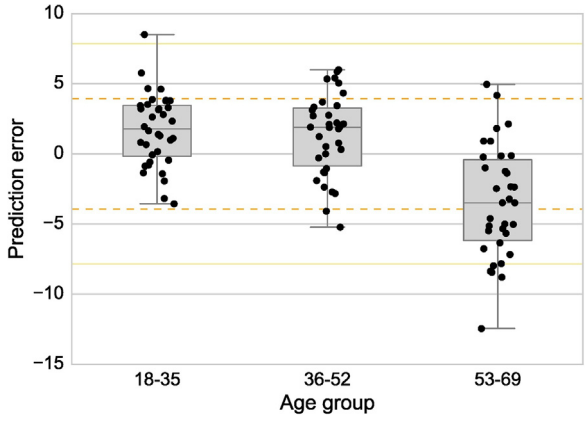

Fig. 3. Deviation of the prediction from the chronological age. The number of samples with prediction errors is plotted for all 104 samples grouped into three age groups of approximately the same size (34–36). The individuals of the young age group are rather overestimated in comparison to older individuals that tend to be underestimated more often. The two lines show the RMSE (orange, dashed) and 2x RMSE range (yellow). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Table 2
Age prediction of samples for a specific age group (top: grouping according to chronological age, bottom: grouping according to predicted age). Number of samples falling within the RMSE and 2×RMSE of the test set was calculated. Additionally, the group-specific RMSE was calculated. To get an indication of the distribution, the maximal absolute deviation is also given. chr: chronological age, pred: predicted age.

Age group	Number of samples	% Samples within RMSE (3.93 y)	% Samples within 2xRMSE (7.86 y)	Group-specific RMSE	Maximal absolute deviation
18–69 (all)	104	70.19	94.2	3.93 (overall RMSE)	12.5
18–35 (chr)	36	88.89	97.2	3.06	8.5
36–52 (chr)	34	76.47	100	3.15	6.0
53–69 (chr)	34	47.06	85.29	5.24	12.5
18–35 (pred)	35	91.43	100	2.75	5.8
36–52 (pred)	37	75.68	91.89	4.03	12.5
53–69 (pred)	32	40.00	90.63	4.82	8.8

rows). The age prediction of younger individuals is superior to older individuals, which is in concordance with other studies [20,25,41]. The chronological age of younger individuals can be predicted with high confidence within the 2xRMSE range of the test set. The group specific RMSE values are also below the general RMSE of 3.93 years. The grouping of individuals to perform these calculations was based on the chronological age of the individuals. However, to get a better age-indication of the "unknown", the same calculations were done based on the obtained prediction value (counting the age of predicted full years of age) (Table 2, lower 3 rows). The lower group-specific RMSE for the age groups 18–35 and 53–69, and the higher RMSE for 36–52 could be caused by the two strongest outliers (cf. Fig. 2b) as these samples are now grouped into the age range 36–52 instead of 18–35 (S253: age: 35, predicted: 43.7) and 53–69 (S306: age: 64, predicted: 51.4). These results show that outliers that occur in the test set can have an impact on age-group specific calculations.

Additionally, age-group specific RFR models were created as the modelling on a smaller age range could improve the model and a-priori knowledge about the age of an unknown sample could then be incorporated. The age group used for training included an additional +/− 10 years compared to the range of the test set to allow a broader age range for prediction. For the age group 18–35 only +10 years and for the group 53–69 only −10 years was possible (Table 3). Limiting the age range of the RFR model slightly improved the prediction accuracy, even though fewer samples are available to train the model. In the case of limiting the age range, the maximum deviation is also lower compared to testing the RFR model based on the full age range. That can be due to better prediction and/or limitation of prediction outcome due to the age range of the training set.

3.4. Conditions that effect DNAm and age prediction

3.4.1. Bisulfite conversion efficiency

Bisulfite conversion is a crucial step, as a low conversion efficiency can lead to a bias of the results (leading to DNAm overestimation due to failed C to U conversion). We noted that the mean conversion efficiency for the train- and test set was 99% (min = 98.5% (training set) and 97% (test set)). We calculated the mean conversion efficiency per marker to exclude marker-specific differences in bisulfite conversion efficiency. All markers showed the same mean conversion efficiency between 98.8% – 99.2% for the train and test set. These findings give no indication that the DNA sequence has an impact on the bisulfite conversion efficiency. Two samples in the test set showed with 97% and 97.6% a lower conversion efficiency compared to the other samples, especially for the markers SAMD10 (S315: 92%, S316: 96%) and KLF14 (S315: 95%, KLF14: 96%). Nevertheless, a deviation between chronological age and predicted age was only observed for S316 (error of 5.3 years) but not for S315 (error of 0.5 years). Therefore, deviation of the age prediction could be due to other reasons and the effect of decreased bisulfite conversion of each marker could also have a different strong effect on the age prediction.

3.4.2. Gender effect

We analyzed if a gender-dependent difference between DNAm at the selected CpG sites exists. Two female-male pairs were created for the training set per year of age showing statistically significant differences (Wilcoxon test, $p < 0.05$) in ELOVL2_6, GRM2_9, KLF14_2, HOXC4_1, LDB2_3, RPA2_3, SAMD10_2, TRIM59_5, and MEIS1_1. However, no statistically significant difference was observed between the absolute prediction error when pairing female and male samples of the test set per age (Wilcoxon test, $p = 0.63$ (test set)) or the per se prediction error, (Wilcoxon test, $p = 0.15$). Nevertheless, a tendency to slightly overestimate the age of males was observed in the test set (Suppl. Fig. S3). Neither the inclusion of the gender variable nor the creation of two separate models improved the age prediction significantly. The observation that differences between gender exist but that the model is not improved based on gender effects, was also made by Zbieć-Piekarska [20]. Eventually, the observed differences of DNAm possibly due to gender effects did not have an impact on the accuracy of our model prediction. An increased number of samples per gender or other mathematical approaches could perhaps make a model improvement more evident by inclusion of gender.

3.4.3. SNP analysis

It is known that the occurrence of SNPs can have an impact on DNAm [42]. Especially loss or gain of a CpG site would alter the measured DNAm [32] at that site and could also lead to a changed

Table 3
Age prediction of samples for age-group specific models. The RMSE for each model was calculated based on the test samples. Number of samples falling within the RMSE and 2xRMSE, and the maximal absolute deviation is provided for each model. Predictions were also evaluated based on classification by the obtained age prediction. Chr: chronological age, pred: predicted age.

Age group	Number of samples	RMSE	% of samples within RMSE	% of samples within 2xRMSE	Maximal absolute deviation
18–35 chr. (trained: 18–45)	36	2.64	50.00	97.2	6.2
36–52 chr. (trained: 26–62)	34	2.8	70.59	100	5.6
53–69 chr. (trained: 43–69)	34	4.67	58.82	97.06	9.6
18–35 (pred)	35	2.62	51.43	97.14	6.2
36–52 (pred)	34	3.02	70.59	97.06	7.9
53–69 (pred)	35	4.49	62.86	97.14	9.6

methylation pattern due to the known correlation between CpG sites within a range of a few 100 bp [42,43]. Therefore, 450 K probes that are known to target a CpG with a putative SNP were removed from marker selection [32]. However, the final chosen CpG site was not in all cases the original 450 K site or were not yet in the used dbSNP137 which was available at the beginning of the study. Additionally, other SNPs can occur in the amplicon. All MPS results were checked for SNPs and some samples that showed an aberrant DNAm or age prediction were analyzed using unconverted DNA and Sanger sequencing. All results and affected positions can be found in Suppl. Table S5 and Suppl. Fig. S2. Only the SNPs observed in samples with a high aberrant DNAm will be discussed below. Sanger sequencing did not reveal any additional SNPs within the amplicon or the primer binding site of the MPS primer.

A heterozygous C > G mutation removed in S306 (decreased DNAm for KLF14_2) the CpG KLF14_1 site next to the age-dependent KLF14_2 and created a new CpG site (CCG > CGC). An additional unexpected CpG site on one of the alleles also appeared for NKIRAS2 (4 samples) and ZYG11A (51 samples). In case of NKIRAS2, these four samples also showed an increased DNAm level especially for S087 (37.34%) and S265 (27.42%). However, S260 which shows a very high DNAm for the given age (age: 68, DNAm: 35.98%) did not reveal any detectable SNP for NKIRAS2 as confirmed with Sanger sequencing. Additional CpG sites can lead to difficulties to be identified via the MPS run, as they can behave differently dependent on the methylation status of these newly generated CpGs. A SNP in a non-CpG within SAMD10 was in most cases heterozygous and the corresponding DNAm of SAMD10_2 was at the lower boundary or within the general observed range. A higher discrepancy from the mean DNAm was observed for the homozygous variant in S014. In general, a more thorough analysis is needed for SNPs, some SNPs only occurred in a few samples or were heterozygous, not providing enough data to draw a reliable conclusion. Our analysis is limited as bisulfite conversion can cover C > T SNPs (or a G > A SNP if the opposite strand is analyzed) within the sequence and at CpG sites (T SNPs would be analyzed as a non-methylated read). Only a restricted number of samples were additionally analyzed with Sanger sequencing. Furthermore, interaction due to more distant SNPs or indirect influences by SNPs in other genes can neither be proved nor rejected. However, the performance of an additional analysis of SNPs in casework could be advantageous allowing an additional check of the reliability of the obtained DNAm.

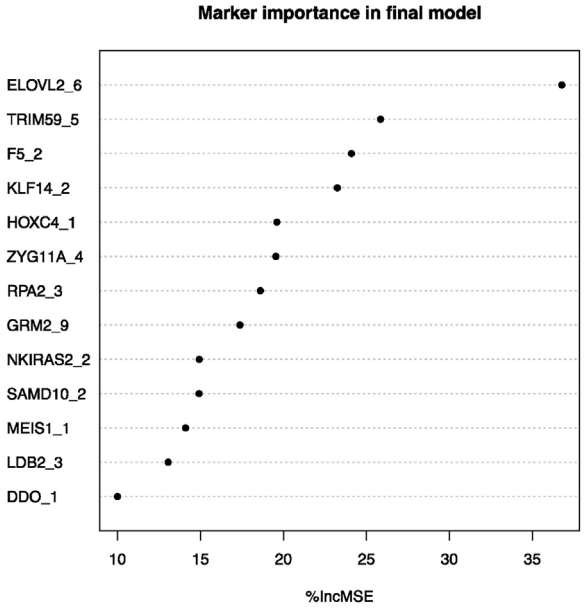

Fig. 4. Increase of mean square error (MSE) in% if the marker would be random assigned values (no age-dependency). It can be seen that ELOVL2_6 contributes most information to the model, followed by TRIM59_5, F5_2, and KLF14_2.

3.5. Possible evaluation of the "unknown" age

In the case of an unknown sample, the expected possible range of deviation needs to be considered from the results of the test set. The conformity of the RMSE of 3.93 years of the test set with the cross-validation of the training set confirms the robustness of the model and can be used to evaluate predictions of an unknown sample. The use of the RMSE and 2xRMSE range allows information about the probability that the chronological age is within these ranges of the predicted age. Around 70% fall into one RMSE range of the trainset and 94% in the 2xRMSE range, 50% of samples were correctly predicted within a range of 2.73 years. Nevertheless, outliers occur, in our test set with a maximal deviation of 12.5 years (S306) using all markers. A visual inspection could help to check if the DNAm values of the predicted age fall within the expected range or if one aberrant DNAm marker could have influenced the prediction. Furthermore, the mentioned SNP analysis can provide further indication of outliers.

3.6. Marker importance and reduced RFR model

We analyzed the variable importance of each single marker provided in the *randomForest* R package by the extracting the normalized% increase of the mean square error (%IncMSE). ELOVL2 (36.8%), TRIM59 (25.8%), F5 (24,1%), and KLF14 (23.2%) had the highest impact on model performance, and DDO (10%) the lowest (Fig. 4).

Using only the top four markers to build a model for age prediction resulted in a slight decrease of model performance obtaining a RMSE 4.63 years, and a MAD 3.64 years for the training set using a ten times repeated 5x-CV. For the test set a RMSE of 4.19 years (73.08% of samples within the RMSE range, and 96.15% within 2xRMSE) and a MAD of 3.24 years was obtained. However, the highest absolute deviation within the test set increased to 14.1 years. The additional use of other markers seems favorable for a more stable analysis and is also not restricted due to the ability of MPS for multiplexing.

The top four age-dependent loci were also thoroughly investigated in previous studies on blood [18,20,25]. Our selected CpG sites do not overlap with the sites chosen by Zbieć-Piekarska et al. (also ELOVL2, TRIM59 and KLF14 in the final assay), and Vidaki et al. (also KLF14). Cg16867657 (ELOVL2_6) was already demonstrated as good age predictor by Zbieć-Piekarska et al. and used by Park et al. [41,44]. TRIM59_5 and ELOVL2_6 were also analyzed by Cho et al. (2017) within their "model 3" (TRIM59_-Pos.5) and "model 4" (ELOVL2_Pos.4), respectively. Discrepancies between position calls of CpG sites occur due to different primer design. We observed for one sample a SNP at the CpG site (cg14361627) of the KLF14 CpG site chosen by Zbieć-Piekarska et al. (2015b) for their final assay [20] which is directly upstream of our final KLF14_2 CpG site. However, this SNP has a very low minor allele frequency within the 1000 Genome project (G = 0.0008) (dbSNP150). Freire-Aradas et al. used the EpiTyper system approach for age determination. ELOVL2, TRIM59, ZYG11A, and F5 were also part of their initial selected sites. TRIM59 and ZYG11A were not further considered due to primer design challenges, and also F5 was not in the final assay. ELOVL2 was represented in their final assay using the CpG site cg21572722 [45]. FHL2, which was detected in other studies as [20,45], was not included in our study. FHL2 was removed during pre-processing due a potential blood cell-type dependent DNAm pattern. A measured correlation of the

DNAm with age could then be caused by a cell count shift with age rather than a direct DNAm change [33].

3.7. Potential, limitations, and further improvement of the model

The use of MPS allows the parallel investigation of multiple markers. It should be mentioned, that only one of the bisulfite converted strands was amplified and analyzed, not considering the rare occurring phenomenon of hemi-methylation of CpG sites (especially after replication) [46,47]. This phenomenon could lead to a discrepancy between the methylation status of a CpG cytosine on one strand compared to the one on the opposite strand. The analyzed strand was chosen based on the best possible primer design and did not always overlap with the strand analyzed in the 450K data (cf. Suppl. Table S2).

The variation of the measured DNAm due to DNA amount, bisulfite conversion, and MPS coverage needs to be considered. We observed with a few exceptions the aimed coverage of 800 x in our training set samples (6 DNAm values >700 x, 95% CI: 50% +/− 3.78%), and a lower coverage for a few amplicons in the test set (15 DNAm values >400 x, leading to a slight increased 95% CI of 50% DNAm +/− 4.9%). The results were nevertheless kept as all markers are needed for age prediction and as also only limited possibilities for additional analysis are given in a forensic case. We also did not observe a correlation of a lower coverage with a higher prediction error for our samples. Nevertheless, a too low coverage needs to be avoided as it would increase the measurement variation too much. The effect of measurement variability on age prediction will also be marker dependent, as a lower change over life leads to smaller changes per year resulting in a higher needed accuracy. A high improvement of the confidence interval is difficult, as the CI will only slightly narrow down with an increased coverage (e.g. 1000 x coverage: 95% CI 1000 x: 50% DNAm +/− 3.16% compared to 800 x coverage: 50% +/− 3.54% DNAm).

The RFR model represents an easy to integrate machine learning algorithm with only a few parameters to consider (tree depth, number of features chosen at each step of tree splitting and number of trees). The random sub-sampling during model building reduces the risk of overfitting the model on the train data. The obtained results for all 13 markers (Test set: RMSE 3.93 years and MAD 3.16 years) show an improved MAD compared to the studies mentioned based on other methods. Our result is also an improvement to the study of Vidaki et al. (2017) which got a MAD of 7.45 years for their test set of 46 samples using MPS analysis However, model training was performed using public 450 K data, which could have led to a technical bias [18]. Moreover, models can be created for specific age-groups if a-priori knowledge of the individual/sample is available. A reduced model version using only the top four markers still gave a reliable result for the training set (RMSE 4.63 years, MAD 3.64 years) and the test set (RMSE 4.19 years, MAD 3.24 years).

The use of other machine learning algorithms or combinations of different models as well as the use of multiple sites from one locus could be advantageous for a better detection and prediction of outliers. An interesting approach is also the combination with other age prediction tools, as the recently published studies that combine DNA methylation with the sjTREC marker, mRNA expression level, and telomere length, respectively [25,48].

Our model is currently restricted to the analyzed age range of 18–69 years. Inclusion of younger and older individuals would broaden the predictable age range. Although no statistically significant difference was found for the accuracy of the prediction of the age dependent on gender, more samples per gender could perhaps lead to a further improvement when including gender. In the future, samples of other populations are needed to prove the use of all markers for worldwide application. Some markers as ELOVL2, F5, TRIM59, and KLF14 already proved to be useful for age prediction as mentioned before and were identified using other datasets from other populations, as well as different gender and age compositions. Especially ELOVL2 was investigated extensively [17,18,20,25,41,45]. However, different CpG sites were used in these studies, and it is important to find out if the same sites can be used or if it is favorable to choose other sites dependent on the population. Cho et al. revealed that another CpG site of ELOVL2 (Pos.1) showed higher age-correlation in Koreans, compared to the site chosen by Zbieć-Piekarska et al. (C7) [25]. Profound experiments are still needed for the analysis of low template samples and also for validation of technical variation between experiments and laboratories. For validation, a clear nomenclature will also be needed for inter-laboratory comparisons as currently the CpG sites are labelled according to the 27K/450K identifier, the position within the PCR amplicon or the chromosome location (GRCh37 or GRCh38). The identifier of the 27K/450K are unique in contrast to the chromosomal location which can change with the GRCh. However, the microarrays do not cover all CpG sites. Additional labelling with the number of bp distance to the identifier could be a solution (e.g. KLF_2: cg08097417–15 bp), but can be replaced with an identifier in case of newer microarrays covering that position.

4. Conclusion

An epigenetic DNAm-based age prediction model for whole blood completely based on MPS was created for the age range of 18–69 years. The use of RFR allowed the selection and incorporation of linear and non-linear markers. The top markers we found overlap with markers identified in previous studies and we confirm their applicability for individuals living in the Netherlands. The created pipeline allows for simultaneous analysis of DNAm for multiple CpG sites, as well as a quality control for bisulfite conversion, read coverage, and SNP detection. The use of 13 markers for age prediction resulted in a robust model, and also the reduction to four markers still gave good predictions. Inter-individual differences and the use of only a restricted number of markers however limits the ability to further reduce MAD and RMSE. Further optimization in marker and model choice, and a better understanding of other genomic phenomena such as SNPs will help to comprehend the occurrence of outliers and to improve their detection.

Conflicts of interest

The authors declare no conflict of interest.

Funding

This study was funded by the Ministry of Justice and Security (NCTV grant, project "Leeftijdsbepaling") and the Swammerdam Institute for Life Sciences, University of Amsterdam. Further supplementary funding of material was obtained from the Amsterdam University fund.

Acknowledgements

We want to thank Andrea Venema and Adri Mul (both AMC) for their introduction into the 450K data analysis and DNA methylation analysis using MPS. Special thanks go to the Sanquin blood donors who allowed the use of their rest-blood for research purposes.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.fsigen.2017.07.015.

References

[1] A. Vidaki, B. Daniel, D.S. Court, Forensic DNA methylation profiling—potential opportunities and challenges, Forensic Sci. Int. Genet. 7 (2013) 499–507, doi:http://dx.doi.org/10.1016/j.fsigen.2013.05.004.

[2] F. Kader, M. Ghai, DNA methylation and application in forensic sciences, Forensic Sci. Int. Genet. 249 (2015) 255–265, doi:http://dx.doi.org/10.1016/j.forsciint.2015.01.037.

[3] C. Bock, I. Beerman, W.-H. Lien, Z.D. Smith, H. Gu, P. Boyle, A. Gnirke, E. Fuchs, D. J. Rossi, A. Meissner, DNAmethylation dynamics during In vivo differentiation of blood and skin stem cells, Mol. Cell. 47 (2012) 633–647, doi:http://dx.doi.org/10.1016/j.molcel.2012.06.019.

[4] B.A. Benayoun, E.A. Pollina, A. Brunet, Epigenetic regulation of ageing: linking environmental inputs to genomic stability, Nat. Rev. Mol. Cell Biol. 16 (2015) 593–610, doi:http://dx.doi.org/10.1038/nrm4048.

[5] M.J. Ziller, H. Gu, F. Müller, J. Donaghey, L.T.-Y. Tsai, O. Kohlbacher, P.L. De Jager, E.D. Rosen, D.A. Bennett, B.E. Bernstein, A. Gnirke, A. Meissner, Charting a dynamic DNA methylation landscape of the human genome, Nature 500 (2013) 477–481, doi:http://dx.doi.org/10.1038/nature12433.

[6] J.H. An, A. Choi, K.-J. Shin, W.I. Yang, H.Y. Lee, DNA methylation-specific multiplex assays for body fluid identification, Int. J. Legal Med. 127 (2012) 35–43, doi:http://dx.doi.org/10.1007/s00414-012-0719-1.

[7] J.-L. Park, O.-H. Kwon, J.H. Kim, H.-S. Yoo, H.-C. Lee, K.-M. Woo, S.-Y. Kim, S.-H. Lee, Y.S. Kim, Identification of body fluid-specific DNA methylation markers for use in forensic science, Forensic Sci. Int. Genet. 13 (2014) 147–153, doi:http://dx.doi.org/10.1016/j.fsigen.2014.07.011.

[8] H.Y. Lee, M.J. Park, A. Choi, J.H. An, W.I. Yang, K.-J. Shin, Potential forensic application of DNA methylation profiling to body fluid identification, Int. J. Legal Med. 126 (2011) 55–62, doi:http://dx.doi.org/10.1007/s00414-011-0569-2.

[9] S. Horvath, DNA methylation age of human tissues and cell types, Genome Biol. 14 (R115) (2013), doi:http://dx.doi.org/10.1186/gb-2013-14-10-r115.

[10] G. Hannum, J. Guinney, L. Zhao, L. Zhang, G. Hughes, S. Sadda, 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 (2013) 359–367, doi:http://dx.doi.org/10.1016/j.molcel.2012.10.016.

[11] H. Heyn, N. Li, H.J. Ferreira, S. Moran, D.G. Pisano, A. Gomez, J. Diez, J.V. Sanchez-Mut, F. Setien, F.J. Carmona, A.A. Puca, S. Sayols, M.A. Pujana, J. Serra-Musach, I. Iglesias-Platas, F. Formiga, A.F. Fernandez, M.F. Fraga, S.C. Heath, A. Valencia, I.G. Gut, J. Wang, M. Esteller, Distinct DNA methylomes of newborns and centenarians, Proc. Natl. Acad. Sci. 109 (2012) 10522–10527, doi:http://dx.doi.org/10.1073/pnas.1120658109.

[12] 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, T.M. Consortium, 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 (2012) e1002629, doi:http://dx.doi.org/10.1371/journal.pgen.1002629.

[13] M.F. Fraga, E. Ballestar, M.F. Paz, S. Ropero, F. Setien, M.L. Ballestar, D. Heine-Suñer, J.C. Cigudosa, M. Urioste, J. Benitez, M. Boix-Chornet, A. Sanchez-Aguilera, C. Ling, E. Carlsson, P. Poulsen, A. Vaag, Z. Stephan, T.D. Spector, Y.-Z. Wu, C. Plass, M. Esteller, Epigenetic differences arise during the lifetime of monozygotic twins, Proc. Natl. Acad. Sci. U. S. A. 102 (2005) 10604–10609, doi:http://dx.doi.org/10.1073/pnas.0500398102.

[14] Q. Tan, B.T. Heijmans, J. v B. Hjelmborg, M. Soerensen, K. Christensen, L. Christiansen, Epigenetic drift in the aging genome: a ten-year follow-up in an elderly twin cohort, Int. J. Epidemiol. 45 (2016) 1146–1158, doi:http://dx.doi.org/10.1093/ije/dyw132.

[15] S. Marttila, L. Kananen, S. Häyrynen, J. Jylhävä, T. Nevalainen, A. Hervonen, M. Jylhä, M. Nykter, M. Hurme, Ageing-associated changes in the human DNA methylome: genomic locations and effects on gene expression, BMC Genomics 16 (2015) 179, doi:http://dx.doi.org/10.1186/s12864-015-1381-z.

[16] S. Bocklandt, W. Lin, M.E. Sehl, F.J. Sánchez, J.S. Sinsheimer, S. Horvath, E. Vilain, Epigenetic predictor of age, PLoS One 6 (2011) e14821, doi:http://dx.doi.org/10.1371/journal.pone.0014821.

[17] B. Bekaert, A. Kamalandua, S.C. Zapico, W.V. de Voorde, R. Decorte, Improved age determination of blood and teeth samples using a selected set of DNA methylation markers, Epigenetics 10 (2015) 922–930, doi:http://dx.doi.org/10.1080/15592294.2015.1080413.

[18] 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, doi:http://dx.doi.org/10.1016/j.fsigen.2017.02.009.

[19] 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, doi:http://dx.doi.org/10.1016/j.fsigen.2015.05.007.

[20] 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, doi:http://dx.doi.org/10.1016/j.fsigen.2015.05.001.

[21] S.K. Mawlood, L. Dennany, N. Watson, B.S. Pickard, The EpiTect Methyl qPCR Assay as novel age estimation method in forensic biology, Forensic Sci. Int. Genet. 264 (2016) 132–138, doi:http://dx.doi.org/10.1016/j.forsciint.2016.03.047.

[22] C.I. Weidner, Q. Lin, C.M. Koch, L. Eisele, F. Beier, P. Ziegler, D.O. Bauerschlag, K.-H. 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 (R24) (2014), doi:http://dx.doi.org/10.1186/gb-2014-15-2-r24.

[23] 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 (srep17788) (2015), doi:http://dx.doi.org/10.1038/srep17788.

[24] 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, doi:http://dx.doi.org/10.1016/j.fsigen.2017.04.006.

[25] 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, doi:http://dx.doi.org/10.1016/j.fsigen.2017.04.020.

[26] D.S.B.S. Silva, J. Antunes, K. Balamurugan, G. Duncan, C.S. Alho, B. McCord, Evaluation of DNA methylation markers and their potential to predict human aging, Electrophoresis (2015), doi:http://dx.doi.org/10.1002/elps.201500137.

[27] 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, doi:http://dx.doi.org/10.1016/j.fsigen.2015.05.014.

[28] L. Breiman, Random forests, Mach. Learn. 45 (2001) 5–32, doi:http://dx.doi.org/10.1023/A:1010933404324.

[29] M. van Iterson, E.W. Tobi, R.C. Slieker, W. den Hollander, R. Luijk, P.E. Slagboom, B.T. Heijmans, MethylAid: visual and interactive quality control of large Illumina 450k datasets, Bioinformatics 30 (2014) 3435–3437, doi:http://dx.doi.org/10.1093/bioinformatics/btu566.

[30] M.J. Aryee, A.E. Jaffe, H. Corrada-Bravo, C. Ladd-Acosta, A.P. Feinberg, K.D. Hansen, R.A. Irizarry, Minfi: a flexible and comprehensive Bioconductor package for the analysis of Infinium DNA methylation microarrays, Bioinformatics 30 (2014) 1363–1369, doi:http://dx.doi.org/10.1093/bioinformatics/btu049.

[31] J.-P. Fortin, A. Labbe, M. Lemire, B.W. Zanke, T.J. Hudson, E.J. Fertig, C.M. Greenwood, K.D. Hansen, Functional normalization of 450k methylation array data improves replication in large cancer studies, Genome Biol. 15 (2014) 503, doi:http://dx.doi.org/10.1186/s13059-014-0503-2.

[32] Y. Chen, M. Lemire, S. Choufani, D.T. Butcher, D. Grafodatskaya, B.W. Zanke, S. Gallinger, T.J. Hudson, R. Weksberg, Discovery of cross-reactive probes and polymorphic CpGs in the Illumina Infinium HumanMethylation450 microarray, Epigenetics 8 (2013) 203–209, doi:http://dx.doi.org/10.4161/epi.23470.

[33] A.E. Jaffe, R.A. Irizarry, Accounting for cellular heterogeneity is critical in epigenome-wide association studies, Genome Biol. 15 (R31) (2014), doi:http://dx.doi.org/10.1186/gb-2014-15-2-r31.

[34] L.F.A. Wessels, M.J.T. Reinders, A.A.M. Hart, C.J. Veenman, H. Dai, Y.D. He, L.J. van't Veer, A protocol for building and evaluating predictors of disease state based on microarray data, Bioinformatics 21 (2005) 3755–3762, doi:http://dx.doi.org/10.1093/bioinformatics/bti429.

[35] N. Rohland, D. Reich, Cost-effective, high-throughput DNA sequencing libraries for multiplexed target capture, Genome Res. 22 (2012) 939–946, doi:http://dx.doi.org/10.1101/gr.128124.111.

[36] F. Krueger, Babraham Bioinformatics - Trim Galore!, (n.d.). https://www.bioinformatics.babraham.ac.uk/projects/trim_galore/(Accessed 24, May 2017).

[37] J. Zhang, K. Kobert, T. Flouri, A. Stamatakis, PEAR: a fast and accurate Illumina Paired-End reAd mergeR, Bioinformatics 30 (2014) 614–620, doi:http://dx.doi.org/10.1093/bioinformatics/btt593.

[38] W.Z.- Bioinformatics, biscuit: BISuilfite-seq CUI Toolkit, 2017. https://github.com/zwdzwd/biscuit (Accessed 24, May 2017).

[39] D. Ryan, MethylDackel: A. (mostly) universal methylation extractor for BS-seq experiments, 2017. https://github.com/dpryan79/MethylDackel (Accessed 24, May 2017).

[40] H. Li, B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, R. Durbin, The sequence Alignment/Map format and SAMtools, Bioinformatics 25 (2009) 2078–2079, doi:http://dx.doi.org/10.1093/bioinformatics/btp352.

[41] J.-L. Park, J.H. Kim, E. Seo, D.H. Bae, S.-Y. Kim, H.-C. Lee, K.-M. Woo, Y.S. Kim, Identification and evaluation of age-correlated DNA methylation markers for forensic use, Forensic Sci. Int. Genet. 23 (2016) 64–70, doi:http://dx.doi.org/10.1016/j.fsigen.2016.03.005.

[42] J.T. Bell, A.A. Pai, J.K. Pickrell, D.J. Gaffney, R. Pique-Regi, J.F. Degner, Y. Gilad, J.K. Pritchard, DNA methylation patterns associate with genetic and gene expression variation in HapMap cell lines, Genome Biol. 12 (R10) (2011), doi:http://dx.doi.org/10.1186/gb-2011-12-1-r10.

[43] F. Eckhardt, J. Lewin, R. Cortese, V.K. Rakyan, J. Attwood, M. Burger, J. Burton, T. V. Cox, R. Davies, T.A. Down, C. Haefliger, R. Horton, K. Howe, D.K. Jackson, J. Kunde, C. Koenig, J. Liddle, D. Niblett, T. Otto, R. Pettett, S. Seemann, C. Thompson, T. West, J. Rogers, A. Olek, K. Berlin, S. Beck, DNA methylation profiling of human chromosomes 6, 20 and 22, Nat. Genet. 38 (2006) 1378–1385, doi:http://dx.doi.org/10.1038/ng1909.

[44] 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, doi:http://dx.doi.org/10.1016/j.fsigen.2014.10.002.

[45] A. Freire-Aradas, C. Phillips, A. Mosquera-Miguel, L. Girón-Santamaría, A. Gómez-Tato, M. Casares de Cal, J. Álvarez-Dios, J. Ansede-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, doi:http://dx.doi.org/10.1016/j.fsigen.2016.06.005.

[46] Y. Li, J. Zhu, G. Tian, N. Li, Q. Li, M. Ye, H. Zheng, J. Yu, H. Wu, J. Sun, H. Zhang, Q. Chen, R. Luo, M. Chen, Y. He, X. Jin, Q. Zhang, C. Yu, G. Zhou, J. Sun, Y. Huang, H. Zheng, H. Cao, X. Zhou, S. Guo, X. Hu, X. Li, K. Kristiansen, L. Bolund, J. Xu, W. Wang, H. Yang, J. Wang, R. Li, S. Beck, J. Wang, X. Zhang, The DNA methylome of human peripheral blood mononuclear cells, PLoS Biol. 8 (2010) e1000533, doi:http://dx.doi.org/10.1371/journal.pbio.1000533.

[47] A. Jeltsch, R.Z. Jurkowska, New concepts in DNA methylation, Trends Biochem. Sci. 39 (2014) 310–318, doi:http://dx.doi.org/10.1016/j.tibs.2014.05.002.

[48] D. Zubakov, F. Liu, I. Kokmeijer, Y. Choi, J.B.J. van Meurs, W.F.J. van IJcken, A.G. Uitterlinden, A. Hofman, L. Broer, C.M. van Duijn, J. Lewin, M. Kayser, Human age estimation from blood using mRNA, DNA methylation, DNA rearrangement, and telomere length, Forensic Sci. Int. Genet. 24 (2016) 33–43, doi:http://dx.doi.org/10.1016/j.fsigen.2016.05.014.