Identifying blood-specific age-related DNA methylation markers on the Illumina MethylationEPIC® BeadChip

Hussain Alsaleh[1], Penelope R. Haddrill[1]

1 Pure and Applied Chemistry, University of Strathclyde

Abstract

The past decade has seen rapid development in DNA methylation (DNAm) microarrays, including the Illumina HumanMethylation27 and HumanMethylation450 (450K) chips, which have played an essential role in identifying and evaluating age-related (AR) DNAm markers in different tissues. Recently, a new array, the Illumina MethylationEPIC (EPIC) was introduced, with nearly double the number of probes as the 450K (~850,000 probes). In this study, we test these newly added probes for age association using a large cohort of 754 DNAm profiles from blood samples assayed on the EPIC BeadChip, for individuals aged 0-88 years old. 52 AR CpG sites (Spearman's abs(rho) >0.6 and *P*-value $<10^{-83}$) were identified, 21 of which were novel sites and mapped to 18 genes, nine of which (*LHFPL4, SLC12A8, EGFEM1P, GPR158, TAL1, KIAA1755, LOC730668, DUSP16,* and *FAM65C)* have never previously been reported to be associated with age. The data were subsequently split into a 527-sample training set and a 227-sample testing set to build and validate two age prediction models using elastic net regression and multivariate regression. Elastic net regression selected 425 CpG markers with a mean absolute deviation (MAD) of 2.6 years based on the testing set. To build a multivariate linear regression model, AR CpG sites with $R^2 > 0.5$ at *FDR* < 0.05 were input into stepwise regression to select the best subset for age prediction. The resulting six CpG markers were linearly modelled with age and explained 81% of age-correlated variation in DNAm levels. Age estimation accuracy using bootstrap analysis was 4.5 years, with 95% confidence intervals of 4.56 to 4.57 years based on the testing set. These results suggest that EPIC BeadChip probes for age estimation fall within the range of probes found on the previous Illumina HumanMethylation platforms in terms of their age-prediction ability.

***Keywords**:* DNA methylation, Forensic epigenetics, Forensic age estimation, Illumina MethylationEPIC, Age, CpG sites.

1 Introduction

Aging can be described as the decline in a set of vital cellular functions that occur over time [1]. This consequently disrupts the homeostatic regulation of the body, which leads to various age-related (AR) diseases such as cancer, and cardiovascular disease. Due to its significant association with chronological age, DNA methylation (DNAm) has been the focus of much attention in the field of epigenetics, and in particular forensic epigenetics, which has proven a more significant association than those found with other factors such as telomere length, mitochondrial dysfunction, loss of proteostasis, and stem cell exhaustion [2]. Thus, age-related (AR) DNAm sites have become important and robust biomarkers for accurately measuring biological age. Aging has a profound effect on DNAm pattern, and so this also affects gene expression, which results in susceptibility to diseases and various health outcomes. Predicted age, also known as "DNAm age" or "epigenetic age" has been found to be related to frailty [3], cognitive/physical fitness in the elderly [4], Parkinson's disease, Alzheimer's disease-related neuropathology [5], and can predict overall mortality in humans [6]. In addition to clinical applications, DNAm age has also received a great deal of attention in forensic epigenetic studies, because of its accuracy in age estimation. This can be implemented in forensic investigations to predict the age of unknown individuals, using their biological samples recovered from the crime scenes, which can provide extremely valuable intelligence information for a police investigation [7-9].

DNAm is an epigenetic mechanism that involves the addition of a methyl group to the 5' position of cytosine residues that are mostly found in form of cytosine-guanine dinucleotide sequences (known as CpG sites). Historically, the only way to study the association between chronological age and DNAm level was to measure the global decrease in the content of the 5'-methylated cytosine in aged cultured cells [10]. However, technologies to analyse DNAm in gene-specific and genome-wide manner have developed significantly in recent years. For instance, gene-specific assays such as EpiTect, SNaPshot, EpiTYPER and targeted bisulfite sequencing have become prevalent in DNAm-related studies for their sensitive and reliable quantification of the DNAm

level [11-14]. However, genome-wide assays that provide the opportunity to quantify methylation level at a single base level, such as the Illumina Infinium HumanMethylation BeadChip technology, have become the main choice for many research groups carrying out epigenome-wide association studies (EWAS). The introduction of two Illumina HumanMethylation BeadChips, namely the HumanMethylation27 (27K), and HumanMethylation450K (450K) BeadChips, was crucial for identifying a huge number of AR CpG sites and genes in the literature. In addition, the public genomic databases have become a rich source of epigenome-wide DNAm data, from a large body of epigenetic studies based on different human tissues [7].

The first two blood specific EWAS looking for an association with age, conducted by Rakyan et al. [15] and Bell et al. [16], were based on Illumina Infinium 27K BeadChip. A total of 775 age-differentially methylated regions (aDMRs) were identified, 90% of them located within promoters of genes [15]. Moreover, it was demonstrated that AR CpG markers were predominantly hypermethylated with age, which may indicate that aberrant hypermethylation of the promoter regions of genes is associated with cancer, and AR diseases [17]. The aDMRs identified in both studies were limited to the methylation sites that were covered by the 27K probes (~27,000 CpG sites), which were relatively sparse and promoter-specific [16]. Thus, Illumina developed a new chip, the Infinium 450K BeadChip, which targeted ~450,000 CpG sites covering 99% of RefSeq genes [18], and a greater number of CpG islands, shores, FANTOM4 promoters [19], and enhancers [20]. This has allowed researchers to interrogate more genomic regions spanning a wide range of genes.

Garagnani et al. [21] were the first to study aging in whole blood using the 450K BeadChip, and their study consisted of a small cohort of 64 individuals aged from 9 to 83 years old. Although they stated that they identified 163 AR CpG sites, only the top nine of these were reported in their paper [21]. These CpG sites were mapped to CpG islands located in the promoter region of three genes, namely, *ELOVL2*, *FHL2* and *PENK*. Furthermore, a cross-sectional study of 965 participants (aged 50-75 years), conducted by Florath et al. [22] also found 162 AR CpG sites, eight of which were the same to the nine CpG markers that were reported by Garagnani et al. (2012). The strong relationship between age and methylation level for the identified AR CpG sites prompted other researchers to exploit them for various applications, such as age estimation for forensic and health purposes [6,23-26]. For example, the first blood specific DNAm age prediction

model was built by Hannum et al. [27]. Their model consisted of 71 CpG markers and was trained on 482 DNAm profiles (from individuals aged 19 to 101 years) assayed on the 450K BeadChip, along with clinical variables such as sex and Body Mass Index (BMI). The mean absolute deviation (MAD) between chronological age and estimated age for their model was 3.9 years when based on the training samples, and 4.9 years based on independent 174 testing samples [27].

Recently, a new array, the Illumina MethylationEPIC (EPIC) BeadChip was introduced, containing over 860,000 probes, nearly double the number on the 450K. Not all the 450K probes were included in the new EPIC BeadChip, ~90% of them were included, but others were removed as a result of reports of poor performance [28]. The newly added probes provide a higher coverage of various genomic regions, such as RefSeq genes, ENCODE [29] and FANTOM5 enhancers [30], DNase hypersensitive sites, miRNA promoter regions, differentially methylated sites in tumor tissues, and non-coding regions such as CpG islands, shores, shelves, and open sea [31]. The EPIC BeadChip is a promising tool to further our understanding of DNAm mechanisms in human development and disease, and in particular the DNAm landscape of distal regulatory elements. In this paper, we perform a comprehensive evaluation of blood-specific AR CpG sites found on the new EPIC BeadChip, and identify their associated genes, which will provide new insights for researchers in various epigenetic and genetic disciplines. Enhancing the accuracy of DNAm based age-prediction models, by searching for new AR CpG sites on the EPIC BeadChip with better age prediction accuracy, will aid forensic investigations in criminal cases where biological samples of unknown origin have been recovered. For this reason, we build an age prediction model using the probes on the EPIC BeadChip, which we test to determine how well it performs in comparison to other models constructed using the previous Illumina HumanMethylation platforms (27K and 450K).

2 Materials and methods

2.1 EPIC data sets

A total of 756 DNAm profiles assayed on the EPIC BeadChip in individuals aged 0-88 years old, were assembled by combining three separate data sets retrieved from the National Centre for Biotechnology Information (NCBI) Gene Expression Omnibus (GEO), which is an online genomic data repository. The accession number of each data set and brief description of the samples can be

found in Table 1. To ensure identification of AR CpG sites was not biased towards a specific range of chronological ages, whole cord blood samples were included in this study, which represent time zero in human age (Figure 1).

Table 1 Description of the three data sets used in this study.

Accession number	DNA origin	n (Prop. female)	Median Age(range)	Citation
GSE103189	Whole cord blood	8 (0.38)	0 (0)	Dou et al. [32]
GSE123914	Whole blood	69 (1)	59 (51-65)	Zaimi et al. [33]
GSE116339	Whole blood	679 (0.59)	53 (23-88)	Curtis et al. [34]

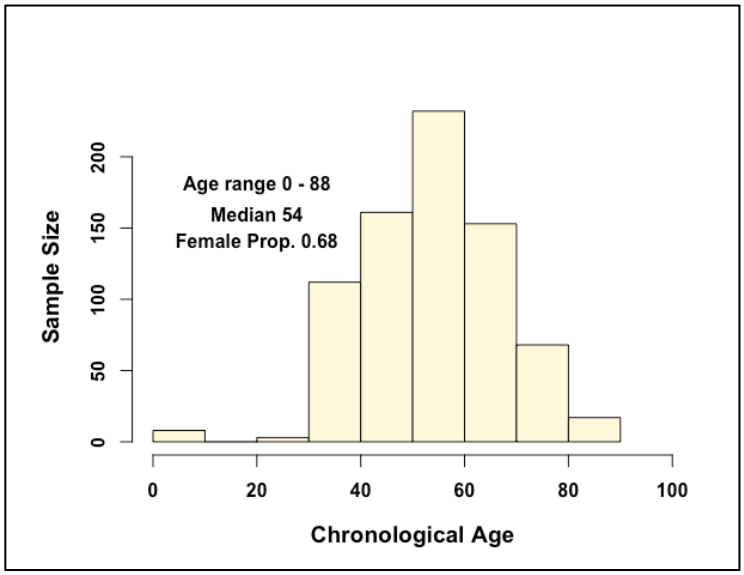

Figure 1 Distribution and descriptive statistics relating to the chronological ages of individuals who provided the samples used in this study.

The first data set (GSE103189) was obtained from the study conducted by Dou et al. [32], which aimed to evaluate the cell composition and DNAm differences between cord blood buffy coat and whole cord blood samples, which revealed no significant differences between them and thus they can be analytically combined and compared. The next data set (GSE123914) was obtained from a longitudinal study by Zaimi et al., which aimed to examine the variation in DNAm level over a 1-year period in whole blood samples collected from 35 healthy women [33]. It was shown in this

study that the median intraclass correlation coefficient across all CpG sites was 0.19, which suggests a wide variation in DNAm stability over a 1-year period. The last data set (GSE116339) contained 679 whole blood samples, retrieved from a study conducted by Curtis et al. [34], which aimed to investigate whether exposure to polybrominated biphenyl (PBB) was associated with DNAm changes in peripheral blood samples. In this study, a total of 1,890 CpG sites were identified that were associated with total PBB levels [34].

2.2 EPIC data processing

The raw files of each data set were downloaded using *GEOquery* package, which runs in R software [35]. The downloaded files consist of raw signal intensities of the red and green channel. The files were imported into R and converted into methylated and unmethylated signals by applying the '*read.metharray.exp'* function in the *minfi* package [36]. Although the number of CpG probes on the EPIC BeadChip is 865,918, the raw file comes with an additional 186,782 probes (giving a total of 1,052,641 probes). These additional probes were designed for quality control measures, such as background correction, negative controls, bisulfite conversion controls, and hybridisation controls [37].

As was the case on the 450K BeadChip, probes on the EPIC BeadChip also have two chemistry designs, Infinium I and II, which possess different DNAm value distributions, introducing unwanted variation into the methylation values [20]. Thus, the two probe designs need to be normalised to render them comparable to each other, which was done using subset quantile normalisation implemented in the *preprocessQuantile* function in the *minfi* package [38]. In addition to the probe type correction, the same function also filtered out probes that did not meet the quality control threshold and had a detection *P*-value > 0.05 in at least one sample. In addition, it filtered out samples with significantly lower values in one of the two signal intensities (red/green channels) compared to the other, which is a quality control measure used to identify sample outliers. The data consist of DNAm signals represented by a Beta value that varies between 0 (hypomethylated) to 1 (hypermethylated), with a bimodal distribution. To ensure that the samples and probes were high quality, any sample that deviated from the normal bimodal pattern was removed from the data set. Finally, probes associated with SNPs and cross-reactive CpGs were

removed from downstream analysis using the *dropLociWithSnps* and *dropCrossReactiveProbes* functions in the *minfi* package [36].

2.2.1 Testing for potential confounders

Given that variation in DNAm has been found to be associated with various factors such as cell type, gender, alcohol intake, smoking, obesity, and certain drugs, it is important to account for these factors as they may cause a confounding effect in EWAS [39]. One of the methods used to discover any hidden relationship between these covariates and the samples is Singular Value Decomposition (SVD). After implementing SVD on the combined data set, segregation was found between the samples, which was based on gender (Figure S2A). For this reason, CpG probes targeting sex chromosomes were filtered out from the downstream statistical analysis (Figure S2B). Another potential confounder in this study was the PBB level, which was measured in the blood samples in the third data set (GSE116339). Since it has been shown that the level of PBB in blood has a significant effect on 1,890 CpG sites, this could also have a confounding effect if it is found to be associated with chronological age. Thus, regression analysis was conducted between PBB level and chronological ages to reveal any linear association between them. Although the P-value of the test was significant (P-value = 1.4×10^{-9}), the R^2 was extremely low (0.05), which indicates that age only explains 5% of the variation in the level of PBB in blood. Finally, batch effects were removed in the data set using a nonparametric empirical Bayes framework method implemented in the *Combat* function within the *SVR* package that runs on R software [40,41].

2.2.2 Estimating and adjusting for cell type composition

It has been demonstrated that the constituents of blood change with aging, thus many DNAm studies adjust for it by including the change in cell composition over time as a covariate in the regression model for statistical analysis [33-35]. The blood cell composition was estimated using an approach proposed by Housemen et al. [42], which is implemented in the *estimateCellCounts* function in the *minfi* package [36]. This function estimated the proportion of the six blood-cell types in each sample: CD8T, CD4T, natural killer cell, B cell, monocyte and granulocyte. The estimated cellular proportions were included in the final multivariate linear regression model.

2.3 Statistical analysis

2.3.1 Identifying AR CpG sites

The DNAm Beta values in the data set were converted to M values ($M= \log_2 Beta/(1-Beta)$) using the *M2Beta* function in the *wateRmelon* package. This transformation was done in order to satisfy the normality and homoscedasticity assumptions of the downstream statistical analyses [43]. Spearman's correlation coefficients between DNAm at each CpG probe and the chronological ages of the samples were calculated using R software. The selection criteria for AR CpG probe candidates were based on two criteria: absolute Spearman's rho ≥0.6, and false discovery rate (FDR) at ≤0.05. The adjusted *P*-value was calculated using *compute.FDR* implemented in the *brainwaver* package. The resulting AR CpG probes were annotated using "Infinium MethylationEPIC v1.0 B4 Manifest File," released by Illumina, which is based on the hg19/GRCh37 human genome assembly.

2.3.2 Building age prediction models

The EPIC BeadChip data were then used to build an age prediction model to determine whether the CpG probes it possesses have better age estimation capabilities compared to the probes found on the old Illumina HumanMethylation platforms such as 27K and 450K. The intended application of the age prediction model will determine the type of method that should be used to build it. For example, if the model will be used for health applications, the number of markers in the model would not pose an issue since the DNA in the sample would usually be in adequate quantities. However, for forensic applications, the number of markers in the model should be kept to a minimum as the quantity of DNA in the majority of forensic samples is low and surveying large numbers of makers requires reasonably large amounts of DNA, due to the destructive procedures involved in DNAm analysis. Therefore, two methods were used to build prediction models, elastic net regression, and multivariate linear regression.

Using the *sample* function in R, the data set was randomly split into a training set and a testing set, with equal relative representation of the various age groups within the sets. The number of samples in the training set was 527, which is 70% of the original set, and 227 samples in the testing set (30%). The sample size of both the training and testing sets in these types of study are important

considerations. The sizes of these data sets in this study were considered to be sufficient, as demonstrated and suggested by Horvath [44] who studied the factors that influence the accuracy of age prediction. They found that the sample size is not significantly (*P*-value = 0.21) correlated with the accuracy of age prediction, as long as it is 100 or greater. That is, the prediction accuracy reaches plateau when the sample sizes of both training and testing sets exceed 100 samples. Elastic net regression was performed using the *glmnet* package in R, and the parameters used were those recommended by Horvath (2013). The feature (CpG marker) selection was based on ten-fold cross-validation, that is, in each fold the training set was split into ten parts, one part served as training set and the rest as validation sets. Then the average error and standard deviation over the ten-folds was computed, and the best subset of markers was determined as the set with the lowest estimation error. The selected subset of CpG markers was then validated on the 227 independent testing samples, and the mean absolute deviation (MAD) between the predicted and chronological age was calculated.

To build an age prediction model with a minimum number of CpG markers, age was linearly regressed on the DNAm level at each CpG site in the training data set, and then markers with R^2 >0.5 at FDR <0.05 were selected. The selected markers were input into a stepwise regression to select the best subset for use in the age prediction model. The stepwise regression was carried out using the *leaps* package in R, which constructs predictive models with all possible subsets of the input CpG sites, then selects the model with the lowest Bayesian Information Criterion (BIC) value, which would have the best predictive ability. The selected CpG markers were then combined in a multivariate linear regression to build the model, and then validated on the testing data set. The model was re-evaluated by bootstrap analysis to ensure its prediction robustness. This involves sampling the testing data set with replacement 10,000 times and calculating MAD values between the predicted age and the chronological age for each bootstrap cohort. From the distribution of the bootstrap observations, the mean of the MAD value was calculated along with the 95% confidence intervals around that mean.

3 Results

3.1 EPIC data sets

The purpose of this investigation was to identify AR CpG markers on the EPIC BeadChip. The analysis initially included 756 samples from three different data sets, however two blood samples (samples GSM3228582, and GSM3228722) from GSE116339 had abnormal Beta value distributions, as shown in the density plot (Figure S1), and thus were removed from the downstream analysis. The number of samples remaining for analysis was 754 samples, and the number of CpG sites after probe filtration was 816,127 probes. Testing for confounding variables was performed by examining how PBB level variation can be explained by age. The results (Table 2) showed that age only explains 5% (P-value $<1.4x10^{-9}$) of the variation in PBB levels in blood. Batch effects were removed using the *Combat* function in R and then visualised using SVD to ensure there was no hidden structure in the data set (Figure S2).

Table 2 Linear regression analysis between PBB levels in each sample and the chronological age of the individual donor.

Term	Estimate (n = 673)	*P*-value	*R*-squared	*P*-value
(Intercept)	-2.20	0.00	0.05	0.00
Age	0.03	0.00		

3.1.1 Estimating and adjusting for cell type composition

The composition of different cell types in each sample was estimated and then tested for association with chronological age. As can be seen in Figure 2, CD4+ T cells, and natural killer (NK) cells had the strongest correlation with age (rho = -0.35 and 0.32 respectively) compared to the other cell types (monocytes, CD4+, granulocytes, and B-cells), which gave rho values of ≤0.19. Therefore, if not adjusted, the change in DNAm level at AR CpG sites could be explained by the change in cell composition, rather than by aging in individuals. For this reason, and to avoid identifying false positive AR markers, each identified AR CpG site in this study was adjusted for cell composition using multivariate linear regression.

276

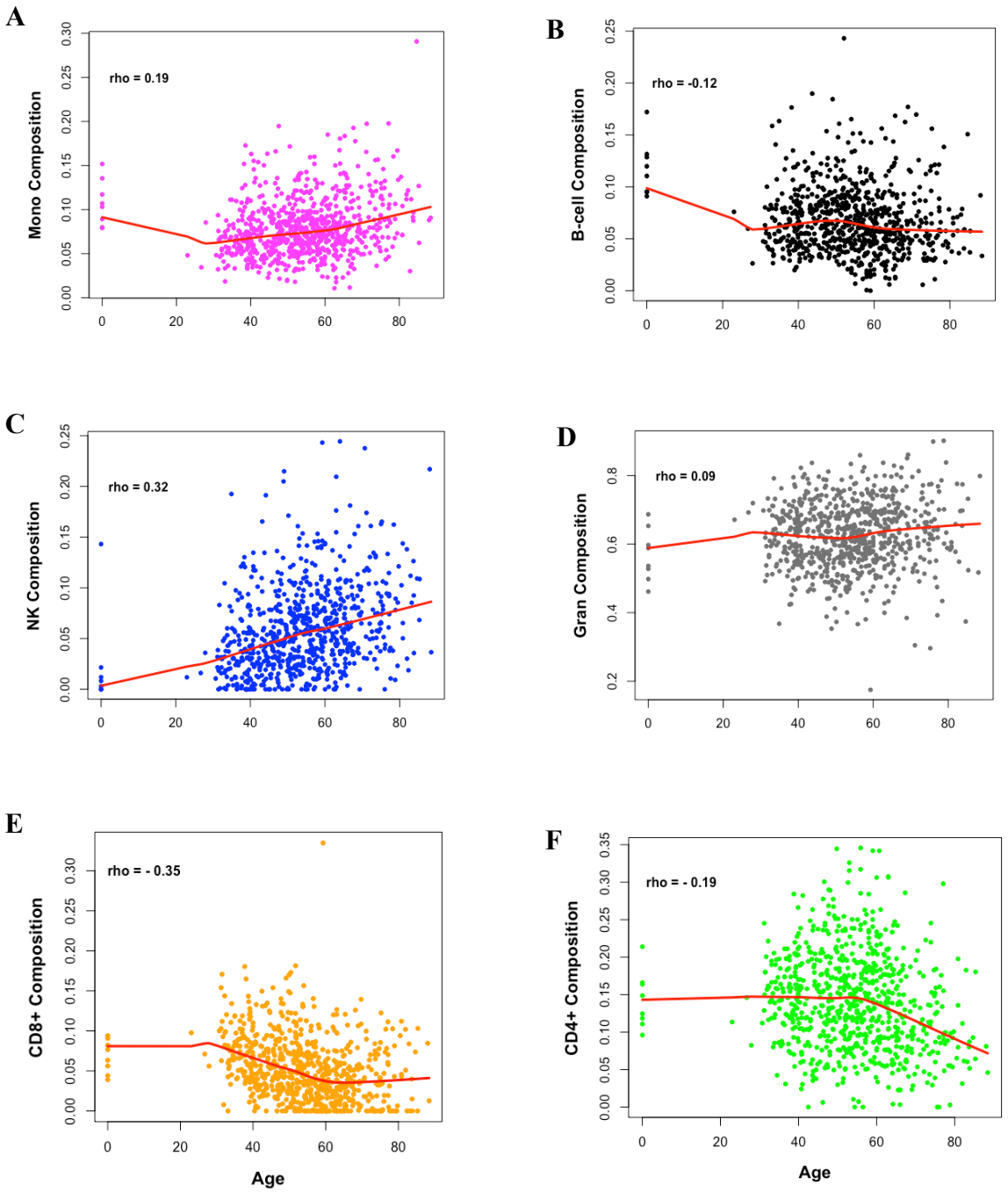

Figure 2 Blood cell composition change with age. The estimated proportions of the six blood cells; (A) monocytes (Mono), (B) B cells, (C) natural killer (NK) cells, (D) granulocytes (Gran), (E) CD8+ T cells, and (F) CD4+ T cells in the samples are plotted against age. Spearman's correlation coefficients are reported for each composition proportion estimate and age. The red lines are weighted regression (Loess) lines fit to the data.

277

3.2 AR CpG markers on the EPIC BeadChip

AR CpG sites were selected on the basis of the Spearman's rank correlation test between chronological age and DNAm level, based on M values. The cut-off value for selecting AR markers was an absolute Spearman's coefficient (rho) > 0.6 at FDR <0.05, as recommended by various studies [45-47]. A total of 52 AR CpG sites passed these conditions (Figure 3A), 19 of which were positively correlated (hypermethylated) and 33 negatively correlated (hypomethylated) with age (Table S1). The AR CpG sites with the top two highest correlation coefficients, were located in the *ELOVL2* gene, which is the most prominent gene associated with age in various tissues, as found in a number of studies (Figure 3B) [21,25,48]. Many of the markers we identified were also identified in other studies that used a similar study design but using the Illumina HumanMethylation450K BeadChip. For example, of the nine AR markers discovered by Garagnani et al. (2012), five were also identified in our study. However, of the remaining four sites, one was dropped by SNP filtration and three had abs(rho) <0.5. In another study, Florath et al. (2015) identified 162 AR CpG sites, of which ten were absent from the EPIC BeadChip, two were dropped after SNP filtration, and only 53 were found with abs(rho) >0.5. In comparing the correlation coefficients of the same AR probes on the two different platforms (450K and EPIC), it was observed that the magnitude of the coefficient values was smaller on the EPIC platform, and for some probes their age association is no longer observed. For instance, nine markers identified by Xu et al. (2015) as highly AR CpG sites (with at least 0.8 abs(rho)) on the 450K platform, were found to have abs(rho) <0.38 on the EPIC BeadChip, which is under the threshold for significant association with age.

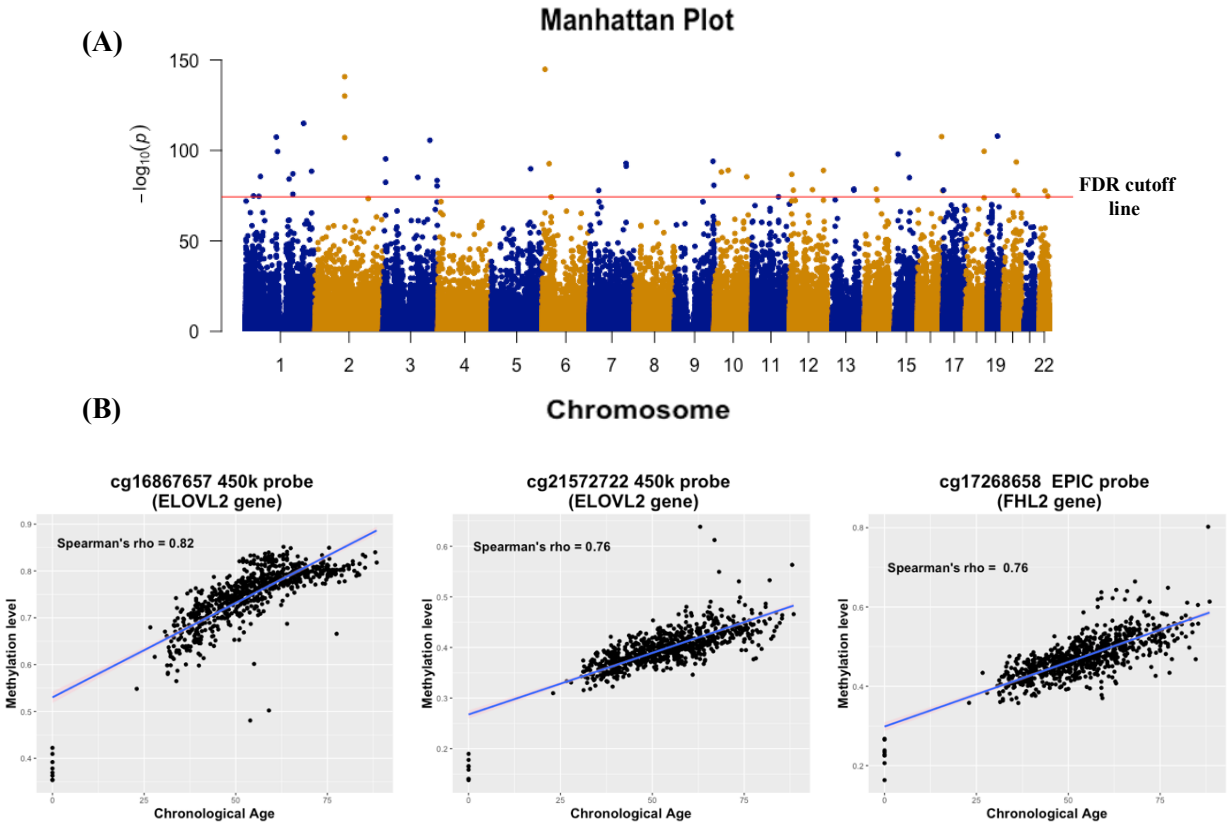

Figure 3 (A) Manhattan plot of *P*-values from Spearman's correlation test between DNAm level at each CpG site and chronological ages in the data set. (B) Scatter plots for the top three AR CpG sites found on the EPIC BeadChip.

3.2.1 Novel AR CpG sites on EPIC BeadChip

From the 52 identified AR CpG sites in this study, 21 CpG sites were from the newly added probes on the EPIC BeadChip, and so these can be considered novel AR CpG sites since they have not been reported in the literature before (Table 3). In addition, they map to 18 genes, nine of which (*LHFPL4, SLC12A8, EGFEM1P, GPR158, TAL1, KIAA1755, LOC730668, DUSP16,* and *FAM65C*) have also never been reported in the literature as being associated with age. The majority of these sites (16) were hypomethylated, and five were hypermethylated with age (Figure 4). The highest positively correlated novel AR CpG site was cg17268658 with rho = 0.76 (*P*-value 1.9×10^{-141}), associated with the *FHL2* gene, which has been reported in many age association studies [22,24,26]. The highest negatively correlated CpG site was cg07323488 with rho = -0.69 (*P*-value

312 $2.6x10^{-106}$), which is linked to a pseudogene known as *EGFEM1P*. Scatter plots of age versus
313 DNAm level for the top four most highly correlated AR CpG sites can be seen in (Figure 5).

314 To account for cell type heterogeneity in blood, the estimated cell composition proportions were
315 included in the multiple linear regression model, and the adjusted estimate was calculated [49].
316 For the 21 novel AR CpG markers, the adjusted estimates after the addition of cell compositions
317 as covariates alongside age in the regression models were within 5% of the original estimates (from
318 the simple regression model that had only age as predictor). This indicates that the DNAm levels
319 at the identified CpG sites were associated with age and not confounded by cell type composition
320 [50].

321

322

323

324

325

326

327

328

329

330

331

332

Table 3 The 21 novel AR CpG sites from the newly added probes on the Illumina EPIC BeadChip, identified by Spearman's correlation test with a cutoff value of abs(rho) >0.6 at FDR <0.05. Probes are sorted from the highest positively to the highest negatively correlated with age.

Probe's ID	UCSC[1] Ref. Gene name	Genomic Location	Chr.[2]	Pos.[3]	Spearman's rho
cg17268658	*FHL2*	TSS200	chr2	106015745	0.76
cg24866418	*LHFPL4*	Body	chr3	9594082	0.66
cg13206721	*GPR158*	TSS1500	chr10	25463350	0.64
cg12841266	*LHFPL4*	Body	chr3	9594093	0.63
cg27099280	*CELF6*	1stExon	chr15	72612204	0.63
cg03650729	*TAL1*	5'UTR	chr1	47692625	-0.6
cg15109150	*FAM65C*	TSS1500	chr20	49308830	-0.6
cg09240238	*LOC730668*	Body	chr22	46402573	-0.6
cg16789844	*PDE1C*	TSS200	chr7	32339213	-0.61
cg01855540	*DUSP16*	TSS1500	chr12	12716653	-0.61
cg17015290	*KIAA1755*	Exon Body	chr20	36850842	-0.61
cg03776853			chr22	36461577	-0.61
cg23719650			chr3	193988507	-0.62
cg25167618	*SLC12A8*	Body	chr3	124840296	-0.63
cg11218872			chr3	193988737	-0.63
cg08587685	*ABLIM1*	Body	chr10	116392206	-0.63
cg08745595	*F5*	TSS1500	chr1	169556012	-0.64
cg05179292	*C1R*	Body	chr12	7244621	-0.64
cg17403084	*PXN*	TSS1500	chr12	120704034	-0.64
cg13552692	*CCDC102B*	5'UTR	chr18	66389447	-0.67
cg07323488	*EGFEM1P*	Body	chr3	168185313	-0.69

[1] Based on UCSC Genome Browser database

[2] Chromosome

[3] Position based on the human assembly GRCh37, also known as hg19.

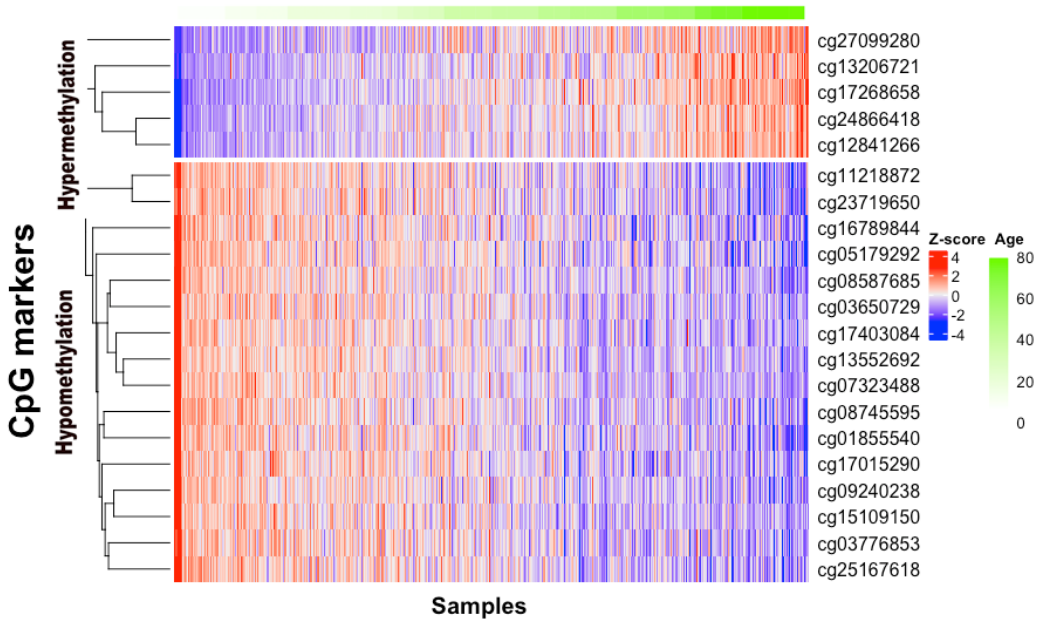

Figure 4 Heat map illustrating methylation level at 21 novel AR CpG markers for each sample in the training data set, ordered by chronological age across samples. The methylation level in each sample is indicated by the Z-score, where red indicates a site is hypermethylated and blue is hypomethylated. Hierarchical clustering of the CpG markers is presented on the left-hand side of the heat map.

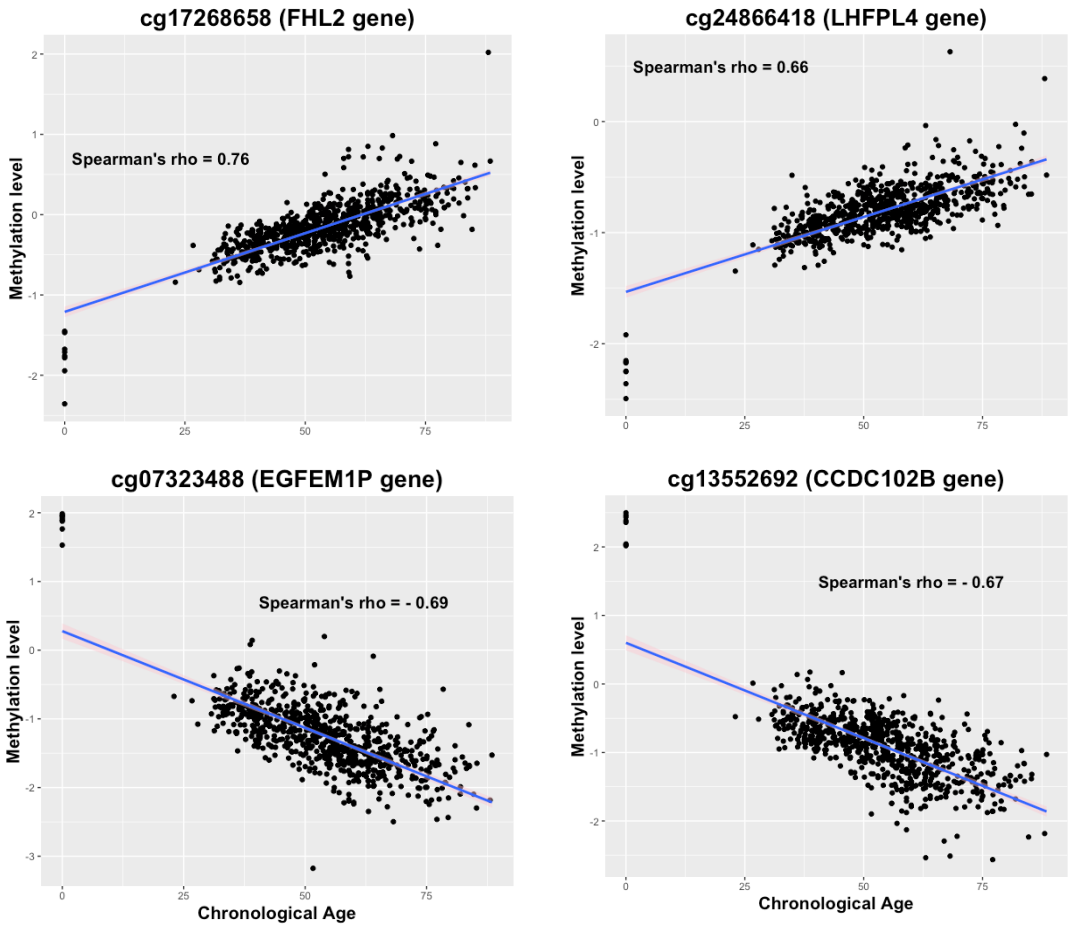

Figure 5 Scatter plots of M values versus chronological age for the top two positively and two negatively correlated AR CpG sites from the newly added probes on the EPIC BeadChip.

3.3 Blood specific age prediction models

In several previous studies where age has been modelled as a function of CpG methylation status for sites in the genome, elastic net regression has been used to perform both feature selection and model building [27,44]. Elastic net regression is ideal for constructing predictive models in cases where the training data has many observations relative to the number of samples [27]. Elastic net regression was performed on the dataset of 816,127 CpG sites and selected 425 AR CpG sites (Table S2) across 527 blood samples. The prediction model containing the selected markers was evaluated on the training data set using one round of ten-fold cross-validation. The prediction accuracy of the model containing the 425 CpG markers based on the training data set was equal to 0.68 years (MAD). Furthermore, its performance was evaluated using an independent validation data set containing 227 blood samples, which resulted in an MAD of 2.6 years, and a Pearson's

correlation coefficient (r) between the predicted and chronological age of 0.97 (95% CI: 0.96–0.98).

To build an age prediction model with the minimum number of AR CpG markers, two steps were carried out. The first step was regressing age on DNAm level at each CpG site in the training data set, and then markers with R^2 >0.5 at FDR <0.05 were selected. Ten CpG markers met this condition, and only two of them were from the newly added probes on the EPIC BeadChip. The second step was to select the best subset of these sites to build an age prediction model. The stepwise regression selected six markers as the best subset for age prediction, which contained only one newly added EPIC BeadChip probe (Table 4). This model explained 81% of the total DNAm levels in the blood samples with prediction accuracy of 4.5 years MAD based on the training data set, and 4.6 years based on the testing data set. The accuracy rate based on bootstrap analysis was 4.5 years, with 95% confidence intervals (CI) of 4.56 to 4.57 years. The correlation (r) between predicted age and chronological age was 0.9 (95% CI: 0.88 – 0.93) (Figure 6). Finally, to avoid gender bias in age prediction, male and female samples in the testing data were separated and their MAD values were assessed separately, to determine whether the difference between them was significant. A t-test showed that there was a non-significant (P-value = 0.3) difference in the prediction accuracy for males (MAD = 4.4 years) compared to females (MAD = 4.9 years).

Table 4 Multivariate linear regression analysis between DNAm levels at six CpG sites and age in the training data set. The CpG marker in bold is the only site exclusively found on the EPIC BeadChip.

Term	Estimate (n = 673)	*P*-value	*R*-squared	*P*-value
(Intercept)	56.10	0.00	0.81	0.00
cg18933331	-9.86	0.00		
cg10501210	-2.68	0.00		
cg06639320	6.58	0.00		
cg24866418	5.55	0.00		
cg16867657	7.50	0.00		
cg17110586	8.14	0.00		

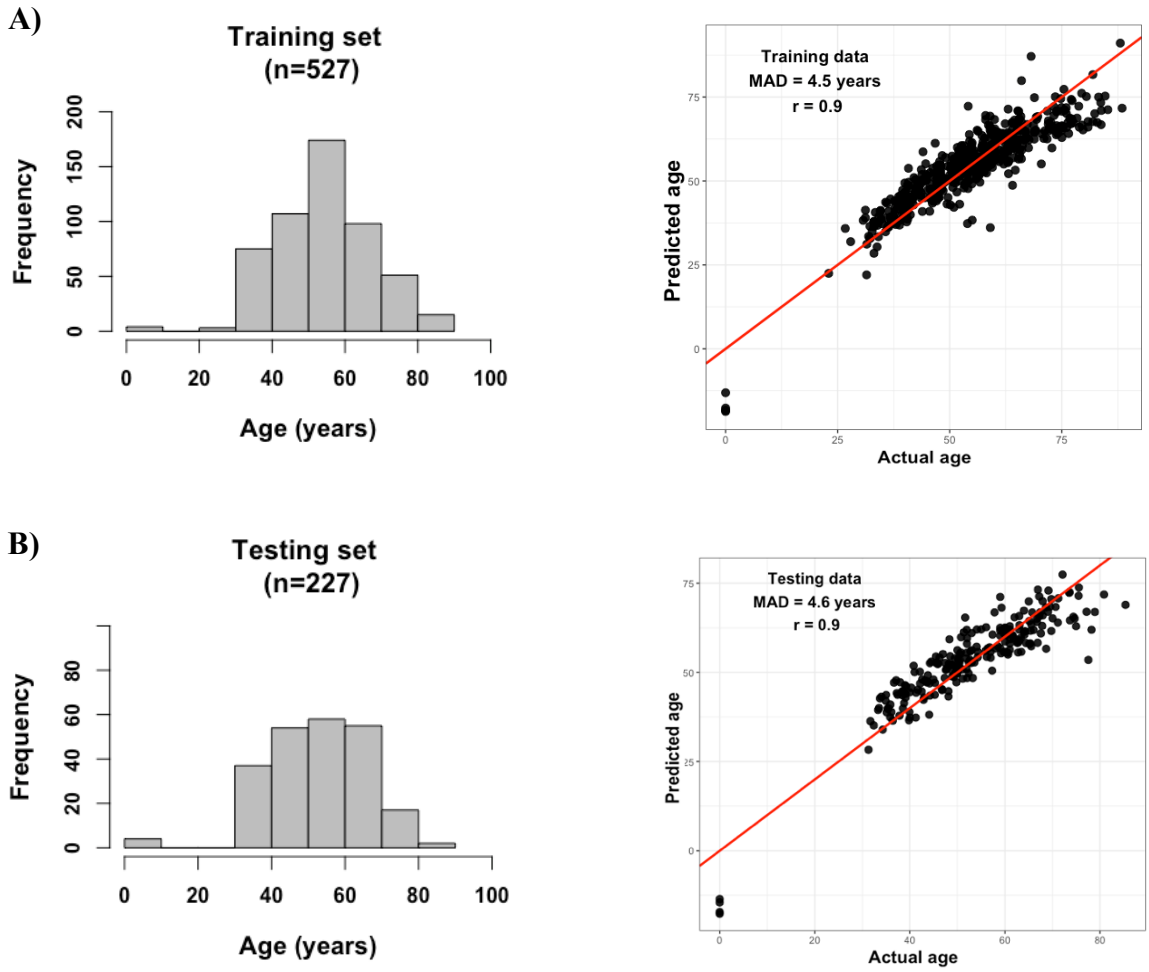

Figure 6 Performance of the multivariate linear regression model consisting of six AR CpG markers. The histograms show age range in the data and the scatter plots show the accuracy of the model in A) The training set of 527 samples, and B) The testing set of 227 samples.

4 Discussion

In this present study, we examined 754 whole blood DNAm profiles assayed on the EPIC BeadChip, and found 52 AR CpG sites, of which 31 were from the 27K and 450K platforms, and 21 were novel AR sites added to the EPIC BeadChip. Apart from these novel sites, all identified AR sites were previously found by different studies who used blood DNAm profiles assayed on the 450K BeadChip. However, their correlation coefficient values on the EPIC BeadChip were lower compared to their values on the 450K array. Although these differences between studies are expected, and may be due to differences in sample size, age range, and the ethnicity of individuals,

an unexpected outcome was that some AR CpG sites with high correlation coefficients on the 450K platform were not associated with age on the EPIC BeadChip. Probes that completely lost their association with age in this study were originally identified in studies with sample sizes below the recommended, which is 100 samples [44]. For example, the number of samples in Xu et al. (2015) was 16 samples, and all their identified AR CpG sites were found to be weakly (abs(rho) < 0.38) associated with age in our study. This suggests that AR probes identified in studies with a small sample size would be more likely to be sample-specific than tissue-specific.

The 21 novel AR CpG sites identified in our study map to 18 genes, nine of which have already been found to be associated with age, namely *ELOVL2, FHL2, CELF6, F5, ABLIM1, PXN, PDE1C, C1R, and CCDC102B*. This indicates that in some cases, adding new probes targeting different genomic locations within the same gene confirms the results obtained from previous EWAS, which, in our case, confirmed the association of these genes with age. In contrast, eight of the remaining nine genes (*LHFPL4, SLC12A8, GPR158, TAL1, KIAA1755, LOC730668, DUSP16, and FAM65C*) were previously targeted by probes that have been shown not to be associated with age. However, by targeting different genomic locations within these same genes, significant age association has been identified. The final gene identified in this study was from a gene newly-targeted on the EPIC BeadChip, *EGFEM1P*.

From the nine newly identified AR genes, two become hypermethylated with age (*LHFPL4* and *GPR158*), and seven become hypomethylated with age. *LHFPL4* is located on chromosome three and encodes a subset of the superfamily of tetraspan transmembrane proteins, which is a critical regulator of postsynaptic GABA clustering in hippocampal pyramidal neurons [51]. Its differential methylation has previously been found to be a biomarker for the early detection of cervical cancer [52]. *GPR158* is located on chromosome ten and encodes a G protein-coupled receptor, which is implicated in many physiological and disease processes [53]. The protein encoded by *DUSP16* on chromosome 12 is a dual specificity phosphatase, implicated in various cellular processes including cell differentiation [54]. *EGFEM1P* is a pseudogene located on chromosome three and was shown by one study to be differentially methylated in obese asthmatics, and by another to be significantly hypermethylated in patients with chronic lymphatic leukemia [55,56]. *KIAA1755* encodes for an uncharacterised protein, and contains a SNP variant (rs6127471) that has been associated with individuals who have increased heart rate [57,58]. *FAM65C* encodes a protein that

is a member of extracellular complex that generally regulates cellular processes in response to stimuli, but its main molecular function is still obscure [59].

The hypomethylated CpG site linked to *LOC730668*, which is a Dynein Heavy Chain-Like pseudogene located on chromosome 22, has been reported in two different studies to be differentially hypomethylated in individuals with temporal lobe epilepsy, and in individuals with psoriatic epidermis [60,61]. Studying genes without knowing how they correlate with chronological age could introduce false positives. Thus, if not adjusted, age could be a potential confounder in case-control studies. For example, a study conducted by Fluhr et al. [62] found *SLC12A8* (which was significantly hypermethylated with age in our study) to be differentially methylated in children with juvenile myelomonocytic leukemia (JMML). However, this study was based on children with JMML versus healthy adults, and the AR markers would be expected to be differentially methylated between children and adults regardless of JMML-status. Another example is the *TAL1* gene located on chromosome one, which encodes a protein that has been associated with Precursor T-Cell Acute Lymphoblastic Leukemia and T-Cell Childhood Acute Lymphocytic Leukemia. In a study conducted by Musialik et al. [63], methylation levels in the promotor of the *TAL1*gene were found to be slightly elevated in patients aged ≥ ten years with Precursor B-cell acute lymphoblastic leukemia, suggesting it was a potential predictor for the disease. Again, since methylation level was not adjusted for age, this association could be confounded by age.

Recently, the search for AR CpG sites and attempts to build DNAm-based age prediction models with high accuracy have been of major interest within the fields of forensic science, and epidemiology. For this reason, this study examined whether the EPIC BeadChip contains AR CpG markers with a better prediction accuracy than those found on the previous Illumina platforms (27K and 450K). Two methods were used to build age prediction models, elastic net regression and multivariate linear regression. The optimum model selected by elastic net regression contained a set of 425 CpG sites, 160 (38%) of them were probes that were newly-added to the EPIC BeadChip. This model had a high prediction accuracy, based on the validation set, of 2.6 years (MAD). Comparing this result with a study conducted by Hannum et al. (2013) that had a similar experimental design but used Illumina 450K data, their prediction model, also selected by elastic net regression, consisted of 71 CpG markers with a prediction accuracy of 4.89 years (MAD).

Building a prediction model for use in forensic investigations requires a small number of markers due to the minute quantities of DNA that is frequently recovered from forensic samples [64]. The six AR CpG sites selected by stepwise regression, which contained only one CpG marker that was newly added to the EPIC BeadChip, had a MAD value of 4.6 years based on the validation set. A review of the literature shows that the range of MAD values achieved by forensic researchers for models based on blood samples was 3.2 to 7.9 years, using two to 17 CpG markers [65-67]. Therefore, the prediction accuracy of data generated using the EPIC BeadChip falls within the MAD values reported in previous studies.

5 Conclusions

The purpose of the study presented here was to use blood-based EPIC BeadChip methylation data to identify AR CpG markers from probes that were new on this platform. We identified 52 AR CpG sites, 21 of which were novel AR CpG sites that mapped to 18 genes, nine of which have never been reported in the literature as being associated with age. This finding will provide new insights for researchers in both clinical and forensic epigenetics. For instance, in clinical epigenetics this will allow researchers to account for the aging effect of these genes, which will significantly limit the false positives in their genome- and epigenome-wide association studies. In addition, although the newly introduced probes on the EPIC BeadChip did not improve the age-prediction accuracy when compared to the other models in the literature, the identification of new genomic sites harboring AR CpG sites can be further studied by forensic geneticists using targeted bisulfite sequencing, which may result in the discovery of additional AR sites with high age prediction accuracy, that can be exploited for forensic purposes.

Conflict of interest statement

The authors declare that they have no conflict of interest.

References:

[1] M. Jung, G.P. Pfeifer, Aging and DNA methylation, BMC Biology. 13 (2015) 1. doi:10.1186/s12915-015-0118-4.

[2] C. López-Otín, M.A. Blasco, L. Partridge, M. Serrano, G. Kroemer, The Hallmarks of Aging, Cell. 153 (2013) 1194–1217. doi:10.1016/j.cell.2013.05.039.

[3] L.P. Breitling, K.-U. Saum, L. Perna, Ben Schöttker, B. Holleczek, H. Brenner, Frailty is associated with the epigenetic clock but not with telomere length in a German cohort, Clinical Epigenetics. 8 (2016) 21. doi:10.1186/s13148-016-0186-5.

[4] R.E. Marioni, S. Shah, A.F. McRae, S.J. Ritchie, G. Muniz-Terrera, S.E. Harris, et al., The epigenetic clock is correlated with physical and cognitive fitness in the Lothian Birth Cohort 1936, Int. J. Epidemiol. 44 (2015) 1388–1396. doi:10.1093/ije/dyu277.

[5] M.E. Levine, A.T. Lu, D.A. Bennett, S. Horvath, Epigenetic age of the pre-frontal cortex is associated with neuritic plaques, amyloid load, and Alzheimer's disease related cognitive functioning, Aging (Albany NY). 7 (2015) 1198–1211. doi:10.18632/aging.100864.

[6] B.H. Chen, R.E. Marioni, E. Colicino, M.J. Peters, C.K. Ward-Caviness, P.-C. Tsai, et al., DNA methylation-based measures of biological age: meta-analysis predicting time to death, Aging (Albany NY). 8 (2016) 1844–1865. doi:10.18632/aging.101020.

[7] W. Parson, Age Estimation with DNA: From Forensic DNA Fingerprinting to Forensic (Epi)Genomics: A Mini-Review, Ger. 64 (2018) 326–332. doi:10.1159/000486239.

[8] A. Vidaki, D. Ballard, A. Aliferi, T.H. Miller, L.P. Barron, D.S. Court, DNA methylation-based forensic age prediction using artificial neural networks and next generation sequencing, Forensic Science International: Genetics. 0 (2017) 225–236. doi:10.1016/j.fsigen.2017.02.009.

[9] A. Freire-Aradas, C. Phillips, M.V. Lareu, Forensic individual age estimation with DNA: From initial approaches to methylation tests, Forensic Science Rev. 29 (2017) 121.

[10] V.L. Wilson, P.A. Jones, DNA methylation decreases in aging but not in immortal cells, Science. 220 (1983) 1055–1057. doi:10.1126/science.6844925.

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

[12] H. Spiers, E. Hannon, S. Wells, B. Williams, C. Fernandes, J. Mill, Age-associated changes in DNA methylation across multiple tissues in an inbred mouse model,

Mechanisms of Ageing and Development. 154 (2016) 20–23. doi:10.1016/j.mad.2016.02.001.

[13] 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 Science International: Genetics. 29 (2017) 118–125. doi:10.1016/j.fsigen.2017.04.006.

[14] A. Aliferi, D. Ballard, M.D. Gallidabino, H. Thurtle, L. Barron, D. Syndercombe Court, DNA methylation-based age prediction using massively parallel sequencing data and multiple machine learning models, Forensic Science International: Genetics. 37 (2018) 215–226. doi:10.1016/j.fsigen.2018.09.003.

[15] V.K. Rakyan, T.A. Down, S. Maslau, T. Andrew, T.-P. Yang, H. Beyan, et al., Human aging-associated DNA hypermethylation occurs preferentially at bivalent chromatin domains, Genome Res. 20 (2010) 434–439. doi:10.1101/gr.103101.109.

[16] J.T. Bell, P.-C. Tsai, T.-P. Yang, R. Pidsley, J. Nisbet, D. Glass, et al., Epigenome-Wide Scans Identify Differentially Methylated Regions for Age and Age-Related Phenotypes in a Healthy Ageing Population, PLOS Genet. 8 (2012) e1002629.

[17] V.K. Rakyan, T.A. Down, N.P. Thorne, P. Flicek, E. Kulesha, S. Gräf, et al., An integrated resource for genome-wide identification and analysis of human tissue-specific differentially methylated regions (tDMRs), Genome Res. 18 (2008) 1518–1529. doi:10.1101/gr.077479.108.

[18] K.D. Pruitt, T. Tatusova, G.R. Brown, D.R. Maglott, NCBI Reference Sequences (RefSeq): current status, new features and genome annotation policy, Nucleic Acids Research. 40 (2012) D130–D135.

[19] J. Severin, A.M. Waterhouse, H. Kawaji, T. Lassmann, E. van Nimwegen, P.J. Balwierz, et al., FANTOM4 EdgeExpressDB: an integrated database of promoters, genes, microRNAs, expression dynamics and regulatory interactions, Genome Biology. 10 (2009) R39. doi:10.1186/gb-2009-10-4-r39.

[20] M. Bibikova, B. Barnes, C. Tsan, V. Ho, B. Klotzle, J.M. Le, et al., High density DNA methylation array with single CpG site resolution, Genomics. 98 (2011) 288–295. doi:10.1016/j.ygeno.2011.07.007.

[21] P. Garagnani, M.G. Bacalini, C. Pirazzini, D. Gori, C. Giuliani, D. Mari, et al., Methylation of ELOVL2 gene as a new epigenetic marker of age, Aging Cell. 11 (2012) 1132–1134. doi:10.1111/acel.12005.

[22] 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, Human Molecular Genetics. 23 (2014) 1186–1201. doi:10.1093/hmg/ddt531.

[23] S. Horvath, M. Gurven, M.E. Levine, B.C. Trumble, H. Kaplan, H. Allayee, et al., An epigenetic clock analysis of race/ethnicity, sex, and coronary heart disease, Genome Biology. 17 (2016) 171. doi:10.1186/s13059-016-1030-0.

[24] R. Zbieć-Piekarska, M. Spólnicka, T. Kupiec, A. Parys-Proszek, Ż. Makowska, A. Pałeczka, et al., Development of a forensically useful age prediction method based on DNA methylation analysis, Forensic Science International: Genetics. 17 (2015) 173–179. doi:10.1016/j.fsigen.2015.05.001.

[25] R. Zbieć-Piekarska, M. Spólnicka, T. Kupiec, Ż. Makowska, A. Spas, A. Parys-Proszek, et al., Examination of DNA methylation status of the ELOVL2 marker may be useful for human age prediction in forensic science, Forensic Science International: Genetics. 14 (2015) 161–167. doi:10.1016/j.fsigen.2014.10.002.

[26] S.-E. Jung, S.M. Lim, S.R. Hong, E.H. Lee, K.-J. Shin, H.Y. Lee, DNA methylation of the ELOVL2, FHL2, KLF14, C1orf132/MIR29B2C, and TRIM59 genes for age prediction from blood, saliva, and buccal swab samples, Forensic Science International: Genetics. 38 (2019) 1–8. doi:doi.org/10.1016/j.fsigen.2018.09.010.

[27] G. Hannum, J. Guinney, L. Zhao, L. Zhang, G. Hughes, S. Sadda, et al., Genome-wide methylation profiles reveal quantitative views of human aging rates, Molecular Cell. 49 (2013) 359–367. doi:10.1016/j.molcel.2012.10.016.

[28] L.M. McEwen, M.J. Jones, D.T.S. Lin, R.D. Edgar, L.T. Husquin, J.L. MacIsaac, et al., Systematic evaluation of DNA methylation age estimation with common preprocessing methods and the Infinium MethylationEPIC BeadChip array, Clinical Epigenetics. 10 (2018) 123. doi:10.1186/s13148-018-0556-2.

[29] L. Siggens, K. Ekwall, Epigenetics, chromatin and genome organization: recent advances from the ENCODE project, Journal of Internal Medicine. 276 (2014) 201–214. doi:10.1111/joim.12231.

[30] M. Lizio, J. Harshbarger, H. Shimoji, J. Severin, T. Kasukawa, S. Sahin, et al., Gateways to the FANTOM5 promoter level mammalian expression atlas, Genome Biology. 16 (2015) 22. doi:10.1186/s13059-014-0560-6.

[31] R. Pidsley, E. Zotenko, T.J. Peters, M.G. Lawrence, G.P. Risbridger, P. Molloy, et al., Critical evaluation of the Illumina MethylationEPIC BeadChip microarray for whole-genome DNA methylation profiling, Genome Biology. 17 (2016) 208.

[32] J. Dou, R.J. Schmidt, K.S. Benke, C. Newschaffer, I. Hertz-Picciotto, L.A. Croen, et al., Cord blood buffy coat DNA methylation is comparable to whole cord blood methylation, (2018) 1–10. doi:10.1080/15592294.2017.1417710.

[33] I. Zaimi, D. Pei, D.C. Koestler, C.J. Marsit, I. De Vivo, S.S. Tworoger, et al., Variation in DNA methylation of human blood over a 1-year period using the Illumina MethylationEPIC array, Epigenetics. 13 (2018) 1056–1071. doi:10.1080/15592294.2018.1530008.

[34] S.W. Curtis, D.O. Cobb, V. Kilaru, M.L. Terrell, E.M. Kennedy, M.E. Marder, et al., Exposure to polybrominated biphenyl (PBB) associates with genome-wide DNA methylation differences in peripheral blood, Epigenetics. 14 (2019) 52–66. doi:10.1080/15592294.2019.1565590.

[35] S. Davis, P.S. Meltzer, GEOquery: a bridge between the Gene Expression Omnibus (GEO) and BioConductor, Bioinformatics. 23 (2007) 1846–1847. doi:10.1093/bioinformatics/btm254.

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

[37] J. Fortin, T. Triche, K. Hansen, Preprocessing, normalization and integration of the Illumina HumanMethylationEPIC array with minfi | Bioinformatics | Oxford Academic, Bioinformatics. 33 (2017) 558–560.

606 [38] N. Touleimat, J. Tost, Complete pipeline for Infinium® Human Methylation 450K
607 BeadChip data processing using subset quantile normalization for accurate DNA
608 methylation estimation, Http://Dx.Doi.org/10.2217/Epi.12.21. 4 (2012) 325–341.
609 doi:10.2217/epi.12.21.
610 [39] A.E. Teschendorff, C.L. Relton, Statistical and integrative system-level analysis of DNA
611 methylation data, Nature Publishing Group. (1AD) 1–19. doi:10.1038/nrg.2017.86.
612 [40] W.E. Johnson, C. Li, A. Rabinovic, Adjusting batch effects in microarray expression data
613 using empirical Bayes methods, Biostatistics. 8 (2007) 118–127.
614 doi:10.1093/biostatistics/kxj037.
615 [41] Z. Sun, H. Chai, Y. Wu, W.M. White, K.V. Donkena, C.J. Klein, et al., Batch effect
616 correction for genome-wide methylation data with Illumina Infinium platform, BMC
617 Medical Genomics 2011 4:1. 4 (2011) 1. doi:10.1186/1755-8794-4-84.
618 [42] E.A. Houseman, W.P. Accomando, D.C. Koestler, B.C. Christensen, C.J. Marsit, H.H.
619 Nelson, et al., DNA methylation arrays as surrogate measures of cell mixture distribution,
620 BMC Bioinformatics 2010 11:1. 13 (2012) 86. doi:doi.org/10.1186/1471-2105-13-86.
621 [43] P. Du, X. Zhang, C.-C. Huang, N. Jafari, W.A. Kibbe, L. Hou, et al., Comparison of Beta-
622 value and M-value methods for quantifying methylation levels by microarray analysis,
623 BMC Bioinformatics 2010 11:1. 11 (2010) 587. doi:10.1186/1471-2105-11-587.
624 [44] S. Horvath, DNA methylation age of human tissues and cell types, Genome Biology. 14
625 (2013) R115. doi:10.1186/gb-2013-14-10-r115.
626 [45] C.I. Weidner, Q. Lin, C.M. Koch, L. Eisele, F. Beier, P. Ziegler, et al., Aging of blood can
627 be tracked by DNA methylation changes at just three CpG sites, Genome Biology. 15
628 (2014) R24. doi:10.1186/gb-2014-15-2-r24.
629 [46] C.M. Koch, W. Wagner, Epigenetic-aging-signature to determine age in different tissues,
630 Aging (Albany NY). 3 (2011) 1018–1027. doi:10.18632/aging.100395.
631 [47] C. Xu, H. Qu, G. Wang, B. Xie, Y. Shi, Y. Yang, et al., A novel strategy for forensic age
632 prediction by DNA methylation and support vector regression model, Scientific Reports.
633 5 (2015) 17788. doi:10.1038/srep17788.
634 [48] R.C. Slieker, C.L. Relton, T.R. Gaunt, P.E. Slagboom, B.T. Heijmans, Age-related DNA
635 methylation changes are tissue-specific with ELOVL2 promoter methylation as exception,
636 Epigenetics & Chromatin. (2018) 1–11. doi:10.1186/s13072-018-0191-3.

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

[50] Y. Liu, M.J. Aryee, L. Padyukov, M.D. Fallin, E. Hesselberg, A. Runarsson, et al., Epigenome-wide association data implicate DNA methylation as an intermediary of genetic risk in rheumatoid arthritis, Nature Biotechnology. 31 (2013) 142–147. doi:10.1038/nbt.2487.

[51] E.C. Davenport, V. Pendolino, G. Kontou, T.P. McGee, D.F. Sheehan, G. López-Doménech, et al., An Essential Role for the Tetraspanin LHFPL4 in the Cell-Type-Specific Targeting and Clustering of Synaptic GABAA Receptors, Cell Rep. 21 (2017) 70–83. doi:10.1016/j.celrep.2017.09.025.

[52] S.S. Wang, D.J. Smiraglia, Y.-Z. Wu, S. Ghosh, J.S. Rader, K.R. Cho, et al., Identification of novel methylation markers in cervical cancer using restriction landmark genomic scanning, Cancer Res. 68 (2008) 2489–2497. doi:10.1158/0008-5472.CAN-07-3194.

[53] N. Patel, T. Itakura, J.M. Gonzalez, S.G. Schwartz, M.E. Fini, GPR158, an orphan member of G protein-coupled receptor Family C: glucocorticoid-stimulated expression and novel nuclear role, Plos One. 8 (2013) e57843. doi:10.1371/journal.pone.0057843.

[54] T. Musikacharoen, K. Bandow, K. Kakimoto, J. Kusuyama, T. Onishi, Y. Yoshikai, et al., Functional involvement of dual specificity phosphatase 16 (DUSP16), a c-Jun N-terminal kinase-specific phosphatase, in the regulation of T helper cell differentiation, The Journal of Biological Chemistry. 286 (2011) 24896–24905. doi:10.1074/jbc.M111.245019.

[55] D. Rastogi, M. Suzuki, J.M. Greally, Differential epigenome-wide DNA methylation patterns in childhood obesity-associated asthma, Scientific Reports. 3 (2013) 2164. doi:10.1038/srep02164.

[56] C. Baer, R. Claus, L.P. Frenzel, M. Zucknick, Y.J. Park, L. Gu, et al., Extensive promoter DNA hypermethylation and hypomethylation is associated with aberrant microRNA expression in chronic lymphocytic leukemia, Cancer Res. 72 (2012) 3775–3785. doi:10.1158/0008-5472.CAN-12-0803.

[57] M. den Hoed, M. Eijgelsheim, T. Esko, B.J.J.M. Brundel, D.S. Peal, D.M. Evans, et al., Identification of heart rate–associated loci and their effects on cardiac conduction and rhythm disorders, Nature Genetics. 45 (2013) 621–631.

[58] L. Weinhold, S. Wahl, M. Schmid, A Statistical Model for the Analysis of Beta Values in DNA Methylation Studies, (2016).

[59] W. Sun, T. He, C. Qin, K. Qiu, X. Zhang, Y. Luo, et al., A potential regulatory network underlying distinct fate commitment of myogenic and adipogenic cells in skeletal muscle, Scientific Reports. 7 (2017) 11G.

[60] S.F.C. Miller-Delaney, K. Bryan, S. Das, R.C. McKiernan, I.M. Bray, J.P. Reynolds, et al., Differential DNA methylation profiles of coding and non-coding genes define hippocampal sclerosis in human temporal lobe epilepsy, Brain. 138 (2015) 616–631. doi:10.1093/brain/awu373.

[61] D. Verma, A.-K. Ekman, C. Bivik Eding, C. Enerbäck, Genome-Wide DNA Methylation Profiling Identifies Differential Methylation in Uninvolved Psoriatic Epidermis, J. Invest. Dermatol. 138 (2018) 1088–1093. doi:10.1016/j.jid.2017.11.036.

[62] S. Fluhr, M. Boerries, H. Busch, A. Symeonidi, T. Witte, D.B. Lipka, et al., CREBBP is a target of epigenetic, but not genetic, modification in juvenile myelomonocytic leukemia, Clinical Epigenetics. 8 (2016) 50. doi:10.1186/s13148-016-0216-3.

[63] E. Musialik, M. Bujko, P. Kober, A. Wypych, K. Gawle-Krawczyk, M. Matysiak, et al., Promoter methylation and expression levels of selected hematopoietic genes in pediatric B-cell acute lymphoblastic leukemia, Blood Res. 50 (2015) 26–32. doi:10.5045/br.2015.50.1.26.

[64] A. Vidaki, B. Daniel, D.S. Court, Forensic DNA methylation profiling—Potential opportunities and challenges, Forensic Science International: Genetics. 7 (2013) 499–507. doi:10.1016/j.fsigen.2013.05.004.

[65] J.-L. Park, J.H. Kim, E. Seo, D.H. Bae, S.-Y. Kim, H.-C. Lee, et al., Identification and evaluation of age-correlated DNA methylation markers for forensic use, Forensic Science International: Genetics. 23 (2016) 64–70. doi:10.1016/j.fsigen.2016.03.005.

[66] 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 Science International: Genetics. 17 (2015) 129–136. doi:10.1016/j.fsigen.2015.05.007.

[67] J. Naue, H.C.J. Hoefsloot, O.R.F. Mook, L. Rijlaarsdam-Hoekstra, M.C.H. van der Zwalm, P. Henneman, et al., Chronological age prediction based on DNA methylation:

Massive parallel sequencing and random forest regression, Forensic Science International: Genetics. 31 (2017) 19–28. doi:10.1016/j.fsigen.2017.07.015.