DOI:
10.1016/j.fsigen.2021.102637

Document Version
Peer reviewed version

Link to publication record in King's Research Portal

Citation for published version (APA):
Aliferi, A., Sundaram, S., Ballard, D., Freire-Aradas, A., Phillips, C., Lareu, M. V., & Court, D. S. (2022). Combining current knowledge on DNA methylation-based age estimation towards the development of a superior forensic DNA intelligence tool. *Forensic Science International: Genetics*, *57*, [102637]. https://doi.org/10.1016/j.fsigen.2021.102637

Combining current knowledge on DNA methylation-based age estimation towards the development of a superior forensic DNA intelligence tool

Anastasia Aliferi[a], Sudha Sundaram[a], David Ballard[a1], Ana Freire-Aradas[b], Christopher Phillips[b], Maria Victoria Lareu[b], Denise Syndercombe Court[a]

[a]King's Forensics, Department of Analytical, Environmental and Forensic Sciences, Faculty of Life Sciences and Medicine, King's College London, London, United Kingdom

[b]Forensic Genetics Unit, Institute of Forensic Sciences, University of Santiago de Compostela, Galicia, Spain

[1]Corresponding author: Dr David Ballard, Senior Lecturer, Department of Analytical, Environmental and Forensic Sciences, Faculty of Life Sciences and Medicine, King's College London, Franklin-Wilkins Building, 150 Stamford Street, London, SE1 9NH, david.ballard@kcl.ac.uk

Abstract

The estimation of chronological age from biological fluids has been an important quest for forensic scientists worldwide, with recent approaches exploiting the variability of DNA methylation patterns with age in order to develop the next generation of forensic 'DNA intelligence' tools for this application. Drawing from the conclusions of previous work utilising massively parallel sequencing (MPS) for this analysis, this work introduces a DNA methylation-based age estimation method for blood that exhibits the best combination of prediction accuracy and sensitivity reported to date. Statistical evaluation of markers from 51 studies using microarray data from over 4,000 individuals, followed by validation using in-house generated MPS data, revealed a final set of 11 markers with the greatest potential for accurate age estimation from minimal DNA material. Utilising an algorithm based on support vector machines, the proposed model achieved an average error (MAE) of 3.3 years, with this level of accuracy retained down to 5 ng of starting DNA input (~1 ng PCR input). The accuracy of the model was retained (MAE=3.8 years) in a separate test set of 88 samples of Spanish origin, while predictions for donors of greater forensic interest (<55 years of age) displayed even higher accuracy (MAE=2.6 years). Finally, no sex-related bias was observed for this model, while there were also no signs of variation observed between control and disease-associated populations for schizophrenia, rheumatoid arthritis, frontal temporal dementia and progressive supranuclear palsy in microarray data relating to the 11 markers.

Highlights

- Evaluation of methylation age markers using microarray data and targeted sequencing revealed a set of 11 'optimal' markers
- The prediction model showed high prediction accuracy in both a UK (MAE=3.3 years) and Spanish sample cohort (MAE=3.8 years)
- Prediction accuracy improved for under 55-year-olds (MAE=2.6), with 81% predicting with an error of less than 4 years
- The accuracy of DNA methylation quantification and age prediction was retained down to 5ng of DNA input (~1ng in PCR stage)

Keywords: age prediction, DNA methylation, machine learning, forensic, DNA intelligence

1 Introduction

A key aspect of forensic science research is the inference of information regarding a person's visible appearance, geographical origin and age using biological stains recovered from crime scenes. This information, commonly referred to as 'DNA intelligence', can provide law enforcement organisations with leads for investigations, taking on the role of a 'biological witness'. Following the successful implementation of DNA-based methods for the inference of ancestry and phenotype (e.g. eye, hair, and skin colour) in forensic investigations, the focus of DNA intelligence research has recently shifted towards the accurate prediction of chronological age. Whilst multiple biomarkers, including protein and nucleic acid-based candidates, have been trialled for use in age estimation, recent studies have focused on the correlation between chronological age and methylation status at certain cytosine residues present in the human genome. Since methods for DNA methylation-based age prediction made their debut in forensic science in 2014[1], a significant amount of research has focused on forensically-relevant tissues as well as targeted sequencing technologies, that offer high potential for sensitivity and are more accessible to forensic laboratories than high-cost genome-wide analysis. However, while DNA methylation-based age prediction rose to become one of the priorities for forensic researchers worldwide, a consensus on the most informative marker sets has yet to be reached.

Despite the domination of targeted sequencing in recent literature on age estimation, *de novo* marker discovery and evaluation are still highly dependent on microarray data available in online depositories. However, the use of such data does not come without challenge, with the presence of batch effects being one of the biggest issues. Batch effects observed between different methylation analysis platforms, as well as between different datasets developed using the same technology, have been shown to introduce bias when comparing data derived from multiple studies [2-4]. In efforts to account for known and unknown batch effects in the Illumina methylation microarray platforms, multiple normalisation packages have been developed, as previously outlined by Dedeurwaerder *et al.* [5]. However, whilst the effect of some of these normalisation approaches can be beneficial for within-array normalisation, the available between-array normalisation methods have proven unsuitable for the Illumina arrays, producing no significant benefits [5]. In addition, large scale transformations of methylation data have been shown to result in an overall decline of data quality, often masking directional methylation patterns [5, 6]. Furthermore, the developed normalisation algorithms can only be applied to raw microarray data, not provided for most of the publicly available datasets, thus significantly limiting the number of available samples. On the other hand, whilst advanced normalisation can be crucial for training prediction algorithms, as batch effects present both within and between arrays could be interpreted as true variation and prevent the algorithm from identifying age-related patterns, its importance significantly decreases when microarray data is used for assessing correlation and identifying potential markers. In such cases, validation of the proposed marker sets using targeted methods and subsequent use of data solely deriving from this targeted analysis for the development of prediction models, can balance out the lack of extensive normalisation in the marker discovery stage.

This stage has, so far, been based almost exclusively on the interrogation of the observed correlation between age and methylation for the different CpGs, usually according to Pearson's or Spearman's correlation coefficient. However, neither of these measures considers the range of methylation over the human lifespan. Whilst not immediately obvious, the importance of this range becomes evident when addressing the issue of sensitivity, which remains one of the most important factors hindering the wider application of DNA methylation-based age prediction in forensic casework. Whilst non-binomial nature of CpG methylation, that represents a percentage, introduces a significant challenge, markers showing large differences between the different age groups can potentially allow for a certain loss of accuracy during the quantification of DNA methylation, offering an 'escape' from the 1000 sequencing reads limit per marker and sample, that has previously been set for this type of methods [7]. The fact that larger methylation ranges allow for higher method accuracy overall is evident in the success of CpG markers relating to the *ELOVL2* gene. Since their discovery, the *ELOVL2* markers have been incorporated in almost every DNA methylation-based age prediction method, while successful age estimation models have been also developed on *ELOVL2* CpGs alone [8]. Looking at the characteristics of these markers, what sets them apart is the combination of high correlation with age and large methylation range over the human lifespan, rather than correlation alone. This indicates that the inclusion of methylation range as a factor during marker selection can increase the potential of a DNA methylation-based age prediction method in terms of its success with samples of low DNA content.

In addition to correlation with age and sensitivity, another thing that needs to be considered when developing forensic age estimation tools is the potential association of the utilised biomarkers with factors other than age. Since their discovery, DNA methylation biomarkers have been widely investigated in medical studies for their association with medical conditions, infections and diseases such as cancer [9], Alzheimer's disease and dementia [10-12], Huntington's disease [13], Parkinson's disease [14], Hutchinson Gilford progeria [15] and Werner syndrome [16].These associations, together with indications of correlation between DNA methylation biomarkers and smoking [17-19], body mass index [20, 21] and socioeconomic status and education [22-24], paved the path for the emergence of the term 'epigenetic age', the distance of which from chronological age has been proposed as a measure of 'biological age' [25].

Biological age, also referred to as functional, physiological or phenotypic age, has been the focus on many recent studies aiming to provide a measure of 'health' and life expectancy through the analysis of DNA methylation [22, 26]. Interestingly, it was also the estimation of biological rather than chronological age that motivated Horvath's work on the 'human epigenetic clock' [3], even though some of the 353 markers proposed in this study have been widely used for the estimation of chronological age in further studies [27, 28]. Whilst the high correlation between these two 'ages' has often blurred the lines between the terms, it is important to address them separately, especially in a forensic setting.

Forensic science often deals with samples for which there is little or no information regarding the donor. Furthermore, strict ethical guidelines apply for the inference of intelligence-related information from human samples, in order to safeguard human privacy and wellbeing. These facts highlight the need to address potential biases in forensic DNA methylation-based age prediction, that could also result in significant inaccuracies.

Drawing upon the recent literature, this work aims to take the first step towards reaching a much-needed consensus in terms of the most informative and sensitive markers for DNA methylation-based age prediction in forensics. Using independent microarray datasets addressing a total of over 4,000 samples, candidate markers were assessed on both their correlation and methylation range, providing a marker selection that was further validated by targeted sequencing using a separate sample cohort. Furthermore, this work represents one the first attempts to scrutinizing forensic DNA methylation-based age prediction markers in terms of their association with sex and disease on both a CpG and gene/protein level.

2 Materials and methods

2.1 Compilation of CpG sites associated with age

A systematic review of the available literature up to 2017 was conducted to identify CpG markers exhibiting methylation patterns associated with chronological age in human samples. In cases where studies investigating large marker sets have provided a marker sub-selection that reveal superior correlation with age, only these most informative subsets were included in the analysis. A comprehensive list of the 51 studies [1, 3, 6, 8, 20, 21, 29-73] can be found in Supplementary_Table_S1.

Following this analysis, a total of 36,137 CpG candidates were identified as potential biomarkers of aging. A subset of 5,364 CpGs, independently validated in at least 2 of the 51 studies or previously included in age-prediction algorithms were selected for further analysis.

2.2 Collection of methylation data from publicly available datasets

Methylation data for the 5,364 CpG candidates was extracted from datasets available in the public repository of the National Centre for Biotechnology Information Gene Expression Omnibus (NCBI GEO, [74]). The R Project for Statistical Computing software in combination with the R Studio platform was employed for this analysis. The 24 datasets used for this analysis had originally been developed using blood samples (including whole blood, blood leukocytes and blood lymphocytes samples) analysed with the Illumina microarray technology (including both the Illumina Infinium HumanMethylation 27K and 450K BeadChip arrays) [21, 30, 31, 36, 37, 39, 43, 45, 50, 59, 75-81]. For studies investigating the methylome of diseased individuals, only data for the control samples was collected at this stage. More information on these datasets is detailed in Supplementary_Table_S2.

As the two Illumina arrays offer different levels of coverage, samples analysed with the 27K array contained information for 1702 out of the 5364 CpGs, whilst those analysed with the 450K array provided with values for 5317 sites. Furthermore, in the 450K data 7 CpG sites containing missing values for 600-1300 samples were removed from the analysis of this array and, as 2 of these CpGs were only available in the 450K array, the overall number of analysed CpGs was subsequently reduced to 5362. Finally, for samples with obvious familial relationships, such as twins or triplets, only one member of the relationship was retained in the dataset in order to avoid bias deriving from genetic similarities unrelated to age.

2.3 Data normalisation

Data for the two different platforms was analysed separately in order to avoid potential bias introduced by different sample sizes between the unique probes for 27K, the unique probes for 450K and the overlapping probes. Methylation data was extracted in the form of β-values (representing the percentage of methylation for a specific CpG site), but for the purposes of correlation analysis these were subsequently converted to M-values following the equation:

$$Mi = log_2\,(\frac{\beta i}{1-\beta i})$$

where Mi represents the M-value for a certain marker in a specific sample and βi represents the equivalent β-value. Whilst β-values have a direct biological meaning and were employed for addressing the methylation range of the different CpGs over the human lifespan, it has been shown that M-values are more appropriate for statistical analysis purposes as they are much more homoscedastic [2, 82].

Given the lack of consensus in the literature regarding the normalisation of microarray data and the fact that normalisation packages require raw microarray data that are not provided for most of the publicly available datasets, none of the previously developed normalisation packages were used in this study. As an alternative, methylation M-values were centred around the overall median value for each platform (27K or 450K) according to the equation:

$$Mi\ centred = Mi - median\ M\ value\ for\ the\ platform\ (27K\ or\ 450K)$$

where Mi represents the M-value for a certain marker in a specific sample and median M represents the median M-value for all samples for this marker in the relevant platform.

2.4 Marker evaluation

Using the normalised methylation data, a further shortlisting of markers was performed in order to identify a subset exhibiting the highest correlation with age (Pearson’s correlation coefficient r) and largest methylation range over the human lifespan (β-value range), while also maintaining functionality for a targeted sequencing approach based on multiplexing (ideally under 20 markers). In order to achieve this, an original subset of 244 markers with |r|≥0.70, or |r|≥0.65 and methylation range above 70% over the human lifespan in either the 27K or 450K dataset (Supplementary_Table_S3), was further reduced to 24 markers with |r|>0.70 and overall methylation range above 60%. Finally, following additional examination of the correlation plots that revealed ‘tailing’ of the data in the younger ages in 5 markers that, when taken into account, reduced the methylation range, this was reduced to a final set of 19 markers (Table 1).

Table 1. Chromosomal location (GRC37/hg19) and genetic information on the 19 selected markers. The Pearson's correlation (R) calculated from the 450K and 27K data, as well as the absolute range of beta values observed for the relevant markers over the different ages are also displayed.

CpG site	Chromosomal location	Associated Gene name	Pearson's correlation (r) in 450K data (n=2976)	Pearson's correlation (r) in 27K data (n=1299)	Beta value range
cg16867657	6:11044877	*ELOVL2*	0.9080	N.A.	0.7507
cg22454769	2:106015767	*FHL2*	0.8713	N.A.	0.8346
cg10501210	1:207997020	*MIR29B2CHG (C1orf132)*	-0.8403	N.A.	0.8957
cg19283806	18:66389420	*CCDC102B*	-0.8265	N.A.	0.8744
cg06639320	2:106015739	*FHL2*	0.8103	N.A.	0.7450
cg24079702	2:106015771	*FHL2*	0.8029	N.A.	0.7767
cg00329615	3:118706648	*IGSF11*	-0.8008	N.A.	0.6844
cg24724428	6:11044888	*ELOVL2, ELOVL2-AS1*	0.7973	N.A.	0.6403
cg21572722	6:11044894	*ELOVL2*	0.7970	N.A.	0.6226
cg09809672	1:236557682	*EDARADD*	-0.7877	-0.8091	0.7942
cg07553761	3:160167977	*SMC4, TRIM59*	0.7847	N.A.	0.9193
cg22796704	10:49673534	*ARHGAP22*	-0.7712	N.A.	0.6016
cg08128734	1:206685423	*RASSF5*	-0.7619	N.A.	0.6873
cg17372101	7:147500722	*CNTNAP2*	-0.7615	N.A.	0.6772
cg18618815	17:48275324	*COL1A1*	-0.7590	N.A.	0.6928
cg08160331	11:75140865	*KLHL35*	0.7571	N.A.	0.6999
cg08262002	4:16575323	*LDB2*	-0.7565	N.A.	0.6576
cg12934382	3:51741135	*GRM2*	0.7559	N.A.	0.7990
cg17471102	19:5851255	*FUT3*	-0.7546	-0.7283	0.6109

2.5 Sample collection and preparation

Collection of tissues for the purposes of this study was conducted under ethical approval granted by the Biomedical Sciences, Dentistry, Medicine and Natural & Mathematical Sciences Research Ethics Subcommittee (BDM/13/14-30). Whole blood samples were collected from 112 unrelated volunteers, aged between 11 and 92.9 years, through venepuncture performed by a trained phlebotomist. Prior to sampling, full informed consent regarding the analysis was acquired from the donors, or their parents or legal guardians for the cases of under-aged individuals (<18 years). No information on medical history was collected during this process in an attempt to create an inclusive, unbiased dataset, representative of the general population. Samples were stored at 4^{o}C.

Additionally, a set of 88 DNA extracts from whole blood samples deriving from adults (19-99 years old) obtained from the 'Carlos III' Spanish National DNA Bank, University of Salamanca, under ethical approval granted by the ethics committee of investigation in Galicia, Spain (CAEI: 2013/543), were shared by the Forensic Genetics unit of University of Santiago de Compostela (USC, Spain).

2.6 DNA methylation standards

Premixed standards of known methylation were purchased from EpigenDx (Massachusetts, USA) for methylation levels of 0%, 5%, 10%, 25%, 50%, 75% and 100% at concentration of 50 ng/μL.

2.7 DNA extraction and quantification

Genomic DNA extractions were carried out using a BioRobot EZ1 automated purification instrument (Qiagen, Hilden, Germany) in combination with the EZ1 Blood kit (Qiagen, Hilden, Germany). Following extraction, DNA samples were stored at -20°C. Quantification of DNA extracts was conducted using the Quantifiler Trio DNA Quantification kit in combination with the ABI PRISM® 7500 Sequence Detection System, both produced by Thermo-Fisher Scientific (Massachusetts, USA). The manufacturer's guidelines [83] were followed throughout the protocol in half volumes and all samples were quantified in duplicate.

2.8 Sodium bisulphite conversion

Treatment with sodium bisulphite was employed for the conversion of unmethylated cytosines to uracils in the DNA samples. A total of 50ng of DNA from each sample or standard was converted using the MethylEdge Bisulphite Conversion System (Promega Corporation, Wisconsin, USA) and the treated DNA was eluted in 10μL of the elution buffer provided according to the manufacturer's specifications [84]. Eluates were processed immediately (see next session). The approximate recovery of DNA following bisulphite conversion using this chemistry has been calculated as 52% [85] and therefore the final concentration of the eluate was estimated at approximately 2.6 ng/μL.

2.9 Amplification of the bisulphite-converted DNA

Primers for this study were designed using the MethPrimer online software [86] for bisulphite-sequencing PCR based on the GRCh37/hg19 human genome (Ensembl genome browser [87]). Individual primer pairs were designed for each CpG of interest, with the exception of cg16867657, cg21572722, cg24724428 and cg06639320, cg22454769, cg24079702 that are located in close proximity inside the regulatory regions of *ELOVL2* and *FHL2* respectively and thus could be interrogated in the same amplicons. Furthermore, as the high abundance of CpG sites in the *ELOVL2* regulatory region complicates primer design, two previously published primer pairs were tested [8, 88]. The primers suggested by Zbieć-Piekarska *et al*. [8] were selected as they exhibited lower amplification bias, but instead of the misalignment employed in the original design to account for the CpG in the primer location, a wobble site (equimolar mix of pyrimidines) was included for that location as suggested by Naue *et al.* [88] in their design. More information on the primers can be found in Supplementary_Table_S4).

Optimum annealing temperature for each primer set was determined by analysing singleplex reactions for each pair at different annealing temperatures using agarose gel electrophoresis. Primers for cg12934382 (*GRM2*) failed to provide amplification products at this point and were therefore excluded from further analysis. Following this analysis, primers were combined in two multiplex reactions using the Qiagen Multiplex PCR kit (Qiagen, Hilden, Germany) for both reactions in half volume (25 μL). Each reaction comprised of 12.5 μL of 2x Qiagen Multiplex PCR Master Mix (providing a concentration of 3 mM $MgCl_2$), an additional 1 μL of 25 mM $MgCl_2$ solution for a final concentration of 4 mM, 2μL (~5 ng) of bisulphite treated DNA or calibration standard and 9.5 μL of primer mix. The final concentration of primers in the two multiplex reactions ranged from 0.08 to 0.7 μM depending on the efficiency of the primers (Table 2*)*. The reaction conditions were: (1) 95°C for 15min, (2) 32 cycles consisting of 94°C for 30s, Tm (see Table 2) for 30s and 72°C for 30s, (3) 72°C for 4min followed by a hold at 4°C.

Table 2. Details on the multiplex reactions employed in this study.

CpG	Associated Genes	Primer concentration in PCR (μM)	Annealing temperature
cg16867657	*ELOVL2*	0.7	59ºC
cg21572722			
cg24724428			
cg06639320	*FHL2*	0.4	
cg22454769			
cg24079702			
cg22796704	*ARHGAP22*	0.1	
cg17372101	*CNTNAP2*	0.2	
cg19283806	*CCDC102B*	0.5	
cg07553761	*SMC4, TRIM59*	0.2	
cg08262002	*LDB2*	0.08	
cg17471102	*FUT3*	0.3	
cg18618815	*COL1A1*	0.7	
cg00329615	*IGSF11*	0.2	56ºC
cg08128734	*RASSF5*	0.2	
cg10501210	*MIR29B2CHG (C1orf132)*	0.6	
cg09809672	*EDARADD*	0.1	
cg08160331	*KLHL35*	0.4	

2.10 Post-PCR Purification and Quantification

Following amplification, samples were purified using the MinElute PCR Purification kit (Qiagen, Hilden, Germany) in order to remove unincorporated primer residues [89]. Elution was performed in 11 μL PCR-grade water. Prior to library preparation all samples were quantified using the Qubit dsDNA HS Assay kit (ThermoFisher, Massachusetts, USA) according to the manufacturer's guidelines [90] and in combination with the Qubit 2.0 Fluorometer instrument and clear thin-walled 0.5 mL PCR tubes.

2.11 Library preparation and quantification

The preparation of sequencing libraries was performed with the NEBNext Ultra II DNA Library Prep Kit for Illumina (New England BioLabs, Massachusetts, USA), starting with 50 ng of purified PCR product per sample. Library preparation was performed according to the manufacturer's specifications [91] in half volumes, while the size selection steps were performed as per the KAPA Hyper Prep protocol [92]. For the size selection stages, AMPure XP Beads (Beckman Coulter Genomics, California, USA) and Illumina Resuspension Buffer (Illumina, California, USA) were used. Finally, library amplification was performed for 8 cycles (up to 15 cycles can be used at this stage according to the NEBNext Ultra II protocol).

Quantification of the libraries was conducted with the KAPA Library Quantification Kit for Illumina platforms (Roche, Basel, Switzerland) [93]. Libraries were diluted 1:100,000 in PCR-grade water prior to quantification and analysed in duplicate. Following quantification, DNA libraries were normalised to 20 nM using Tris-HCL 10 mM/pH 8.5 with 0.1% Tween (EBT buffer) and were pooled together in equal amounts to a final volume of 240 μL (for a typical 24-samples run). Following denaturation and dilution to 10 pM, 500 μL of library was mixed with 100 μL of denatured 20 pM PhiX control (Illumina, CA) and loaded in the MiSeqFGx instrument (Illumina, California, USA) using the MiSeq version 2 (300 cycles) cartridge and reagents.

2.12 Sequencing

Sequencing of the libraries was performed using the Illumina MiSeqFGx benchtop instrument (Illumina, California, USA). Sample sheets and sample plates were created in the Illumina Experiment Manager software and the instrument was set to perform paired-end sequencing of 201-101 bp for the forward and reverse directions, while the analysis workflow was set to ‘FASTQ only’. The online platform Basespace (https://euc1.sh.basespace.illumina.com) was used for monitoring the performance of the runs as well as retrieve the sequencing files.

2.13 Data analysis and normalisation

Analysis of the FASTQ files was conducted with the Burrows-Wheeler Aligner (BWA) [94], Sequence Alignment/Map (SAMtools) [95], and Genome Analysis Toolkit (GATK, Broad Institute, Massachusetts, USA) [96] software. Reads were aligned to a custom genome containing only the 18 (cg12934382 (*GRM2*) was removed from the analysis as primers failed to yield products) amplicon sequences, where all non-CpG cytosines were replaced by thymines. For CpG positions, information was collected for the presence of both cytosines and thymines. Files were exported in variant call format (VCF) using GATK and data was subsequently extracted from these files with the R Project for Statistical Computing software in combination with R Studio platform and were finally processed with Microsoft Office Excel software. The methylation percentage (β-values) for the 18 targeted CpGs was calculated by comparing the number of cytosine reads (suggesting the presence of methylation) to the combined total of cytosine and thymine (suggesting the absence of methylation) reads at each CpG. A similar analysis was carried out for all non-CpG cytosine sites in each amplicon in order to establish the conversion efficiency of the bisulphite treatment. Non-CpG cytosines are expected to be free of methylation [97, 98] and therefore should be converted to uracils and subsequently to thymines following bisulphite treatment and amplification. Any cytosines therefore detected in those positions were indicative of incomplete conversion and the methylation percentages for the relevant CpGs were corrected according to the formula:

$$\begin{aligned}&Corrected\ methylation\ value\ for\ CpGi\\ &\qquad = 1-\left(\frac{(1-CpGi\ Methylation\ Value)}{Amplicon\ Conversion\ Rate}\right)\end{aligned}$$

where CpGi corresponds to a specific marker, and the amplicon conversion rate corresponds to the percentage of non-CpG cytosines successfully converted in the relevant amplicon. For blood samples analysed in duplicate, average methylation values between duplicates was calculated based on the number of sequencing reads for each duplicate and each marker, where the methylation value of the duplicate with the higher number of sequencing reads contributed accordingly high to the final methylation score for the relevant marker following the equation:

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

Where CpGi corresponds to a specific marker and a and b correspond to the two replicates of the specific sample. Prior to statistical analysis and modelling, methylation β-values were converted to M-values as previously described (see section 2.3). Finally, the entire dataset was subsequently normalised by centring of the M-values around the median M-value according to the equation:

$$Mi\ centred = Mi - median\ M\ value\ for\ the\ dataset$$

where M$_i$ represents the M-value for a certain marker in a specific sample and median M represents the median M-value for all dataset samples for this marker.

2.14 Marker elimination and age prediction

Final marker elimination was performed based on the in-house developed dataset (n=112). Using the R project for statistical computing software version 3.3.3 [99] in combination with the *caret* package [100], CpG selection was based on the results obtained from 8 independent algorithms assessing marker informativeness. These included forward selection, backward elimination, Boruta, 2 separate genetic algorithms (one of 10 iterations and one with 200 iterations), as well as LASSO, ridge and elastic net regression. These algorithms were used for assessing which CpG markers (variables) or marker sets were most useful in age estimation, with their results taking the form of suggested CpG subsets performing best for age estimation and/or ranking of the individual markers. Briefly, forward selection and backwards elimination produced subsets of 'most important' CpGs for age prediction selected through stepwise regression, Boruta produced a CpG ranking from most to least informative in regard to age through random forest regression, the genetic algorithms produced sets of 'fittest' CpG predictors using an algorithm that mimics the theory of natural selection and the three regression algorithms, LASSO, ridge and elastic net, defined subsets of most important CpG age predictors, while also assigning scores indicating the 'importance' of each individual CpG in age estimation.

Analysis of the results produced by these marker selection algorithms, revealed a subset of 11 markers that scored highly on all occasions. These markers were cg21572722 (*ELOVL2*), cg24724428 (*ELOVL2*), cg06639320 (*FHL2*), cg09809672 (*EDARADD*), cg22796704 (*ARHGAP22*), cg08128734 (*RASSF5*), cg17372101 (*CNTNAP2*), cg10501210 (*MIR29B2CHG*), cg19283806 (*CCDC102B*), cg07553761 (*SMC4*, *TRIM59*) and cg08262002 (*LDB2*).

Following a split of the dataset into training (n=77) and validation (i.e. blind, n=35) sets, two support vector machine models with polynomial function (SVMp) were trained simultaneously for all 18 markers and for the selection of 11 markers. The two models were assessed based on both the absolute prediction error (MAE) and root mean square error (RMSE) of the test set.

In cases where samples failed to obtain reads for certain markers in the sensitivity experiment, an imputation of the missing values was performed based on K nearest neighbours.

2.15 Sequencing adapter-tagged primers

Following the formation of the 11-CpG marker set, the 10 primer pairs relating to these CpGs were re-designed in order to include the adaptor sequences used for the MiSeq platform. This re-design was performed in order to reduce the number of steps required for library preparation, allowing for reduced processing time, elimination of adaptor dimer formation issues and removal of one of the two clean-up steps that are associated with loss of product. This process included the addition of the relevant sequences in the 5' end of the forward (ACACTCTTTCCCTACACGACGCTCTTCCGATCT) and reverse (GACTGGAGTTCAGACGTGTGCTCTTCCGATCT) primers. Primer concentrations in the protocol were adjusted based on the amplification efficiency of the new primers (Table 3), whilst amplification conditions remained the same.

Table 3. Details on the multiplex reactions for the final 11 markers, using the sequencing adapter-tagged primers.

CpG	Associated Genes	Primer concentration in PCR (µM)	Annealing temperature
cg21572722	*ELOVL2*	0.7	59°C
cg24724428			
cg06639320	*FHL2*	0.4	
cg22796704	*ARHGAP22*	0.08	
cg07553761	*SMC4, TRIM59*	0.1	
cg19283806	*CCDC102B*	0.04	
cg17372101	*CNTNAP2*	0.04	
cg08262002	*LDB2*	0.08	
cg08128734	*RASSF5*	0.7	56°C
cg10501210	*MIR29B2CHG (C1orf132)*	0.5	
cg09809672	*EDARADD*	0.4	

As these primers were pre-tagged with the adaptor sequence, the first steps of the NEB Next Ultra II library preparation protocol, including end prep and adaptor ligation, were subsequently omitted.

2.16 Sex association

Following marker selection and method development, the need to conduct more extensive validation and address potential issues that can hinder the wider application of this method was identified. The first such issue investigated was that of potential bias introduced by the sex of the donors. Firstly, methylation data collected from the analysis of blood samples obtained from 107 out of the 112 unrelated volunteers was also employed for this analysis (for the remaining 5 samples data on sex was not available). Furthermore, given the limited number of samples in the targeted sequencing dataset, methylation data previously collected for the age markers from 14 studies conducted on the Illumina Infinium HumanMethylation 450K BeadChip technology were also utilised (Supplementary_Table_S5). This data was selected over that from the HumanMethylation 27K BeadChip due to the larger number of samples and more balanced ratio between male (n=1311) and female (n=1433) donors.

2.17 Disease association using publicly available datasets

Similarly, investigation of potential bias introduced in DNA methylation-based age estimation due to disease status was again conducted using methylation data collected from studies conducted with the Illumina Infinium HumanMethylation 450K BeadChip technology. This data derives from the non-control samples of studies previously used for the evaluation of age markers and relates to the conditions of schizophrenia (n=62) [37], rheumatoid arthritis (n=354) [78], frontal temporal dementia (FTD) (n=121) and progressive supranuclear palsy (PSP) (n=42) [80] (Supplementary_Table_S6). These datasets were chosen based on the facts that they contained data on over 30 samples covering a large age range, they were developed using blood samples, and they contained information on donor age, rather than there being a pre-established link between the described conditions and the age-associated markers included in this model.

Condition-related datasets were compared, at first instance, to the combined control dataset (n=2796) deriving from the 15 studies developed on the Illumina Infinium HumanMethylation 450K BeadChip technology as previously described. Datasets showing potential deviation from the combined controls were subsequently compared to control data from the same study in order to account for inter-study variability. Variability related to sex was not investigated at this instance as no evidence of sex-related bias in this marker set was observed in the previous section.

2.18 Gene annotation and ontological analysis of age prediction markers

Annotation of the CpG markers to their relevant genes was performed using the Epigenome-Wide Association Study (EWAS) Data Hub [101] based on the cg numbers (e.g. cg17885226). The gene identifiers obtained through this process (in Ensembl format, e.g. "ENSG00000126243") were subsequently used as inputs for the PANTHER [102-104] and DAVID [105-107] online software. The gene list analysis function of PANTHER was primarily used for the functional classification of the relevant genes, while similar analysis was performed using the DAVID software's functional annotation tool for comparison. Furthermore, association of the relevant genes with biological pathway networks was conducted using the KEGG (Kyoto Encyclopedia of Genes and Genomes) [108] and GAD (Genetic Association Database) [109] pathway annotation in DAVID.

This analysis was performed for both the initial selection of 244 CpGs identified for their association with age in blood and the sub-selection of 11 markers included in the final blood model.

3 Results and discussion

3.1 Marker selection

3.1.1 Age-correlated CpG sites in the literature

Review of the current literature on DNA methylation-based age prediction revealed a total of 36,137 CpG sites exhibiting methylation patterns correlated with age in 51 independent studies. While this work focuses on whole blood, information on potential markers was collated independently of the tissue of focus for the different studies, as multi-tissue applicability of certain methylation markers has been previously demonstrated. A subset of 5,364 CpG markers identified by at least two studies or included in DNA methylation-based age prediction models were shortlisted for further analysis, while information on the 18 markers appearing most frequently in the literature can be found in Table 4.

Table 4. Information on the 18 age-associated CpGs appearing most times in the literature.

No. of study mentions	CpG site	Associated Genes	Associated Gene name	No. of age prediction models CpG is present in
14 [8, 38, 43, 45, 51, 53, 59, 63-65, 68, 71-73]	cg16867657	*ELOVL2*	Fatty Acid Elongase 2	7 [8, 38, 43, 53, 63, 68, 71]
12 [8, 38, 48, 53, 63-65, 68, 69, 71-73]	cg21572722	*ELOVL2*	Fatty Acid Elongase 2	10 [6, 8, 38, 48, 53, 63, 64, 68, 69, 71]
11 [8, 38, 48, 51, 53, 63-65, 68, 71, 73]	cg24724428	*ELOVL2*	Fatty Acid Elongase 2	7 [8, 38, 48, 53, 63, 68, 71]
10 [3, 39, 51, 61, 63, 64, 66, 68, 72, 73]	cg09809672	*EDARADD*	EDAR Associated Death Domain	3 [3, 66, 68]
9 [32, 35, 39, 40, 49, 63-65, 69, 73]	cg00059225	*GLRA1*	Glycine Receptor Alpha	3 [32, 49, 69]
9 [43, 50, 63-65, 68, 69, 71, 72]	cg07553761	*SMC4, TRIM59*	Structural Maintenance of Chromosomes 4, Tripartite Motif Containing 59	4 [43, 68, 69, 71]
9 [43, 51, 63-65, 69, 71-73]	cg10501210	*C1orf132*	Chromosome 1 Open Reading Frame 132	3 [43, 69, 71]
9 [48, 51, 61, 63-65, 68, 69, 73]	cg17110586	*Unknown*	Unknown	3 [48, 68, 69]
8 [39, 40, 49, 51, 53, 63, 64, 66]	cg02228185	*ASPA*	Aspartocylase	5 [6, 49, 53, 64, 66]
8 [43, 51, 63-65, 68, 69, 73]	cg07547549	*MMP9, SLC12A5*	Matrix Metallopeptidase 9, Solute Carrier Family12 Member 5	2 [43, 69]
8 [35, 39, 48, 49, 51, 63, 68, 73]	cg08090640	*IFI35*	Interferon-induced 35kDA protein	2 [48, 49]
8 [31, 35, 39, 49, 51, 60, 61, 73]	cg16363586	*BST2*	Bone Marrow Stromal Cell Antigen 2	1 [49]
8 [3, 39, 40, 59, 63-65, 69]	cg22736354	*NHLRC1*	E3 Ubiquitin-protein Ligase	2 [3, 69]
7 [43, 51, 63-65, 68, 69]	cg04875128	*OTUD7A*	OTU Deubiquitinase 7A	3 [43, 68, 69]
7 [38, 43, 51, 63-65, 73]	cg06639320	*FHL2*	Four and a Half LIM Domains 2	4 [6, 38, 43, 64]
7 [43, 63-65, 68, 69, 71]	cg08097417	*KLF14*	Krüppel-like factor 14	3 [43, 69, 71]
7 [38, 43, 51, 63-65, 73]	cg22454769	*FHL2*	Four and a Half LIM Domains 2	2 [38, 43]
7 [38, 43, 51, 63-65, 68]	cg24079702	*FHL2*	Four and a Half LIM Domains 2	2 [38, 43]

3.1.2 Microarray datasets

In total, methylation data from 1229 samples from individuals aged between 2-88 years were collated from studies employing the 27K platform, while 2796 samples from individuals aged between 8 months and 101 years were collated for the 450K platform. In the 27K data a minimum of 75 samples were collected per age decade up to the age of 80 years, whilst a minimum of 120 samples per age decade up to the age of 90 in the 450K data. For both datasets the oldest age group (80-90 years in the 27K and 90-100 years in the 450K) contained a limited number of samples (n<20). Finally, a balanced male to female ratio was observed for most age groups in the 450K data, as opposed to the 27K data where the majority of the samples in the younger age groups belong to male donors and a large number of samples containing no information on sex appear in the older age groups.

3.1.3 Marker evaluation

In the first step of marker selection, using microarray data from the 27K and 450K Infinium platforms independently, a subset of 244 markers were identified for their high correlation with chronological age and large range of methylation values over the human lifespan. Evaluation of markers based on the observed methylation range over the human lifespan was included in this analysis in an effort to increase sensitivity, as larger methylation differences between the age groups can potentially eliminate the effect of technical noise during the quantification of DNA methylation from low quantities of template [53].

Out of the 244 shortlisted markers, 88 have been already incorporated in published DNA methylation-based age prediction models. Whilst data from the two microarray platforms were analysed independently, 188 markers were unique for the 450K platform and 56 were present in both platforms but no markers unique for the 27K fulfilled the strict thresholds applied for this analysis. This result can be traced back to the fact that the number of unique probes for the 27K is limited, as well as the fact that the dataset collated from this microarray is smaller and more unbalanced than the 450K one. Nonetheless, for the 56 common markers the observed methylation trends were consistent in the two datasets.

Additionally, in the 244 CpG marker set, CpGs associated with the same promoter/gene, such as *ELOVL2* (3 CpGs), *FHL2* (3 CpGs) and *ASPA* (2 CpGs), showed consistent methylation trends (hyper- or hypomethylation with age). Furthermore, 210 markers (86%) exhibited hypomethylation trends with age, an observation that contradicts previous findings suggesting an enrichment of hypermethylation trends for age-associated CpGs [6]. The most likely origin of these opposing observations relates to the fact that the majority of the markers identified in this study are unique for the 450K platform, whilst the work by Koch *et al.* focuses exclusively on datasets developed with the 27K platform [6]. Looking at the main differences between the two microarray platforms, it is evident that the extended probe set of the 450K platform targets significantly more CpGs located outside CpG islands (CGIs) than the 27K probes, that mainly target CGIs. Annotation of the 244 selected markers, showed that only 14% were located in CGIs, 94% of which showed hypermethylation with age, whilst the remaining 86% were located outside CGIs with 99% of them revealing age-related hypomethylation. These observations are concordant with previous reports suggesting that age-associated hypermethylation is enriched in CGIs and hypomethylation is predominant in CpGs outside CGIs [45, 51] and provide with an explanation for the discordance with the observations by Koch *et al*. [6].

Finally, since this analysis focuses on blood, it is worth noting that whilst it has been suggested that hypomethylation trends with age in whole blood can represent changes in the cell composition of this tissue [41], studies have repeatedly proven that such effects, when present, are minor and do not affect the observed age-correlated methylation patterns [3, 36, 43, 48]. In this study, the use of multiple datasets, with some deriving from specific blood cell types rather than whole blood, combined with the investigation of markers that have been previously identified for their correlation with age by multiple independent studies, practically eliminates the chance of selecting markers with false association with age.

3.1.4 Final marker set

Further analysis of the data obtained for the 244 CpG marker set revealed a set of 19 markers with superior combination of correlation with age and methylation range over the human lifespan (Table 5). Comparison of this marker set with the set of 18 most popular markers in the literature (Table 4) reveals that the two sets are over 50% identical, sharing 10 markers, a finding that may be unsurprising. Notably, even though 86% of the markers in the 244 CpG selection were hypomethylated with age, in the final selection the markers are split almost 50-50 between those exhibiting hypomethylation (10 CpGs) and hypermethylation (9 CpGs) trends. However, the 19 markers correspond to 15 different genes, with *ELOVL2* and *FHL2* genes represented by 3 CpGs each that all exhibit hypermethylation trends with age. Taking this into account, when looking at the markers at the gene level, the ratio of hypomethylated to hypermethylated changes to 2:1, which is still higher than expected based on the low representation of markers exhibiting hypermethylation with age in the original selection.

Out of the 19 markers 14 have been previously incorporated in DNA methylation-based age prediction models, while comparison of the correlation coefficients obtained for the selected markers in this study and that observed for the same markers in previous publications revealed high concordance of the results.

Table 5. Information on the 19 markers selected for further analysis. Pearson's correlation (r) for this study is based on data from the 450K array. This table also includes Pearson's correlation (r) observed in previous studies, as well as the absolute range of beta values observed for the relevant markers over the different ages, and the number of times these markers have been used in age estimation models in the literature. *Highlighted markers were included in the final model proposed by this study after validation (see section 3.2.3)

CpG site	Associated Gene name	Pearson's correlation (r) in 450K	Pearson's correlation (r) in other studies	Beta value range	No. of age prediction models CpG is present in
cg16867657	*ELOVL2*	0.91	0.83	0.7507	7 [8, 38, 43, 53, 63, 68, 71]
cg22454769	*FHL2*	0.87	0.74	0.8346	2 [38, 43]
cg10501210*	***MIR29B2CHG (C1orf132)***	**-0.84**	**–0.74**	**0.8957**	**3** [43, 69, 71]
cg19283806*	***CCDC102B***	**-0.83**	**–0.72, -0.89, -0.64**	**0.8744**	**4** [6, 43, 63, 64]
cg06639320*	***FHL2***	**0.81**	**0.75, 0.90, 0.74**	**0.7450**	**4** [6, 38, 43, 64]
cg24079702	*FHL2*	0.80	0.74, 0.66	0.7767	2 [38, 43]
cg00329615	*IGSF11*	-0.80	-0.58	0.6844	0
cg24724428*	***ELOVL2, ELOVL2-AS1***	**0.80**	**0.67**	**0.6403**	**7** [8, 38, 48, 53, 63, 68, 71]
cg21572722*	***ELOVL2***	**0.80**	**0.79, 0.94**	**0.6226**	**10** [6, 8, 38, 48, 53, 63, 64, 68, 69, 71]
cg09809672*	***EDARADD***	**-0.79**	**-0.94, -0.61**	**0.7942**	**3** [3, 66, 68]
cg07553761*	***SMC4, TRIM59***	**0.78**	**0.72, 0.65**	**0.9193**	**4** [43, 68, 69, 71]
cg22796704*	***ARHGAP22***	**-0.77**	**-0.64**	**0.6016**	**1** [43]
cg08128734*	***RASSF5***	**-0.76**	**-0.59**	**0.6873**	**0**
cg17372101*	***CNTNAP2***	**-0.76**	**-0.54**	**0.6772**	**0**
cg18618815	*COL1A1*	-0.76	-0.58	0.6928	0
cg08160331	*KLHL35*	0.76	0.65	0.6999	1 [48]
cg08262002*	***LDB2***	**-0.76**	**-0.55**	**0.6576**	**1** [48]
cg12934382	*GRM2*	0.76	0.56	0.7990	0
cg17471102	*FUT3*	-0.75	-0.59	0.6109	2 [49, 66]

3.2 *Validation of the MPS-based assay*

3.2.1 Linearity

Pre-mixed standards at 0%, 5%, 10%, 25%, 50%, 75% and 100% methylation were used in order to assess the ability of this 18-marker method (14 amplicons) to accurately quantify different levels of methylation at the selected CpG sites. All standards were processed in duplicate and sequenced simultaneously. Comparison between the expected and observed methylation fraction showed high coefficient of determination between the two for 8 out of 14 markers (markers present on the same amplicon, such as cg16867657, cg24724428, cg21572722 for *ELOVL2* and cg06639320, cg22454769, cg24079702 for *FHL2*, were analysed together) with R^2>0.87. Noticeable bias towards overestimation of methylation was observed for markers associated with the *FHL2* gene (cg06639320, cg22454769, cg24079702, R^2=0.72), cg08128734 (*RASSF5*) (R^2=0.69), cg10501210 (*MIR29B2CHG*) (R^2=0.63), cg18618815 (*COL1A1*) (R^2=0.44) and cg22796704 (*ARHGAP22*) (R^2=0.19), while marker cg08160331 (*KLHL35*) failed to provide with any distinction between methylation levels and was thus excluded from further analysis (Supplementary_Fig_S1). Furthermore, a second primer set, previously described by Naue *et al.* [88], was investigated for the *ELOVL2* markers but demonstrated higher bias (R^2=0.75) compared to the design proposed here (R^2=0.96). The bias towards the methylated allele observed for some of the markers did not result in a significant skewing of the overall linearity when results for 17 markers (excluding cg08160331 (*KLHL35*)) were combined (R^2=0.96) (Figure 1). Furthermore, whilst high bias practically results in the observed methylation being 0 or 100%, eliminating the chance of distinction between the different methylation levels, a low level of bias can be accounted for in the subsequent analysis as long as it is consistent.

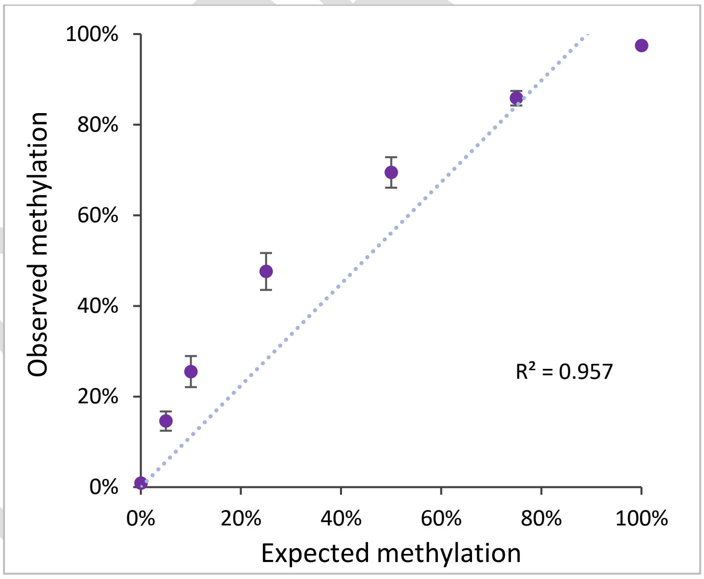


Figure 1. Comparison between the expected and average observed methylation fraction (β-values expressed as percentage of methylation) for the 17 selected markers. The 'observed' methylation values represent the average observed methylation for all 17 CpGs for each of the standards (at 0%, 5%, 10%, 25%, 50%, 75% and 100% methylation). Error bars represent the standard deviation for the different CpG sites and the R^2 value for the linear correlation is displayed on the chart.

3.2.2 Reproducibility

The reproducibility of the developed assay for the quantification of DNA methylation at the 17 CpGs was assessed by comparing the methylation values obtained for these sites in 20 blood samples analysed in duplicate post DNA extraction and quantification. The average absolute difference observed between the duplicates for all markers was calculated at 4%, with approximately 71% of the markers (12 out of 17) exhibiting an average difference below that point (Supplementary_Fig_S2). The largest differences were observed at cg18618815 (*COL1A1*) and the 3 CpGs related to the *ELOVL2* gene (cg16867657, cg24724428, cg21572722). This increased variation can potentially be traced back to the low amplification efficiency of the 2 corresponding amplicons (cg16867657, cg24724428, cg21572722 are part of the same amplicon), that resulted in limited reads for one or both duplicates. The sequencing coverage obtained for those two amplicons (targeting *ELOVL2* and *COL1A1*) averaged at 702 and 562 reads per sample respectively and was consistently lower than the remaining 11 amplicons that obtained an average of 3,841-34,178 reads per sample (Supplementary_Fig_S3). Nonetheless, despite the increased variation observed between duplicates for certain markers in this assay, the reproducibility results were considered satisfactory for this method given the fact that the overall methylation range over the human lifespan for these markers is at least 11 times higher than the relevant variation between replicates.

3.2.3 Age prediction

Following a final marker elimination, based on statistical predictor variable selection using the 112-sample dataset analysed in-house with the previously outlined method, a set of 11 markers (cg24724428, cg21572722, cg06639320, cg09809672, cg22796704, cg08128734, cg17372101, cg10501210, cg19283806, cg07553761 and cg08262002) relating to 10 different genes (*ELOVL2*, *ELOVL2*, *FHL2*, *EDARADD*, *ARHGAP22*, *RASSF5*, *CNTNAP2*, *MIR29B2CHG*, *CCDC102B*, *SMC4*/*TRIM59* and *LDB2* respectively) were selected. Using the same split of the dataset as previously described by Aliferi *et al.* [28], a support vector machine model with polynomial function was trained on 77 samples and was further tested using the remaining 35 samples (2 additional samples added in this set compared to previous work [28]). The mean absolute prediction error was calculated at 3.6 years (RMSE=5.1 years) for the training set and at 3.3 years (RMSE=4.4 years) for the test set, with the similarity between these values for the two sets suggesting high model generalizability and no presence of overfitting (Figure 2). Furthermore, over 71% of the samples present in the test set predicted with an absolute error of less than 4 years, while 89% predicted with an absolute error of less than 7 years. Compared to the previously published model [28], this model does not only achieve increased accuracy, with the mean absolute error reduced by 0.4 and 0.7 years in the training and test sets respectively, but also demonstrates a ~1.4 times higher percentage of samples predicting with an error range of ±4 years, as the relevant score for the previous model was 52%.

Additionally, a separate SVMp model trained on all 18 markers and on the same dataset showed identical RMSE values with the 11-marker model (5.1 years for the training and 4.4 years for the test set), providing further evidence in support of the proposed marker elimination.

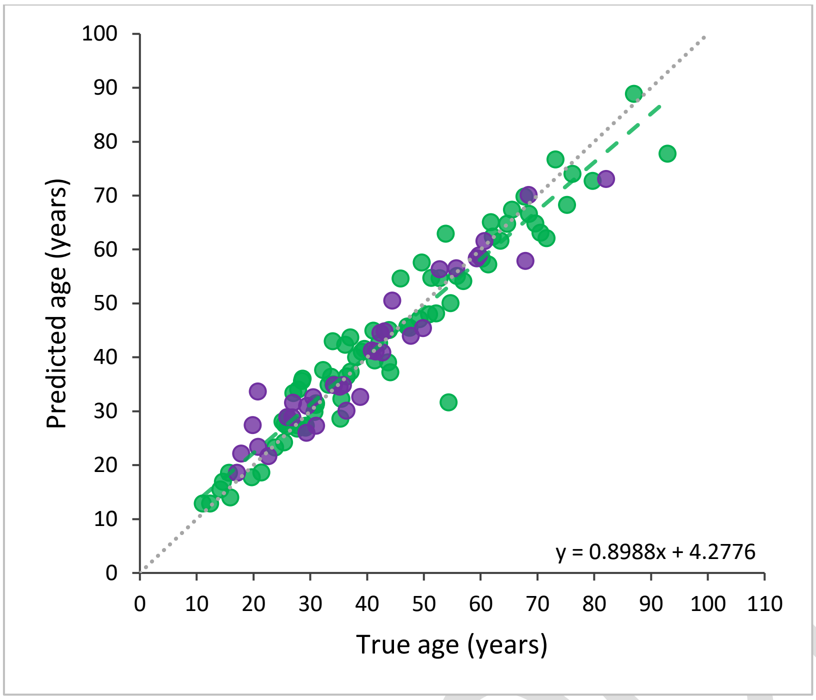


Figure 2. Comparison between the predicted and the given age for the training (green, n=77) and blind test set (purple, n=35) in the SVMp model. The mean absolute prediction error was calculated at 3.6 and 3.3 years respectively. The equation of the linear trendline fitting the training set (green dashed line) can be seen on the graph, while the grey dotted line represents the 'perfect' predictions where predicted and true age overlap (y=x).

Furthermore, a separate set of 88 DNA extracts from whole blood samples, obtained as part of a collaboration with the University of Santiago de Compostela in Spain (USC) [110], were also processed in-house following the previously outlined 11-marker method. Given that the number of samples in this set was larger than the original training set of the prediction model, the SVMp algorithm was re-trained using the entire KCL dataset (n=112) and the USC dataset was introduced as a blind test. The MAE for the USC set was calculated at 3.8 years (RMSE=5 years), closely matching the expected prediction accuracy based on the results obtained by the original training and test set. Further analysis of the predictions for this dataset revealed a loss of prediction accuracy for individuals aged over 60 years (Figure 3), possibly relating to the low number of samples for the age groups of 60-70 years (n=13), 70-80 years (n=6), 80-90 years (n=2) and 90-100 years (n=1) included in the training set. At the same time, a loss of accuracy in the prediction of age for older individuals has been reported by multiple studies [3, 8, 53, 64, 71, 88, 111, 112] and has been associated with an increased effect of non-genetic factors in the methylation patterns of older individuals [3], as well as a lower variation in age-related methylation for older ages, that makes it hard to distinguish between them [112]. Nonetheless, according to the national DNA database for the UK, as of June 2019, 95% of the profiles belong to individuals under the age of 55 years at the time of inclusion [113]. Given the forensic scope of this work, age-estimation statistics were calculated for the 'forensically-relevant' age group (<55 years) from the USC dataset. The results reveal high accuracy with a MAE of 2.6 years (RMSE=3.1 years) and 81% of the samples predicting with an absolute error of less than 4 years.

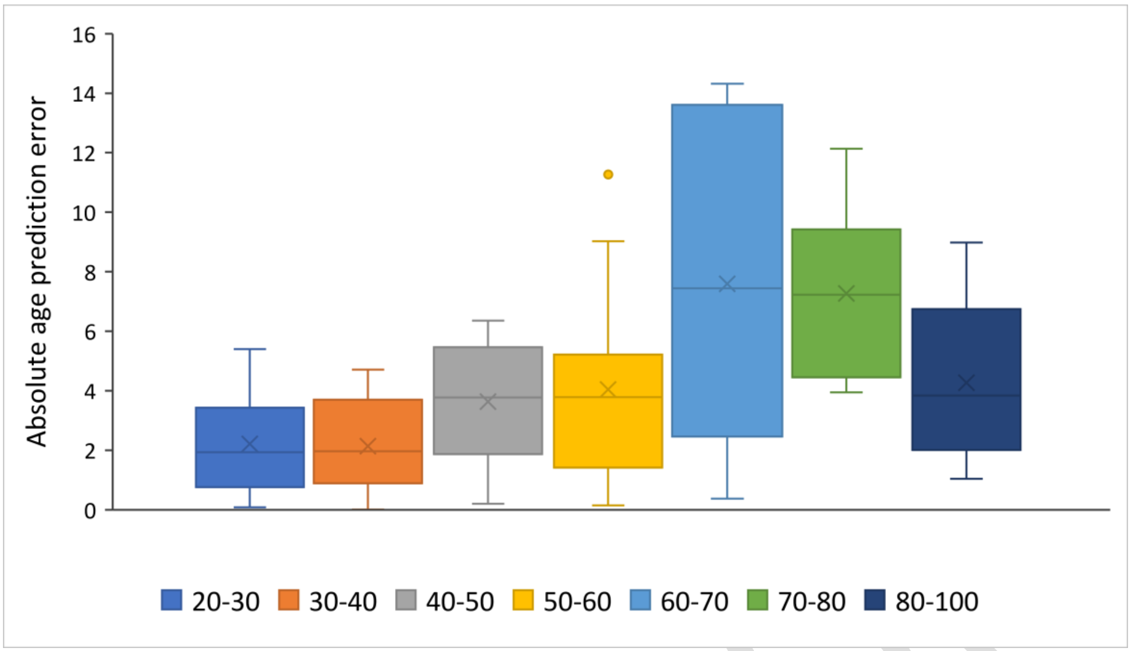


Figure 3. Box plots representing the spread of absolute prediction error for samples in 7 distinct age groups separated by decade between the ages of 20 and 100 years. The vertical line inside each box represent the median absolute error for the relevant age group, while the x mark represents the average absolute error for the same group.

3.2.4 Sensitivity

In order to assess the sensitivity of the final 11-marker method (10 amplicons), six whole blood samples from the test dataset, belonging to individuals aged 17, 27, 36, 43, 53 and 61 years, were re-analysed starting with 6 different DNA inputs for bisulphite conversion. The DNA inputs used were 50 ng, as previously used for the initial analysis, 25 ng, 10 ng, 5 ng, 2.5 ng and 1 ng. Taking into account the loss of template in the bisulphite conversion state (~52% recovery [85]), the elution volume and the two multiplex reactions required for the amplification of all markers this translates to approximately 10, 5, 2, 1, 0.5 and 0.2 ng in the PCR stage.

In terms of precision in the quantification of DNA methylation itself, the values obtained for most markers did not vary between the 50, 25, 10 and 5 ng inputs, but increased variation was observed for the 2.5 and 1 ng inputs in all markers (Supplementary_Fig_S4). This is also reflected in the average difference in methylation observed for the entire marker set at the different DNA inputs and correlates with a loss of sequencing reads at these levels (Figure 4). At this point it is worth noting that all 1 ng replicates obtained less than 100 reads in at least 3 markers, while for cg21572722 (*ELOVL2*), cg24724428 (*ELOVL2*) and cg19283806 (*CCDC102B*) virtually no reads (under 10) were obtained at this input with 6 methylation values requiring imputation *in silico*.

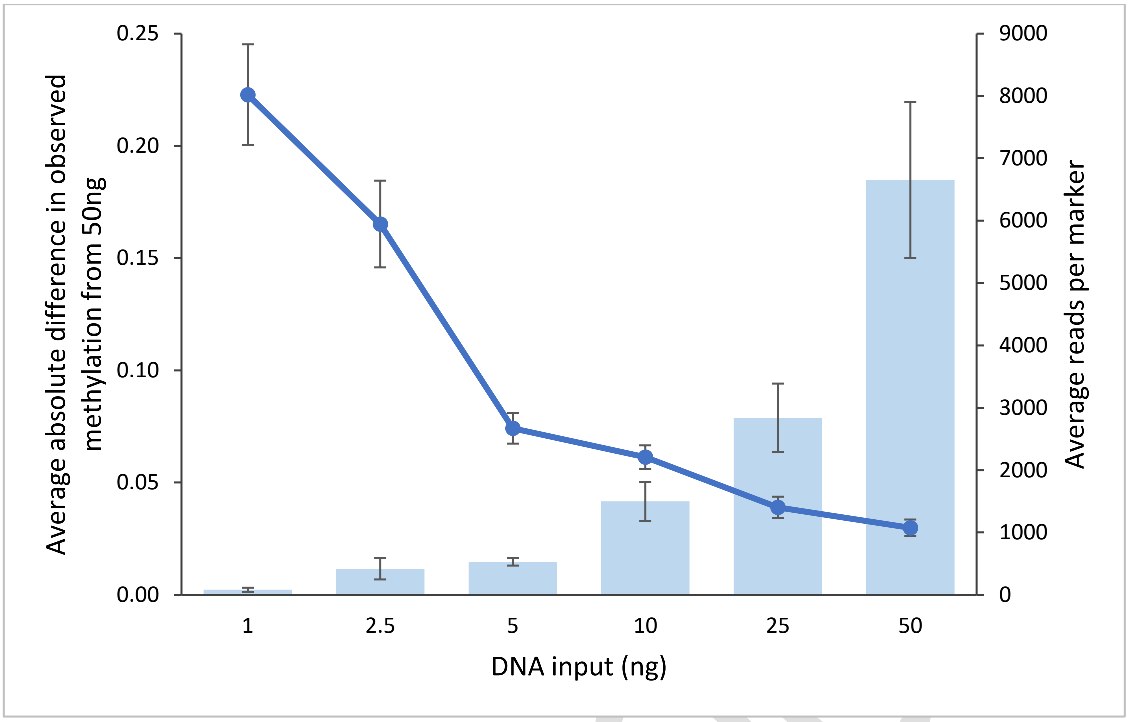


Figure 4. Average absolute difference between the methylation β-values observed when using inputs of 50, 25, 10, 5, 2.5 and 1 ng and those observed during the original quantification of methylation (50 ng) for 6 whole blood samples at all 11 markers (blue line – note that the line is included to aid with visual representation and no measurements were taking between the 6 points). The bars represent the average number of sequencing reads obtained for each marker for the different inputs. Error bars represent the standard deviation observed at each point.

In terms of accuracy in age prediction, this was successfully retained down to 5 ng of DNA input, whilst the error increased drastically at the 2.5 and 1 ng inputs following the trend seen in the precision analysis (Figure 5). An important observation at this point relates to the fact that, whilst both precision in the quantification of DNA methylation and prediction accuracy are highly retained down to 5ng of DNA input, the slight slope in the precision graph between 5 and 50 ng, relating to a slight increase in variation as the input is reduced, is not reflected in the predictions, with both the MAE and RMSE values remaining practically identical for the 25 (MAE=4.3 years, RMSE=5.6 years), 10 (MAE=4.3 years, RMSE=6.2 years) and 5ng (MAE=3.9 years, RMSE=5 years) inputs. These results suggest that the prediction algorithm is able to successfully cope with loss of accuracy in the quantification of DNA methylation, with issues only appearing when a significant loss of sequencing power, resulting in complete loss of reads for some markers, is observed. Furthermore, at these levels of DNA input, stochastic effects that can skew the observed methylation values are expected due to the low number of template molecules.

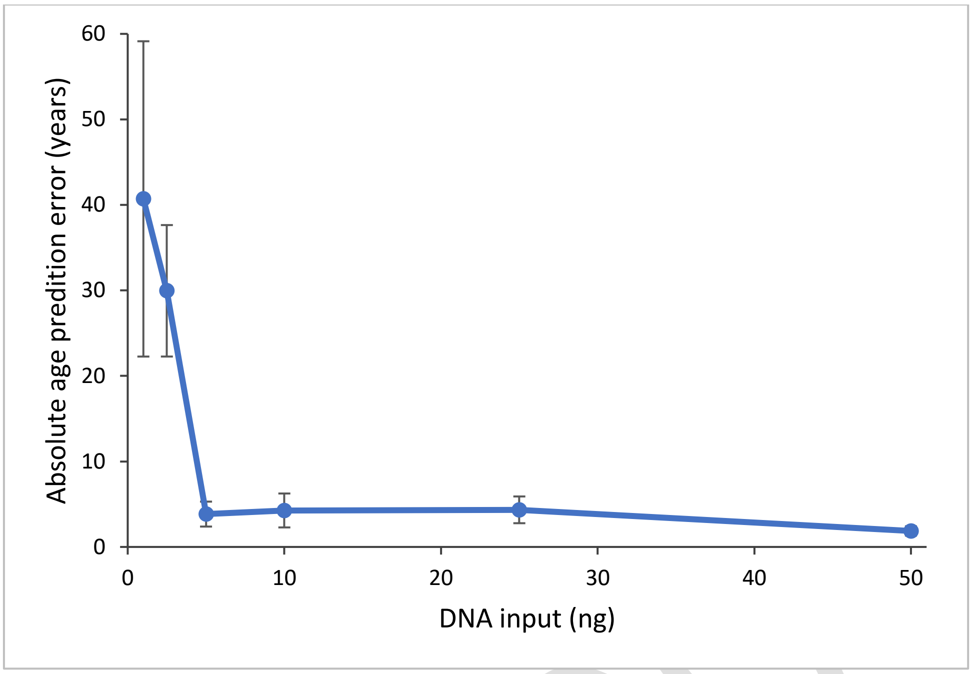


Figure 5. Average absolute error in age prediction observed for a set of samples (n=6) analysed at different DNA inputs corresponding to 50, 25, 10, 5, 2.5 and 1 ng. Error bars represent the standard deviation of the prediction error between the 6 samples.

3.2.5 Sex association

In order to investigate potential sex-specific bias for this method, prediction accuracy was assessed independently for the two sexes in the training and test sets of the SVMp age prediction model based on the same markers. The observed mean absolute prediction error (MAE) was similar for the two sexes in the training set (3.7 years for males and 3.6 years for females), however, the difference increased in the test set with females predicting with increased accuracy (MAE=3.7 years in males and 2.9 years in females). Despite this, the limited size of these datasets (n=34 in the test set) makes it hard to draw any conclusions from these results. Furthermore, the fact that 2 out of 3 individuals over the age of 65 years in the test set are males, introduces a potential bias as a decrease in prediction accuracy has previously been observed for samples deriving from older individuals. At the same time, a slight decrease in the accuracy of age estimation in males has been previously reported in the literature for DNA methylation-based age prediction, albeit not representing a statistically significant variation [88].

Furthermore, in order to investigate this in a larger scale, the correlation between age and methylation was examined separately for males and females for the 11 age-associated markers in the combined 450K microarray dataset (n=2,744). The correlation coefficient (r) values obtained, indicated strong (|r|>0.6) to very strong (|r|>0.8) correlation between age and methylation status for all markers independently of sex, in concordance with the results previously obtained for the combined dataset. However, with the exception of marker cg24724428 (*ELOVL2*), absolute correlation values obtained for the female cohort were slightly lower than those of the male cohort (Figure 6), an observation that further suggest that the slight decrease in the accuracy of age estimation in males observed in the targeted sequencing data is not of statistical significance.

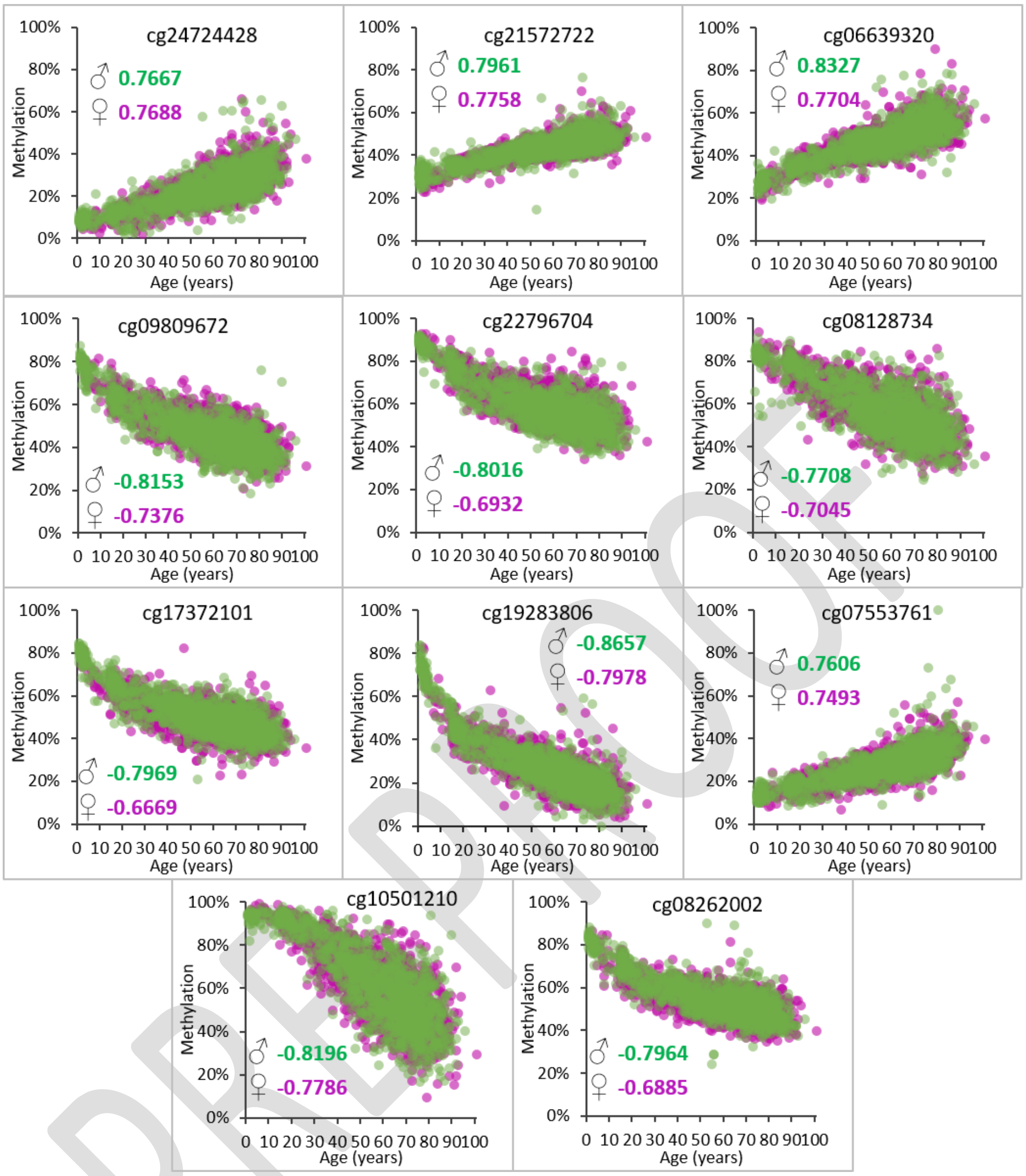


Figure 6. Comparison between the methylation trends (β-values expressed as a methylation percentage, not normalised) of male (green, n=1311) and female (purple, n=1433) blood samples in the 450K microarray for the 11 markers selected for age prediction in this tissue. The Pearson correlation values (r) for each sex are included in the relevant graphs.

3.3 *Disease association*

3.3.1 Publicly available datasets

Using data from publicly available datasets, methylation trends with age were compared for the 11 age-associated markers between control samples and samples obtained from individuals suffering from conditions such as schizophrenia (n=62), rheumatoid arthritis (n=354), frontal temporal dementia (n=121) and progressive supranuclear palsy (n=42). Both control and diseased samples exhibited similar methylation trends and β-value range with age for each of the 11 markers (Figure 7), indicating an absence of additional variation in relation to these conditions for this marker set. Methylation values obtained for schizophrenia samples in cg24724428 (*ELOVL2*), borderline falling out of range for this marker, were further compared to those obtained for control samples in the same study. This comparison revealed a clear overlap between the two sets, indicating the presence of a study-specific rather than condition-specific effect.

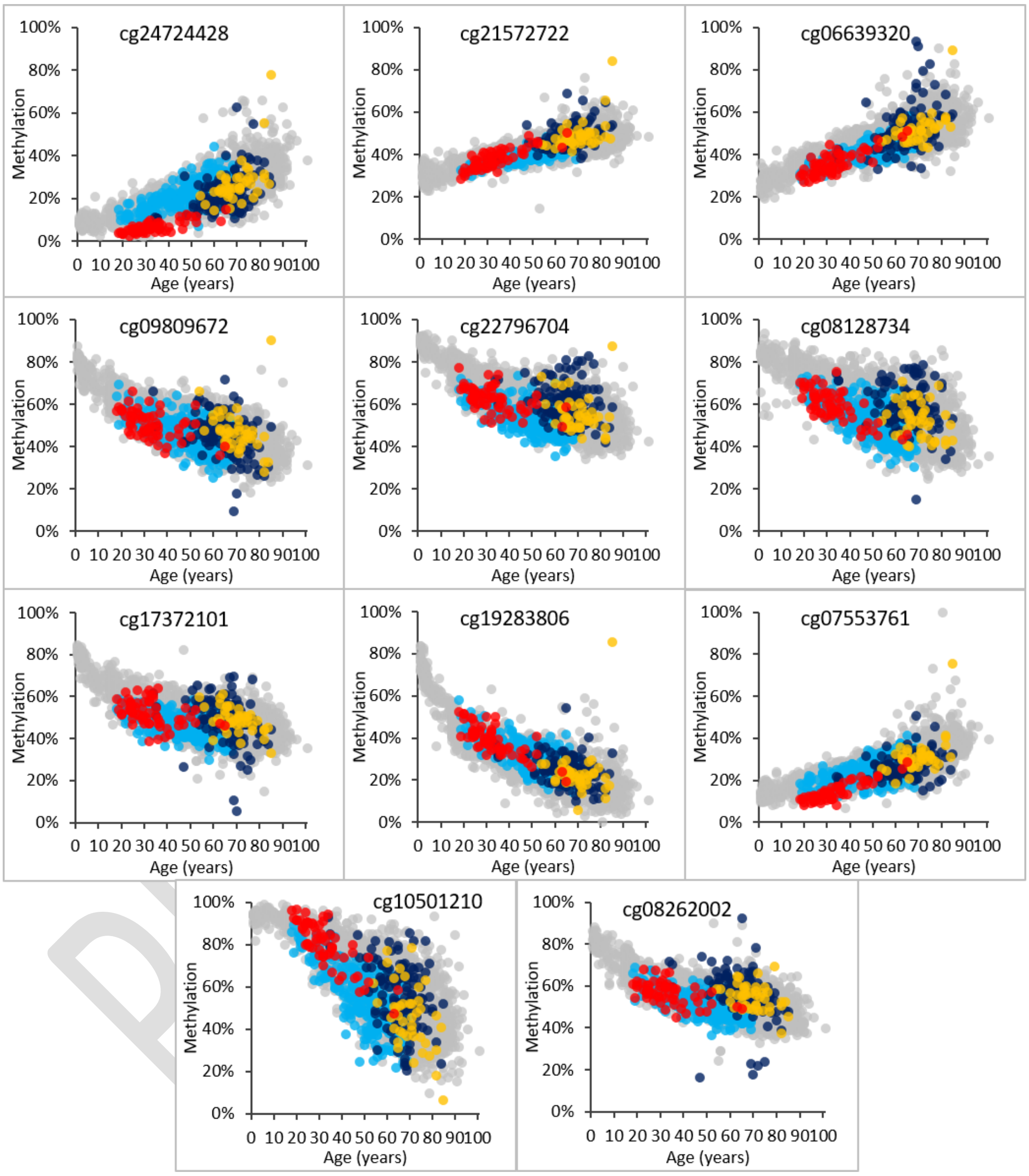


Figure 7. Comparison between the methylation trends (β-values expressed as a methylation percentage) with age for control populations (grey) and cohorts of individuals diagnosed with schizophrenia (red), rheumatoid arthritis (light blue), frontal temporal dementia (dark blue) and progressive supranuclear palsy (yellow), for the 11 CpGs included in the DNA methylation-based age prediction model. For the control populations the data represent a compilation from control samples from 15 different datasets (n=2796), whilst data for each disease group originate from a single study. All data derive from Infinium 450K arrays.

Finally, the correlation between age and methylation values obtained for controls and condition-related samples was compared for the different markers (Supplementary_Table_S7). Whilst, when compared to the correlation observed for the combined controls dataset, weaker correlations were observed for all markers for the frontal temporal dementia and progressive supranuclear palsy samples, comparison with the same-study controls revealed similar r scores. These results further indicate that variation observed for these datasets derives from batch effects related to the relevant studies and no condition-related variation is observed for these 4 conditions in this marker set.

3.3.2 Biological pathways

In addition to the disease association analysis conducted using microarray data, the involvement of the genes related to the age markers in biological pathways was also investigated. This analysis revealed association with a variety of diseases and conditions for the 164 genes relating to the 244 markers previously identified for their correlation with age in the tissue of blood. Over 40 different genes associated with this marker set were involved in biological pathways relating to metabolic (66 genes), cardiovascular (54 genes), chemical dependency (45 genes) and neurological (44 genes) conditions (Supplementary_Fig_S5).

Furthermore, comparison of this gene list with the KEGG pathways, a collection of pathway maps that represent current knowledge of molecular interactions, reactions and relation networks, revealed association with T-cell leukaemia retrovirus infection (HTLV-I), non-alcoholic fatty liver disease (NAFLD), inflammatory bowel disease (IBD) as well as asthma, graft-versus-host disease/allograft rejection and type I diabetes (Supplementary_Fig_S6). Looking further into some of these conditions, such as graft-versus-host disease, it comes as no surprise that its prevalence has been previously associated with age in medical studies [114].

These results highlight a large number of conditions that could affect the methylation levels/trends at these age-correlated CpGs, potentially skewing the prediction accuracy of DNA methylation-based age estimation. This can be related to the use of markers for which the correlation with chronological age is not direct but rather stemming from their association with biological age. However, when this annotation was limited to the 11 markers (10 genes) included in the final age estimation model, association was only indicated for obesity (BMI) (4 genes) and tobacco use (6 genes). Both of these associations have been previously highlighted in the literature for age-related CpG sites [20-22, 26, 115, 116], suggesting that analysis of relevant sample cohorts might be beneficial in further addressing potential issues with this marker set.

3.4 Gene ontology

Annotation of the 244 markers previously identified for their correlation with age in the tissue of blood revealed association with 164 different genes involved mainly in cellular processes (58 genes, 35%), biological regulation (37 genes, 23%) and metabolic processes (37 genes, 23%) (Supplementary_Fig_S7). In terms of molecular function, defined as the function that a protein performs on its direct molecular targets, the main activity categories identified for the proteins associated with this marker group related to binding (33 genes, 20%) and catalytic (31 genes, 19%) activities (Supplementary_Fig_S8).

Looking further into these associations, the two strongest links established for this this set of DNA methylation age markers relate to metabolism and cellular communication, processes that, unsurprisingly, have been previously associated with the 9 ‘hallmarks of aging’ as defined by López-Otín *et al.* [117]. Each of these hallmarks has been associated with undesirable metabolic alterations, with the strongest links observed with ‘deregulated nutrient sensing’ and ‘mitochondrial dysfunction’ [118], while ‘altered intercellular communication’ is a hallmark of its own [117]. Furthermore, the association between various metabolic parameters and longevity has been the focus of multiple studies, both in terms of investigating its underlying mechanisms [118, 119] and assessing the use of this connection to promote healthy aging [120].

4 Conclusions

This work describes an attempt to integrate current research outputs on DNA methylation-based age prediction into an accurate and sensitive tool with high potential for application in forensic casework.

Introducing a new approach to marker selection, aimed towards minimizing the required DNA input for DNA methylation-based age estimation, and combining analyses of both microarray and targeted-sequencing data, a set of 11 CpG sites were identified as the markers with the highest potential for forensically orientated age estimation. Drawing upon previous knowledge on targeted sequencing-based methylation analysis coupled with the use of machine learning for age estimation [28], the developed 11-marker support vector machine model trained on data from the MiSeq platform was able to predict the age of two independent test sets from the UK (n=35) and Spanish (n=88) populations with a MAEs of 3.3 and 3.8 years respectively. Additionally, investigating a more forensically relevant age range (<55 years), an even lower error of 2.6 years was observed, with 81% of the samples predicted with an absolute error of less than 4 years.

Whilst similar levels of age estimation accuracy (MAE 2.9-3.7 years) have been previously recorded by similar studies, the accuracy of this model was successfully retained down to 5 ng of starting DNA material which is 4-1,400 times lower than any other published work to date and approximately half of the limit observed in our previous work conducted on a set of pre-selected markers [28]. Furthermore, in addition to the model's accuracy being retained despite environmental and lifestyle differences between individuals from Spain and the UK, there was also no indication of bias related to sex, in concordance to the relevant literature [32, 53, 54, 64, 111, 121], or conditions such as schizophrenia, rheumatoid arthritis, frontal temporal dementia and progressive supranuclear palsy.

Analysis of the markers at gene level revealed potential association with metabolic and cardiovascular diseases, with the main links highlighted for the 11 markers included in the final model relating to obesity and smoking. Whilst these associations do not necessarily translate to age-estimation bias for individuals with the relevant conditions, they raise questions worth exploring as age estimation panels move towards implementation in forensic casework. Finally, ontological analysis of the relevant genes also revealed strong association with various metabolic processes taking place at a cellular level, highlighting the close relationship between the age-informative markers and the hallmarks of human ageing [117] and raising questions regarding the overlap between methylation markers for chronological and biological age and its potential effect on the prediction accuracy of forensic DNA methylation-based age estimation.

5 References

1. Yi, S.H., et al., *Isolation and identification of age-related DNA methylation markers for forensic age-prediction.* Forensic Sci Int Genet, 2014. **11**: p. 117-25.
2. Zhuang, J., M. Widschwendter, and A.E. Teschendorff, *A comparison of feature selection and classification methods in DNA methylation studies using the Illumina Infinium platform.* BMC Bioinformatics, 2012. **13**: p. 59.
3. Horvath, S., *DNA methylation age of human tissues and cell types.* Genome Biology, 2013. **14**(10): p. 115.
4. Laird, P.W., *Principles and challenges of genomewide DNA methylation analysis.* Nat Rev Genet, 2010. **11**(3): p. 191-203.
5. Dedeurwaerder, S., et al., *A comprehensive overview of Infinium HumanMethylation450 data processing.* Briefings in Bioinformatics, 2014. **15**(6): p. 929-941.
6. Koch, C.M. and W. Wagner, *Epigenetic-aging-signature to determine age in different tissues.* Aging, 2011. **3**(10): p. 1018-1027.
7. Masser, D.R., A.S. Berg, and W.M. Freeman, *Focused, high accuracy 5-methylcytosine quantitation with base resolution by benchtop next-generation sequencing.* Epigenetics Chromatin, 2013. **6**(1): p. 33.
8. Zbiec-Piekarska, R., et al., *Examination of DNA methylation status of the ELOVL2 marker may be useful for human age prediction in forensic science.* Forensic Sci Int Genet, 2015. **14**: p. 161-7.
9. Das, P.M. and R. Singal, *DNA methylation and cancer.* J Clin Oncol, 2004. **22**(22): p. 4632-42.
10. Levine, M.E., et al., *Epigenetic age of the pre-frontal cortex is associated with neuritic plaques, amyloid load, and Alzheimer's disease related cognitive functioning.* Aging (Albany NY), 2015. **7**(12): p. 1198-211.
11. Lunnon, K. and J. Mill, *Epigenetic studies in Alzheimer's disease: current findings, caveats, and considerations for future studies.* Am J Med Genet B Neuropsychiatr Genet, 2013. **162B**(8): p. 789-99.
12. Smith, A.R., et al., *A cross-brain regions study of ANK1 DNA methylation in different neurodegenerative diseases.* Neurobiol Aging, 2019. **74**: p. 70-76.
13. Horvath, S., et al., *Huntington's disease accelerates epigenetic aging of human brain and disrupts DNA methylation levels.* Aging (Albany NY), 2016. **8**(7): p. 1485-512.
14. Horvath, S. and B.R. Ritz, *Increased epigenetic age and granulocyte counts in the blood of Parkinson's disease patients.* Aging (Albany NY), 2015. **7**(12): p. 1130-42.
15. Horvath, S., et al., *Epigenetic clock for skin and blood cells applied to Hutchinson Gilford Progeria Syndrome and ex vivo studies.* Aging (Albany NY), 2018. **10**(7): p. 1758-1775.
16. Maierhofer, A., et al., *Accelerated epigenetic aging in Werner syndrome.* Aging (Albany NY), 2017. **9**(4): p. 1143-1152.
17. Breitling, L.P., et al., *Tobacco-smoking-related differential DNA methylation: 27K discovery and replication.* Am J Hum Genet, 2011. **88**(4): p. 450-7.
18. Jenkins, T.G., et al., *Cigarette smoking significantly alters sperm DNA methylation patterns.* Andrology, 2017. **5**(6): p. 1089-1099.
19. Lee, K.W. and Z. Pausova, *Cigarette smoking and DNA methylation.* Front Genet, 2013. **4**: p. 132.
20. Mansego, M.L., et al., *Differential DNA Methylation in Relation to Age and Health Risks of Obesity.* Int J Mol Sci, 2015. **16**(8): p. 16816-32.
21. Almen, M.S., et al., *Genome-wide analysis reveals DNA methylation markers that vary with both age and obesity.* Gene, 2014. **548**(1): p. 61-7.
22. Levine, M.E., et al., *An epigenetic biomarker of aging for lifespan and healthspan.* Aging (Albany NY), 2018. **10**(4): p. 573-591.

23. Hughes, A., et al., *Socioeconomic Position and DNA Methylation Age Acceleration Across the Life Course.* Am J Epidemiol, 2018. **187**(11): p. 2346-2354.

24. McDade, T.W., et al., *Genome-wide analysis of DNA methylation in relation to socioeconomic status during development and early adulthood.* Am J Phys Anthropol, 2019. **169**(1): p. 3-11.

25. Sprott, R.L., *Biomarkers of aging.* Exp Gerontol, 1988. **23**(1): p. 1-3.

26. Lu, A.T., et al., *DNA methylation GrimAge strongly predicts lifespan and healthspan.* Aging (Albany NY), 2019. **11**(2): p. 303-327.

27. Vidaki, A., et al., *DNA methylation-based forensic age prediction using artificial neural networks and next generation sequencing.* Forensic Sci Int Genet, 2017. **28**: p. 225-236.

28. Aliferi, A., et al., *DNA methylation-based age prediction using massively parallel sequencing data and multiple machine learning models.* Forensic Sci Int Genet, 2018. **37**: p. 215-226.

29. Boks, M.P., et al., *The relationship of DNA methylation with age, gender and genotype in twins and healthy controls.* PLoS One, 2009. **4**(8): p. e6767.

30. Rakyan, V.K., et al., *Human aging-associated DNA hypermethylation occurs preferentially at bivalent chromatin domains.* Genome Res, 2010. **20**(4): p. 434-9.

31. Teschendorff, A.E., et al., *Age-dependent DNA methylation of genes that are suppressed in stem cells is a hallmark of cancer.* Genome Res, 2010. **20**(4): p. 440-6.

32. Bocklandt, S., et al., *Epigenetic predictor of age.* PLoS One, 2011. **6**(6): p. e14821.

33. Koch, C.M., et al., *Specific age-associated DNA methylation changes in human dermal fibroblasts.* PLoS One, 2011. **6**(2): p. e16679.

34. Hernandez, D.G., et al., *Distinct DNA methylation changes highly correlated with chronological age in the human brain.* Human Molecular Genetics, 2011. **20**(6): p. 1164-1172.

35. Martino, D.J., et al., *Evidence for age-related and individual-specific changes in DNA methylation profile of mononuclear cells during early immune development in humans.* Epigenetics, 2011. **6**(9): p. 1085-94.

36. Bell, J.T., et al., *Epigenome-wide scans identify differentially methylated regions for age and age-related phenotypes in a healthy ageing population.* PLoS Genet, 2012. **8**(4): p. e1002629.

37. Horvath, S., et al., *Aging effects on DNA methylation modules in human brain and blood tissue.* Genome Biol, 2012. **13**(10): p. R97.

38. Garagnani, P., et al., *Methylation of ELOVL2 gene as a new epigenetic marker of age.* Aging Cell, 2012. **11**(6): p. 1132-1134.

39. Alisch, R.S., et al., *Age-associated DNA methylation in pediatric populations.* Genome Research, 2012. **22**(4): p. 623-632.

40. Numata, S., et al., *DNA methylation signatures in development and aging of the human prefrontal cortex.* Am J Hum Genet, 2012. **90**(2): p. 260-72.

41. Teschendorff, A.E., J. West, and S. Beck, *Age-associated epigenetic drift: implications, and a case of epigenetic thrift?* Hum Mol Genet, 2013. **22**(R1): p. R7-R15.

42. Day, K., et al., *Differential DNA methylation with age displays both common and dynamic features across human tissues that are influenced by CpG landscape.* Genome Biol, 2013. **14**(9): p. R102.

43. Hannum, G., et al., *Genome-wide methylation profiles reveal quantitative views of human aging rates.* Mol Cell, 2013. **49**(2): p. 359-367.

44. Hollegaard, M.V., et al., *DNA methylome profiling using neonatal dried blood spot samples: a proof-of-principle study.* Mol Genet Metab, 2013. **108**(4): p. 225-31.

45. Johansson, A., S. Enroth, and U. Gyllensten, *Continuous Aging of the Human DNA Methylome Throughout the Human Lifespan.* PLoS One, 2013. **8**(6): p. e67378.

46. Zykovich, A., et al., *Genome-wide DNA methylation changes with age in disease-free human skeletal muscle.* Aging Cell, 2014. **13**(2): p. 360-366.

47. Martino, D., et al., *Longitudinal, genome-scale analysis of DNA methylation in twins from birth to 18 months of age reveals rapid epigenetic change in early life and pair-specific effects of discordance.* Genome Biol, 2013. **14**(5): p. R42.

48. Florath, I., et al., *Cross-sectional and longitudinal changes in DNA methylation with age: an epigenome-wide analysis revealing over 60 novel age-associated CpG sites.* Hum Mol Genet, 2014. **23**(5): p. 1186-201.

49. Weidner, C.I., et al., *Aging of blood can be tracked by DNA methylation changes at just three CpG sites.* Genome Biol, 2014. **15**(2): p. R24.

50. Steegenga, W.T., et al., *Genome-wide age-related changes in DNA methylation and gene expression in human PBMCs.* Age, 2014. **36**(3): p. 9648.

51. Marttila, S., et al., *Ageing-associated changes in the human DNA methylome: genomic locations and effects on gene expression.* BMC Genomics, 2015. **16**(1): p. 179.

52. McClay, J.L., et al., *A methylome-wide study of aging using massively parallel sequencing of the methyl-CpG-enriched genomic fraction from blood in over 700 subjects.* Hum Mol Genet, 2014. **23**(5): p. 1175-85.

53. Bekaert, B., et al., *Improved age determination of blood and teeth samples using a selected set of DNA methylation markers.* Epigenetics, 2015. **10**(10): p. 922-30.

54. Huang, Y., et al., *Developing a DNA methylation assay for human age prediction in blood and bloodstain.* Forensic Sci Int Genet, 2015. **17**: p. 129-136.

55. Lee, H.Y., et al., *Epigenetic age signatures in the forensically relevant body fluid of semen: a preliminary study.* Forensic Sci Int Genet, 2015. **19**: p. 28-34.

56. Soares Bispo Santos Silva, D., et al., *Evaluation of DNA methylation markers and their potential to predict human aging.* Electrophoresis, 2015. **36**(15): p. 1775-80.

57. Yi, S.H., et al., *Age-related DNA methylation changes for forensic age-prediction.* International Journal of Legal Medicine, 2015. **129**(2): p. 237-244.

58. Zaghlool, S.B., et al., *Association of DNA methylation with age, gender, and smoking in an Arab population.* Clin Epigenetics, 2015. **7**: p. 6.

59. Xu, C., et al., *A novel strategy for forensic age prediction by DNA methylation and support vector regression model.* Scientific Reports, 2015. **5**: p. 17788.

60. Acevedo, N., et al., *Age-associated DNA methylation changes in immune genes, histone modifiers and chromatin remodeling factors within 5 years after birth in human blood leukocytes.* Clinical Epigenetics, 2015. **7**(1): p. 34.

61. Peters, M.J., et al., *The transcriptional landscape of age in human peripheral blood.* Nat Commun, 2015. **6**: p. 8570.

62. Zubakov, D., et al., *Human age estimation from blood using mRNA, DNA methylation, DNA rearrangement, and telomere length.* Forensic Sci Int Genet, 2016. **24**: p. 33-43.

63. Park, J.L., et al., *Identification and evaluation of age-correlated DNA methylation markers for forensic use.* Forensic Sci Int Genet, 2016. **23**: p. 64-70.

64. Freire-Aradas, A., et al., *Development of a methylation marker set for forensic age estimation using analysis of public methylation data and the Agena Bioscience EpiTYPER system.* Forensic Sci Int Genet, 2016. **24**: p. 65-74.

65. Kananen, L., et al., *Cytomegalovirus infection accelerates epigenetic aging.* Exp Gerontol, 2015. **72**: p. 227-9.

66. Vidal-Bralo, L., Y. Lopez-Golan, and A. Gonzalez, *Simplified Assay for Epigenetic Age Estimation in Whole Blood of Adults.* Frontiers in Genetics, 2016. **7**: p. 126.

67. Knight, A.K., et al., *An epigenetic clock for gestational age at birth based on blood methylation data.* Genome Biol, 2016. **17**(1): p. 206.

68. Tan, Q., et al., *Epigenetic drift in the aging genome: a ten-year follow-up in an elderly twin cohort.* International Journal of Epidemiology, 2016. **45**(4): p. 1146-1158.

69. Hong, S.R., et al., *DNA methylation-based age prediction from saliva: High age predictability by combination of 7 CpG markers.* Forensic Sci Int Genet, 2017. **29**: p. 118-125.

70. Mayne, B.T., et al., *Accelerated placental aging in early onset preeclampsia pregnancies identified by DNA methylation.* Epigenomics, 2017. **9**(3): p. 279-289.

71. Cho, S., et al., *Independent validation of DNA-based approaches for age prediction in blood.* Forensic Sci Int Genet, 2017. **29**: p. 250-256.

72. Benton, M.C., et al., *Methylome-wide association study of whole blood DNA in the Norfolk Island isolate identifies robust loci associated with age.* Aging (Albany NY), 2017. **9**(3): p. 753-768.

73. Xu, C.J., et al., *The emerging landscape of dynamic DNA methylation in early childhood.* BMC Genomics, 2017. **18**(1): p. 25.

74. Barrett, T., et al., *NCBI GEO: archive for functional genomics data sets--update.* Nucleic Acids Res, 2013. **41**(Database issue): p. D991-5.

75. Anjum, S., et al., *A BRCA1-mutation associated DNA methylation signature in blood cells predicts sporadic breast cancer incidence and survival.* Genome Medicine, 2014. **6**(6): p. 47.

76. Chen, Y.A., et al., *Sequence overlap between autosomal and sex-linked probes on the Illumina HumanMethylation27 microarray.* Genomics, 2011. **97**(4): p. 214-22.

77. Horvath, S. and A.J. Levine, *HIV-1 Infection Accelerates Age According to the Epigenetic Clock.* The Journal of Infectious Diseases, 2015. **212**(10): p. 1563-1573.

78. Liu, Y., et al., *Epigenome-wide association data implicate DNA methylation as an intermediary of genetic risk in rheumatoid arthritis.* Nat Biotechnol, 2013. **31**(2): p. 142-7.

79. Harris, R.A., et al., *Genome Wide Peripheral Blood Leukocyte DNA Methylation Microarrays Identified a Single Association with Inflammatory Bowel Diseases.* Inflammatory bowel diseases, 2012. **18**(12): p. 10.1002/ibd.22956.

80. Bell, J.T., et al., *Differential methylation of the TRPA1 promoter in pain sensitivity.* Nature Communications, 2014. **5**: p. 2978.

81. Horvath, S., et al., *An epigenetic clock analysis of race/ethnicity, sex, and coronary heart disease.* Genome Biology, 2016. **17**(1): p. 171.

82. Du, P., et al., *Comparison of Beta-value and M-value methods for quantifying methylation levels by microarray analysis.* BMC Bioinformatics, 2010. **11**(1): p. 587.

83. Biosystems, A., *Quantifiler™ HP and Trio DNA Quantification Kits User Guide.* Thermo Fisher Scientific, 2017.

84. Corporation, P., *MethylEdge™ Bisulfite Conversion System Instructions for use of product N1301.* Promega Corporation, 2013.

85. Leontiou, C.A., et al., *Bisulfite Conversion of DNA: Performance Comparison of Different Kits and Methylation Quantitation of Epigenetic Biomarkers that Have the Potential to Be Used in Non-Invasive Prenatal Testing.* PLOS ONE, 2015. **10**(8): p. e0135058.

86. Li, L.C. and R. Dahiya, *MethPrimer: designing primers for methylation PCRs.* Bioinformatics, 2002. **18**(11): p. 1427-31.

87. Yates, A.D., et al., *Ensembl 2020.* Nucleic Acids Res, 2020. **48**(D1): p. D682-D688.

88. Naue, J., et al., *Chronological age prediction based on DNA methylation: Massive parallel sequencing and random forest regression.* Forensic Sci Int Genet, 2017. **31**: p. 19-28.

89. Qiagen, *MinElute® PCR Purification Kit Quick Start Protocol.* Qiagen Sample and Assay Technologies, 2011.

90. Technologies, L., *User Guide: Qubit® dsDNA HS Assay Kits for use with the Qubit® Fluorometer (all models).* Life Technologies Molecular Probes, 2015.

91. BioLabs, N.E., *NEBNext® Ultra™ II DNA Library Prep Kit for Illumina® (E7645, E7103) Instruction Manual*, N.E. BioLabs, Editor.

92. Biosystems, K., *KAPA Hyper Prep Kit Technical Data Sheet KR0961 – v5.16.* KAPA Biosystems, 2016.

93. Biosystems, K., *KAPA Library Quantification Kit Technical Data Sheet KR0405 – v8.17.* KAPA Biosystems, 2017.

94. Li, H. and R. Durbin, *Fast and accurate short read alignment with Burrows–Wheeler transform.* Bioinformatics, 2009. **25**(14): p. 1754-1760.

95. Li, H., et al., *The Sequence Alignment/Map format and SAMtools.* Bioinformatics, 2009. **25**(16): p. 2078-2079.

96. McKenna, A., et al., *The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data.* Genome Research, 2010. **20**(9): p. 1297-1303.

97. Li, E. and Y. Zhang, *DNA Methylation in Mammals.* Cold Spring Harbor Perspectives in Biology, 2014. **6**(5): p. a019133.

98. Gruenbaum, Y., et al., *Methylation of CpG sequences in eukaryotic DNA.* FEBS Letters, 1981. **124**(1): p. 67-71.

99. Team, R.C., *R: A language and environment for statistical computing. R Foundation for Statistical Computing*. 2020: Vienna, Austria.

100. Kuhn, M., *Building Predictive Models in R Using the caret Package.* Journal of Statistical Software, 2008. **28**(5): p. 26.

101. Xiong, Z., et al., *EWAS Data Hub: a resource of DNA methylation array data and metadata.* Nucleic Acids Res, 2020. **48**(D1): p. D890-D895.

102. Mi, H., et al., *PANTHER version 14: more genomes, a new PANTHER GO-slim and improvements in enrichment analysis tools.* Nucleic Acids Res, 2019. **47**(D1): p. D419-D426.

103. Mi, H., et al., *Protocol Update for large-scale genome and gene function analysis with the PANTHER classification system (v.14.0).* Nat Protoc, 2019. **14**(3): p. 703-721.

104. Mi, H. and P. Thomas, *PANTHER pathway: an ontology-based pathway database coupled with data analysis tools.* Methods Mol Biol, 2009. **563**: p. 123-40.

105. Huang da, W., B.T. Sherman, and R.A. Lempicki, *Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists.* Nucleic Acids Res, 2009. **37**(1): p. 1-13.

106. Huang da, W., B.T. Sherman, and R.A. Lempicki, *Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources.* Nat Protoc, 2009. **4**(1): p. 44-57.

107. Huang, D.W., B.T. Sherman, and R.A. Lempicki, *Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists.* Nucleic Acids Research, 2009. **37**(1): p. 1-13.

108. Kanehisa, M., et al., *New approach for understanding genome variations in KEGG.* Nucleic Acids Res, 2019. **47**(D1): p. D590-D595.

109. Becker, K.G., et al., *The genetic association database.* Nat Genet, 2004. **36**(5): p. 431-2.

110. Freire-Aradas, A., et al., *A Comparison of Forensic Age Prediction Models Using Data From Four DNA Methylation Technologies.* Front Genet, 2020. **11**: p. 932.

111. Hamano, Y., et al., *Forensic age prediction for dead or living samples by use of methylation-sensitive high resolution melting.* Leg Med, 2016. **21**: p. 5-10.

112. Naue, J., et al., *Proof of concept study of age-dependent DNA methylation markers across different tissues by massive parallel sequencing.* Forensic Sci Int Genet, 2018. **36**: p. 152-159.

113. Office, H., *National DNA Database Strategy Board Biennial Report 2018-2020*. 2020.

114. Atkinson, K., et al., *Female marrow donors increase the risk of acute graft-versus-host disease: effect of donor age and parity and analysis of cell subpopulations in the donor marrow inoculum.* Br J Haematol, 1986. **63**(2): p. 231-9.

115. Zhang, Y., et al., *DNA methylation signatures in peripheral blood strongly predict all-cause mortality.* Nat Commun, 2017. **8**: p. 14617.

116. Nevalainen, T., et al., *Obesity accelerates epigenetic aging in middle-aged but not in elderly individuals.* Clin Epigenetics, 2017. **9**: p. 20.

117. Lopez-Otin, C., et al., *The hallmarks of aging.* Cell, 2013. **153**(6): p. 1194-217.

118. Lopez-Otin, C., et al., *Metabolic Control of Longevity.* Cell, 2016. **166**(4): p. 802-821.

119. Ma, S., et al., *Organization of the Mammalian Metabolome according to Organ Function, Lineage Specialization, and Longevity.* Cell Metab, 2015. **22**(2): p. 332-43.

120. Fontana, L. and L. Partridge, *Promoting health and longevity through diet: from model organisms to humans.* Cell, 2015. **161**(1): p. 106-118.

121. Zbiec-Piekarska, R., et al., *Development of a forensically useful age prediction method based on DNA methylation analysis.* Forensic Sci Int Genet, 2015. **17**: p. 173-179.

SUPPLEMENTARY TABLES

Supplementary_Table_S1. Literature investigated for the identification of age-related CpG sites, presented in chronological order.

No.	Study	Tissue	Number of CpGs	Ref.
1	Boks *et al.* 2009	Whole blood	6	[29]
2	Rakyan *et al.* 2010	Whole blood	131	[30]
3	Teschendorff *et al.* 2010	Whole blood	411	[31]
4	Bocklandt *et al.* 2011	Saliva	3	[32]
5	Koch *et al.* 2011	Multi-tissue	5	[6]
6	Koch *et al.* 2011	Dermal fibroblasts	31	[33]
7	Hernandez *et al.* 2011	Brain tissues	10	[34]
8	Martino *et al.* 2011	Cord and Whole blood	1030	[35]
9	Bell *et al.* 2012	Whole blood	490	[36]
10	Horvath *et al.* 2012	Blood and Brain tissues	1000	[37]
11	Garagnani *et al.* 2012	Whole blood	9	[38]
12	Alisch *et al.* 2012	Whole blood	2078	[39]
13	Numata *et al.* 2012	Prefrontal cortex	300	[40]
14	Teschendorff *et al.* 2013	Multi-tissue	67	[41]
15	Day *et al.* 2013	Multi-tissue	431	[42]
16	Hannum *et al.* 2013	Whole blood	71	[43]
17	Hollegaard *et al.* 2013	Whole blood	68	[44]
18	Johansson *et al.* 2013	White blood cells	1	[45]
19	Zykovich *et al.* 2013	Skeletal muscle	500	[46]
20	Martino *et al.* 2013	Buccal	2632	[47]
21	Horvath 2013	Multi-tissue	353	[3]
22	Almen *et al.* 2014	Whole blood	25	[21]
23	Florath *et al.* 2014	Whole blood	17	[48]
24	Weidner *et al.* 2014	Whole blood	3	[49]
25	Yi *et al.* 2014	Whole blood	16	[1]

26	Steegenga *et al.* 2014	Peripheral blood cells	719	[50]
27	Marttila *et al.* 2014	Peripheral blood cells	8540	[51]
28	McClay et al. 2014	Whole blood	70	[52]
29	Zbiec-Piekarska *et al*. 2015	Whole blood	5	[8]
30	Bekaert *et al.* 2015	Blood and Teeth	4	[53]
31	Huang *et al.* 2015	Whole blood	4	[54]
32	Lee *et al.* 2015	Semen	3	[55]
33	Mansego *et al.* 2015	White blood cells	54	[20]
34	Soares Bispo Santos Silva *et al.* 2015	Blood and Saliva	2	[56]
35	Yi *et al.* 2015	Blood and Saliva	3	[57]
36	Zaghlool *et al.* 2015	Whole blood	674	[58]
37	Xu *et al.* 2015	Whole blood	2965	[59]
38	Acevedo *et al.* 2015	Blood leukocytes	794	[60]
39	Peters *et al.* 2015	Whole blood	1497	[61]
40	Zubakov *et al.* 2016	Whole blood	75	[62]
41	Park *et al.* 2016	Whole blood	582	[63]
42	Freire-Aradas *et al.* 2016	Whole blood	177	[64]
43	Kananen *et al.* 2016	Whole blood	1202	[65]
44	Vidal-Bralo *et al.* 2016	Whole blood	8	[66]
45	Knight *et al.* 2016	Blood tissues	148	[67]
46	Tan *et al.* 2016	Whole blood	2284	[68]
47	Hong *et al.* 2017	Saliva	62	[69]
48	Mayne *et al.* 2017	Placental tissue	62	[70]
49	Cho *et al.* 2017	Whole blood	32	[71]
50	Benton *et al.* 2017	Whole blood	497	[72]
51	Xu *et al.* 2017	Whole blood	14150	[73]

Supplementary_Table_S2. Datasets used for the collection of DNA methylation data on the 5364 selected CpGs.

No.	Accession number	Tissue	Sample size	Age range (years)	Platform	Ref.
1	GSE41037	Whole blood	391	16 - 88	27k[1]	[37]
2	GSE44763	Peripheral whole blood	46	41 - 70	27k	[21]
3	GSE57285	Whole blood	41	19 - 71	27k	[75]
4	GSE19711	Whole blood	268	52 - 78	27k	[31]
5	GSE20236	Whole blood	15	53 - 71	27k	[30]
6	GSE27097	Peripheral blood leukocyte cells	398	3 - 17	27k	[39]
7	GSE20242	Sorted human blood cells	20	16 - 69	27k	[30]
8	GSE23638	Whole blood lymphocytes	23	2 - 33	27k	[76]
9	GSE58045	Blood samples	97	32 - 80	27k	[36]
10	GSE67751	Blood samples	69	35 - 65	450k[2]	[77]
11	GSE40279	Whole blood	656	19 - 101	450k	[43]
12	GSE41169	Whole blood	32	18 - 65	450k	[37]
13	GSE42861	Whole blood	335	20 - 70	450k	[78]
14	GSE32148	Peripheral whole blood	19	3 - 76	450k	[79]
15	GSE36064	Leukocytes	78	1 - 16	450k	[39]
16	GSE40005	Blood samples	10	53 - 68	450k	N.A.[3]
17	GSE53740	Peripheral whole blood	165	37 - 93	450k	[80]
18	GSE49064	Peripheral whole blood	10	30 - 66	450k	[50]

		mononuclear cells (PBMCs)				
19	GSE65638	Blood samples	8	21 - 32	450k	[59]
20	GSE84624	Peripheral blood	24	0.5 - 6	450k	N.A.[3]
21	GSE87571	Whole blood	671	14 - 94	450k	[45]
22	GSE72775	Whole blood	335	36 - 91	450k	[81]
23	GSE72777	Whole blood	46	2 - 35	450k	[81]
24	GSE72773	Whole blood	310	35 - 92	450k	[81]

[1] 27k: assay conducted on Illumina Infinium HumanMethylation27 BeadChip platform

[2] 450k: assay conducted on Illumina Infinium Human Methylation450 BeadChip platform

[3] N.A.: not applicable as no journal article is referenced with this dataset

Supplementary_Table_S3. 244 CpG markers with |r|≥0.70, or |r|≥0.65 and methylation range above 70% over the human lifespan.

cg16867657	cg08262002	cg24892069	cg19344626	cg27401724	cg21120249
cg22454769	cg12934382	cg00602811	cg16193278	cg08877357	cg07164639
cg10501210	cg17471102	cg11649376	cg18651026	cg26725076	cg01719405
cg22736354	cg00503840	cg14359680	cg02867102	cg07027613	cg23320649
cg01820374	cg26685941	cg27015931	cg23124451	cg11807280	cg09118625
cg19283806	cg05308819	cg18150280	cg15804973	cg12580096	cg23341182
cg25256723	cg20273670	cg23744638	cg10221746	cg08713098	cg14956327
cg06639320	cg22016779	cg00101260	cg03224418	cg08644498	cg20067719
cg09809672	cg06247837	cg01243823	cg20153322	cg18034299	cg22768222
cg04875128	cg20822990	cg24847230	cg25538571	cg20988565	cg25809905
cg02228185	cg15948836	cg19761273	cg05331060	cg18568843	cg05379350
cg24079702	cg21296230	cg07388493	cg00863306	cg25994988	cg02838877
cg00329615	cg13033938	cg14556683	cg04503319	cg21186955	cg09636661
cg07082267	cg04604946	ch.2.30415474F	cg12483947	cg16983588	cg03043157
cg24724428	cg06268694	cg22580512	cg15037004	cg01234420	cg00308665
cg21572722	cg20669012	cg06911110	cg26543112	cg18826637	cg27192248

cg07553761	cg20222376	cg23078123	ch.1.171672612F	cg22943590	cg01812894
cg16008966	cg17183905	cg13823169	cg10149533	cg08888956	cg23479922
cg22156456	cg06419432	cg10247798	cg22947000	cg19991948	cg22082462
cg14361627	cg12261786	cg06567855	cg27209729	cg18450254	cg25711003
cg18933331	cg12939283	cg20052760	cg07583137	cg13221458	cg19663246
cg08234504	cg02046143	cg02030542	cg26969888	cg10804656	cg18186343
cg16762684	ch.6.33611621F	cg04742397	cg21469505	cg02872426	cg04123409
cg01974375	cg08468401	cg12711760	cg03746976	cg23836737	cg10872209
cg03996822	cg20816447	cg18079948	cg26894354	cg15894389	cg05042708
cg22796704	cg04581938	cg21990700	cg23715749	cg01459453	cg18797590
cg11741201	cg22483030	cg15845821	cg04474832	cg01282174	cg00664406
cg25533247	cg00573770	cg23950157	cg14042143	cg13327545	cg12317815
cg03431918	cg27320127	cg08453194	cg14583999	cg20102280	cg09124496
cg08090640	cg16054275	cg22730004	cg08553327	cg19848940	cg06493994
cg26350754	cg05207048	cg25428494	cg07211259	cg10835286	cg20747538
cg02286081	cg22273555	cg07080372	cg00292135	cg00548268	cg21801378
cg08128734	cg18738190	cg12623930	cg11436113	cg24768561	cg17168836
cg17372101	cg18215449	cg26608718	cg04411841	cg09552402	cg06285727
cg16744741	cg05156137	cg18182399	cg05619598	cg10917602	cg13959344
cg03725309	cg24212517	cg25537245	cg23500537	cg22737154	cg09278098
cg14195318	cg11693709	cg26815395	cg04425624	cg05412028	cg20692569
cg04208403	cg21922223	cg14314729	cg12079303	cg14747813	cg21878650
cg18618815	cg17457912	cg05584950	cg11194994	cg27210390	cg06279276
cg06874016	cg08097417	cg17721618	cg05404236	cg25413977	
cg08160331	cg19722847	cg04416734	cg10650821	cg15538427	

Supplementary_Table_S4. Primer sequences for the 19 markers. Amplicon lengths are also displayed.

CpG site	Associated Genes	Primer Sequence (5'-3')		Amplicon length (bp)
cg16867657	*ELOVL2*	F	AGGGGYGTAGGGTAAGTGAGG	308
cg21572722		R	AACAAAACCATTTCCCCCTAATAT	
cg24724428				
cg06639320	*FHL2*	F	GTTTTTGGGATTAGGTAGAGATTT	165
cg22454769		R	TTTATTTACCAAAACTCCTTTCTTC	
cg24079702				
cg00329615	*IGSF11*	F	TATGTGTTTGAGATTTGGTAGGTT	181
		R	TTATTCATTCATTATTCTCCTTAAAAAAAT	
cg09809672	*EDARADD*	F	GGTTTGATTTTGGTTAGATAATTAG	148
		R	AAAAACTTTAATACCTCTCCCCATC	
cg22796704	*ARHGAP22*	F	GGATTTAGGGGTAGGTAGAATTTGT	148
		R	TCTAAACTAAACTTAACCACCTTCC	
cg08128734	*RASSF5*	F	ATTTTGGGTATTTGGAAGGTATTT	189
		R	TCCCAATTAAAACCAAAAATAAAAA	
cg17372101	*CNTNAP2*	F	GTTTTAAAGTAGGTTAAGAAGTGGGAGT	124
		R	AAAACAAAAAATATCCCTAAATTTCCT	
cg08160331	*KLHL35*	F	TATTAAGAGGTAGTATTAAAAGATGATGAA	231
		R	CTTACTTCCTAAAAAAAATAAAAAC	
cg10501210	*MIR29B2CHG* *(C1orf132)*	F	AAGAAGGTGAGAAAGATAGAGTATTTATAT	210
		R	TAAAAAATTTAATAAAACCAAATTCTAAAA	
cg19283806	*CCDC102B*	F	GGGTTATAAGTTTTGTTTTGATGAAGT	171
		R	AATAAATTTCTCCTTAAACAATCCC	
cg07553761	*SMC4, TRIM59*	F	GTGGTTTGGGGGAGAGGT	86
		R	CCAAATAAAAAAATAATTCCTCAAAAAC	
cg08262002	*LDB2*	F	TTTTGGGTATTGAGTGAGGTATAGG	110
		R	ACCATTCATACATTCTAACAAAACC	
cg12934382	*GRM2*	F	GTTGGGTTGGGAGTAGGAGAT	284

		R	TAAAATAAAAACCAAAAAAATC	
cg17471102	*FUT3*	F	GGAGATTTTTTAGGAAAGGTTTTTT	144
		R	CTAACCACATTCCAAATCATAAACA	
cg18618815	*COL1A1*	F	GGTTGATAGGGATTTGTTTTTTAATT	180
		R	CCCCAAACCTAAAAATTCTTCTATAA	
Y represents a degenerate or ‘wobble’ base that is an equimolar mix of pyrimidines (T+C).				

Supplementary_Table_S5. Information on the Illumina 450K datasets used for assessing sex association in the age-correlated CpGs described in this work.

Accession number	Tissue	Sample size	♀	♂	Age range (years)	Ref.
GSE67751	Blood	69	45	24	35 - 65	[77]
GSE40279	Blood	656	338	318	19 - 101	[43]
GSE41169	Blood	32	12	20	18 - 65	[37]
GSE42861	Blood	335	239	96	20 - 70	[78]
GSE32148	Blood	19	12	7	3 - 76	[79]
GSE36064	Blood	78	0	78	1 - 16	[39]
GSE40005	Blood	10	4	6	53 - 68	-
GSE53740	Blood	165	102	63	37 - 93	[80]
GSE49064	Blood	10	0	10	30 - 66	[50]
GSE65638	Blood	8	8	0	21 - 32	[59]
GSE87571	Blood	729	388	341	14 - 94	[45]
GSE72775	Blood	335	138	197	36 - 91	[81]
GSE72777	Blood	46	31	15	2 - 35	[81]
GSE72773	Blood	310	150	160	35 - 92	[81]

Supplementary_Table_S6. Information on the Illumina 450K datasets used for assessing disease association in the age-correlated CpGs described in this work.

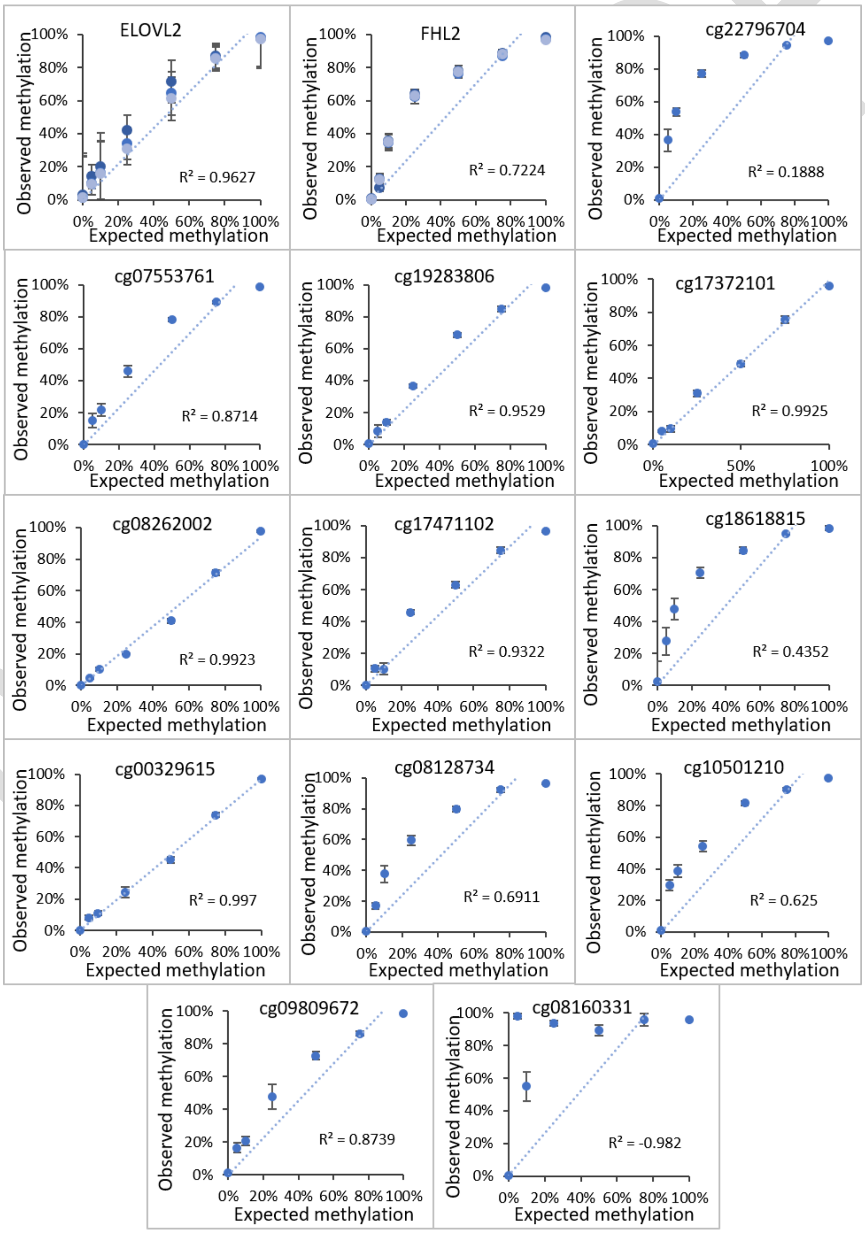
Accession number	Disease/ Condition	Tissue	Sample size	Age range (years)	Ref.
GSE41169	Schizophrenia	Whole blood	62	18 - 65	[37]
GSE42861	Rheumatoid arthritis	Whole blood	354	18 - 69	[78]
GSE53740	Frontal temporal dementia	Peripheral whole blood	121	34 - 85	[80]
GSE53740	Progressive supranuclear palsy	Peripheral whole blood	42	54 - 85	[80]

Supplementary_Table_S7. Comparison between the Pearson's correlation scores (r) observed between methylation and age in the combined dataset of control samples, as well as within the individual datasets for control samples (C) and samples obtained from individuals with schizophrenia (SCZ), rheumatoid arthritis (RA), frontal temporal dementia (FTD) and progressive supranuclear palsy (PSP).

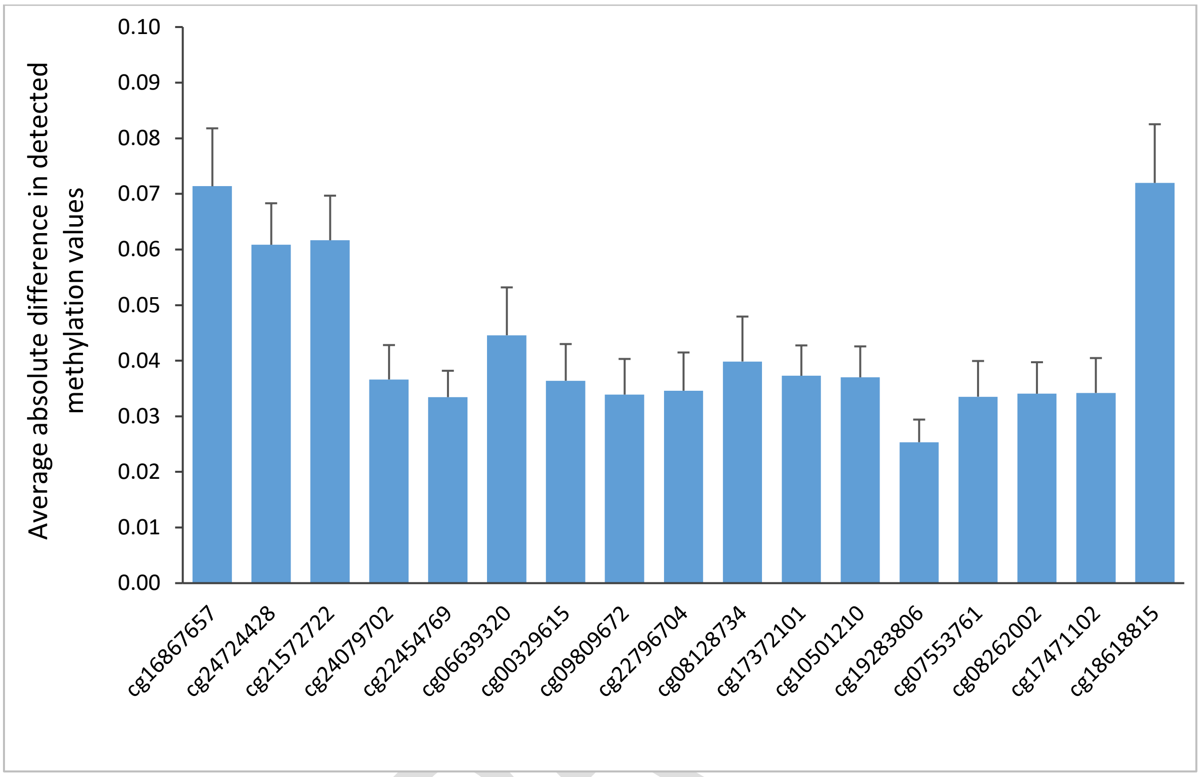
CpG marker	Associated Genes	Controls combined	GSE41169		GSE42861		GSE53740		
			C	SCZ	C	RA	C	FTD	PSP
cg24724428	*ELOVL2*	0.79	0.78	0.76	0.68	0.66	0.44	0.45	0.56
cg21572722	*ELOVL2*	0.76	0.79	0.82	0.76	0.81	0.51	0.44	0.49
cg06639320	*FHL2*	0.80	0.89	0.79	0.71	0.79	0.45	0.40	0.52
cg09809672	*EDARADD*	-0.78	-0.53	-0.51	-0.55	-0.60	-0.46	-0.40	-0.23
cg22796704	*ARHGAP22*	-0.75	-0.72	-0.53	-0.53	-0.57	-0.38	-0.22	-0.21
cg08128734	*RASSF5*	-0.74	-0.70	-0.72	-0.58	-0.58	-0.30	-0.26	-0.27
cg17372101	*CNTNAP2*	-0.73	-0.59	-0.39	-0.50	-0.44	-0.34	-0.33	-0.55
cg19283806	*CCDC102B*	-0.83	-0.54	-0.78	-0.59	-0.71	-0.39	-0.37	0.14
cg07553761	*SMC4, TRIM59*	0.75	0.84	0.88	0.61	0.60	0.41	0.32	0.50
cg10501210	*MIR29B2CHG*	-0.80	-0.90	-0.83	-0.68	-0.73	-0.50	-0.39	-0.52
cg08262002	*LDB2*	-0.74	-0.70	-0.51	-0.54	-0.64	-0.32	-0.35	-0.34

SUPPLEMENTARY FIGURES

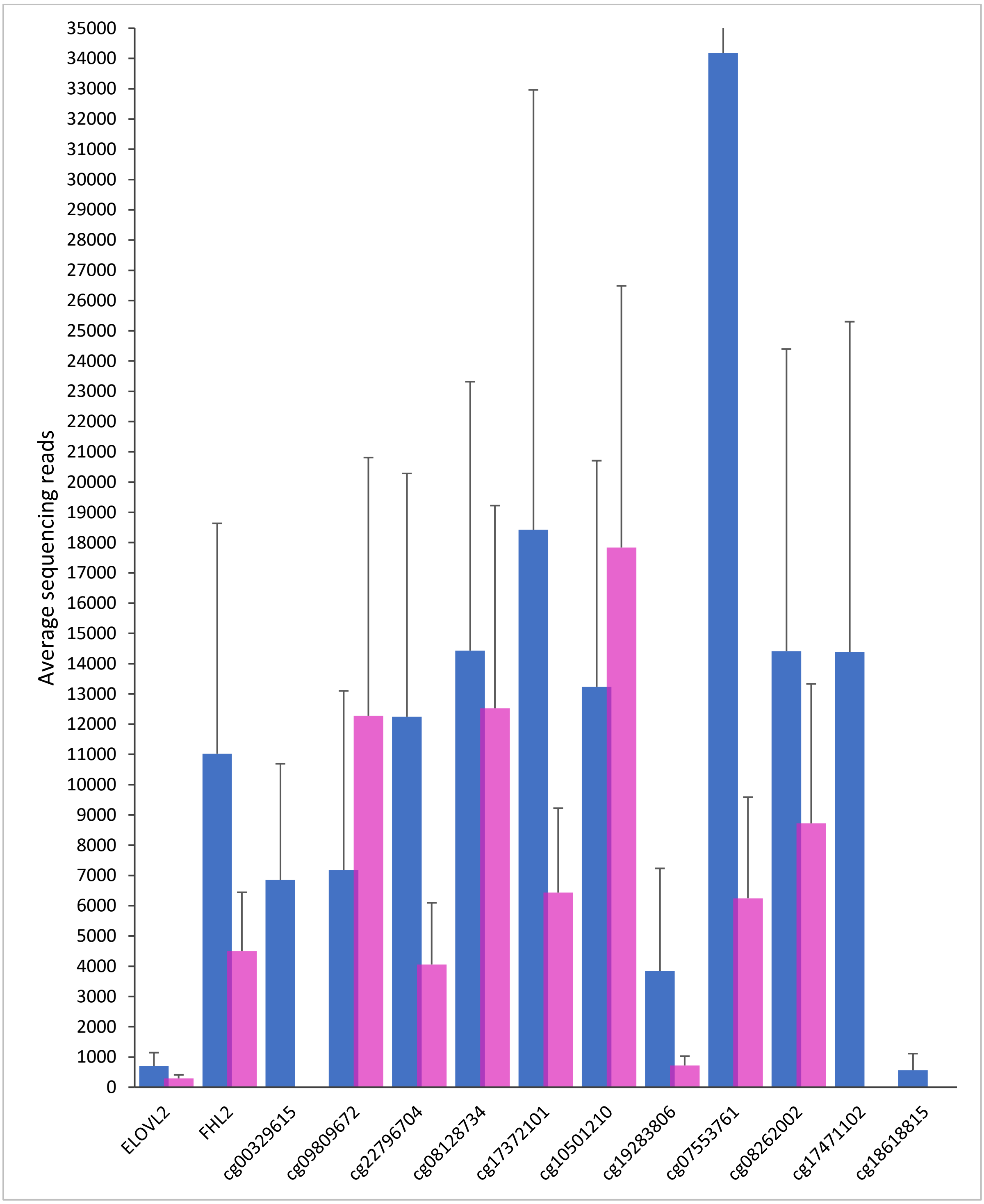
Supplementary_Fig_S1. Comparison between the observed and expected methylation values (β-values expressed as percentage of methylation) for the 18 markers analysed in this part of the study. Markers present in the same amplicon such as cg16867657, cg24724428, cg21572722 for ELOVL2 and cg06639320, cg22454769, cg24079702 for FHL2 are represented in the same graph. Primers for marker cg12934382 failed to yield amplification products and thus this marker is not represented here. Standards of known methylation (at 0%, 5%, 10%, 25%, 50%, 75% and 100% methylation) were processed in duplicate and the average value represents the 'observed' methylation fraction in the graphs. Error bars represent the standard error observed between duplicates and the R^2 values for the linear trendline (intercept set at 0) are displayed in each graph.



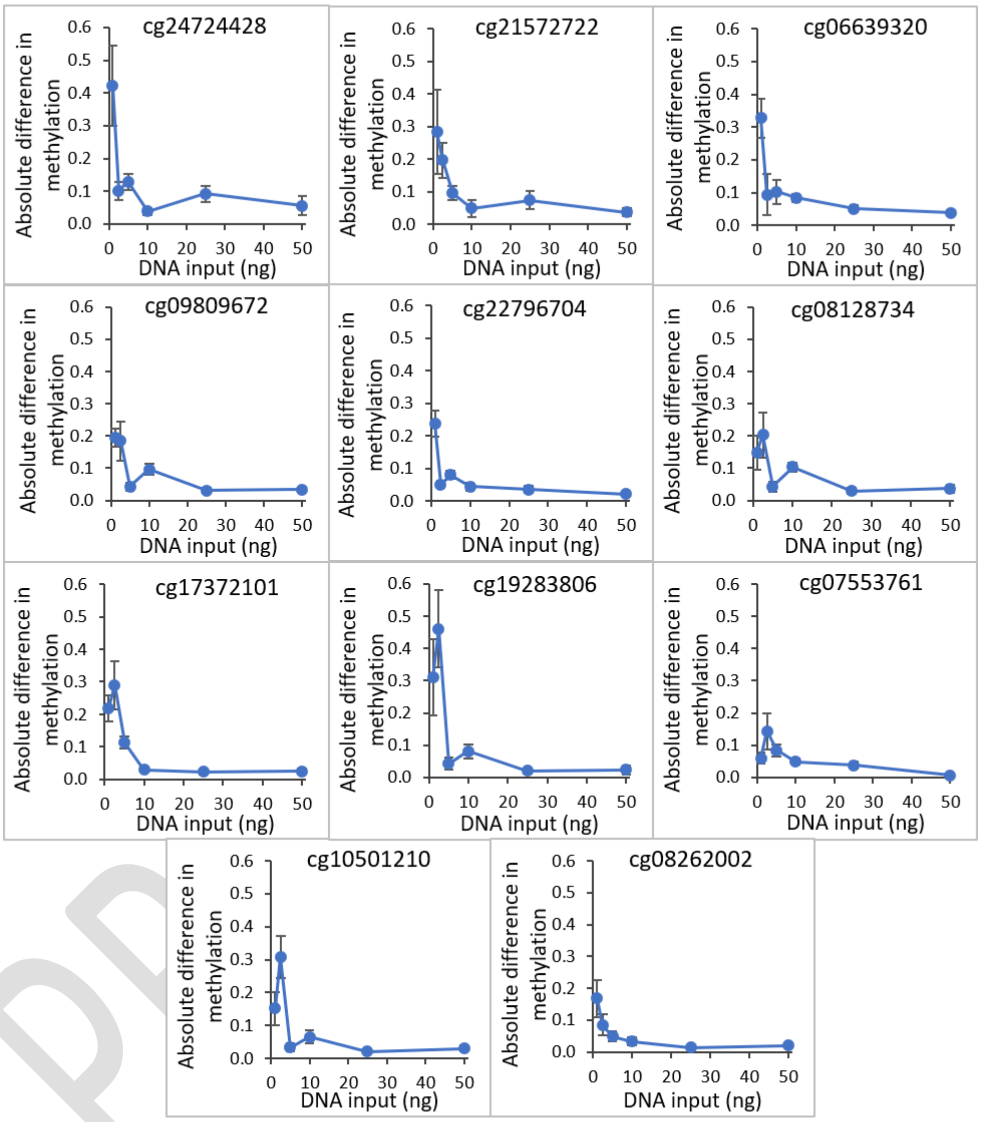
Supplementary_Fig_S2. Average absolute difference between the methylation β-values of samples analysed in duplicate (n=20) for the 17 different markers. The error bars represent the standard deviation.



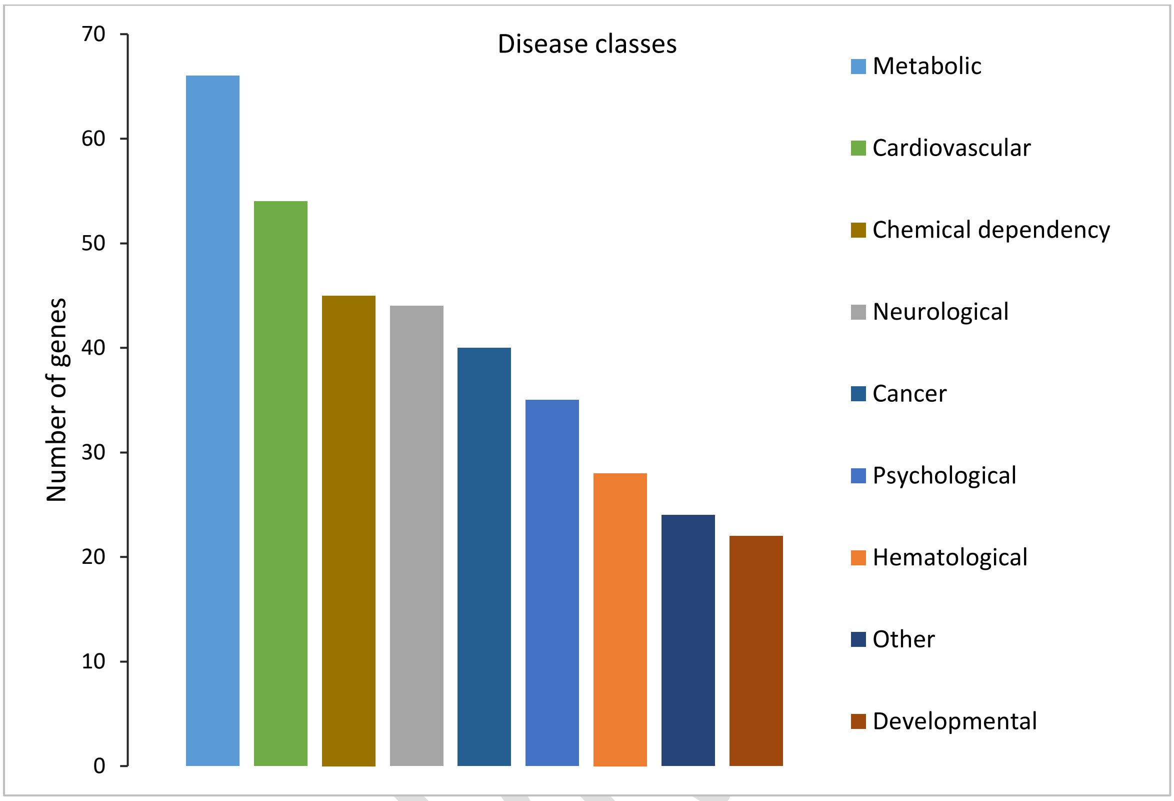
Supplementary_Fig_S3. Average sequencing reads obtained per amplicon in the 13-amplicon (17 markers, blue) and 10-amplicon (11 markers, purple) assays. Data for the 13-amplicon assay derive from the reproducibility study (section 3.2.2, n=40), whilst data from the sensitivity study (section 3.2.4, n=6) represent the 10-amplicon assay using the adapter-tagged primers (section 2.15). Error bars represent standard deviation.



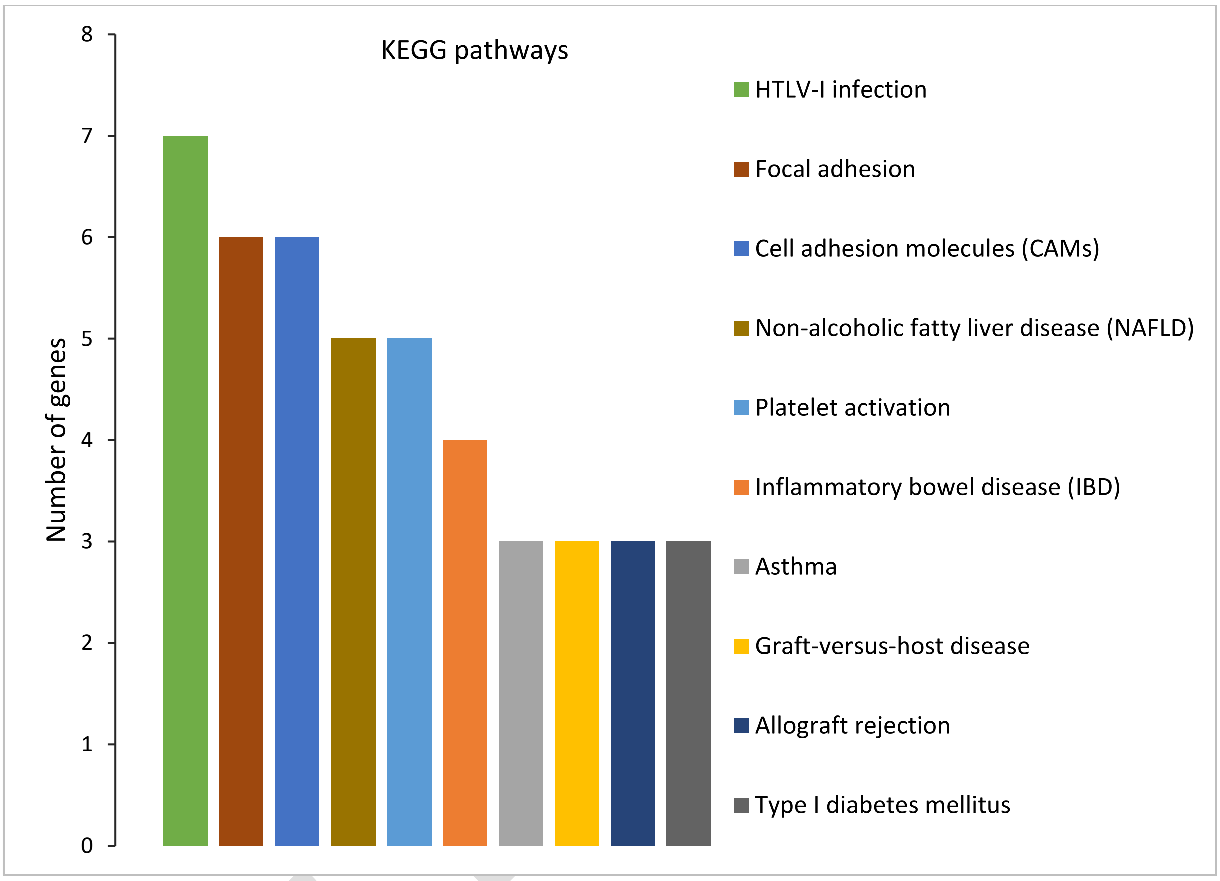
Supplementary_Fig_S4. Average absolute difference in the methylation β-values observed for 6 blood samples (from individuals aged 17, 27, 36, 43, 53 and 61 years) at each marker when 50, 25, 10, 5, 2.5 and 1 ng of DNA input was used as opposed to the original values obtained at 50 ng. The error bars represent the standard error of the difference between the methylation observed for each of the six samples and the average methylation observed for the original 50 ng input for the same sample during the development of the test set.



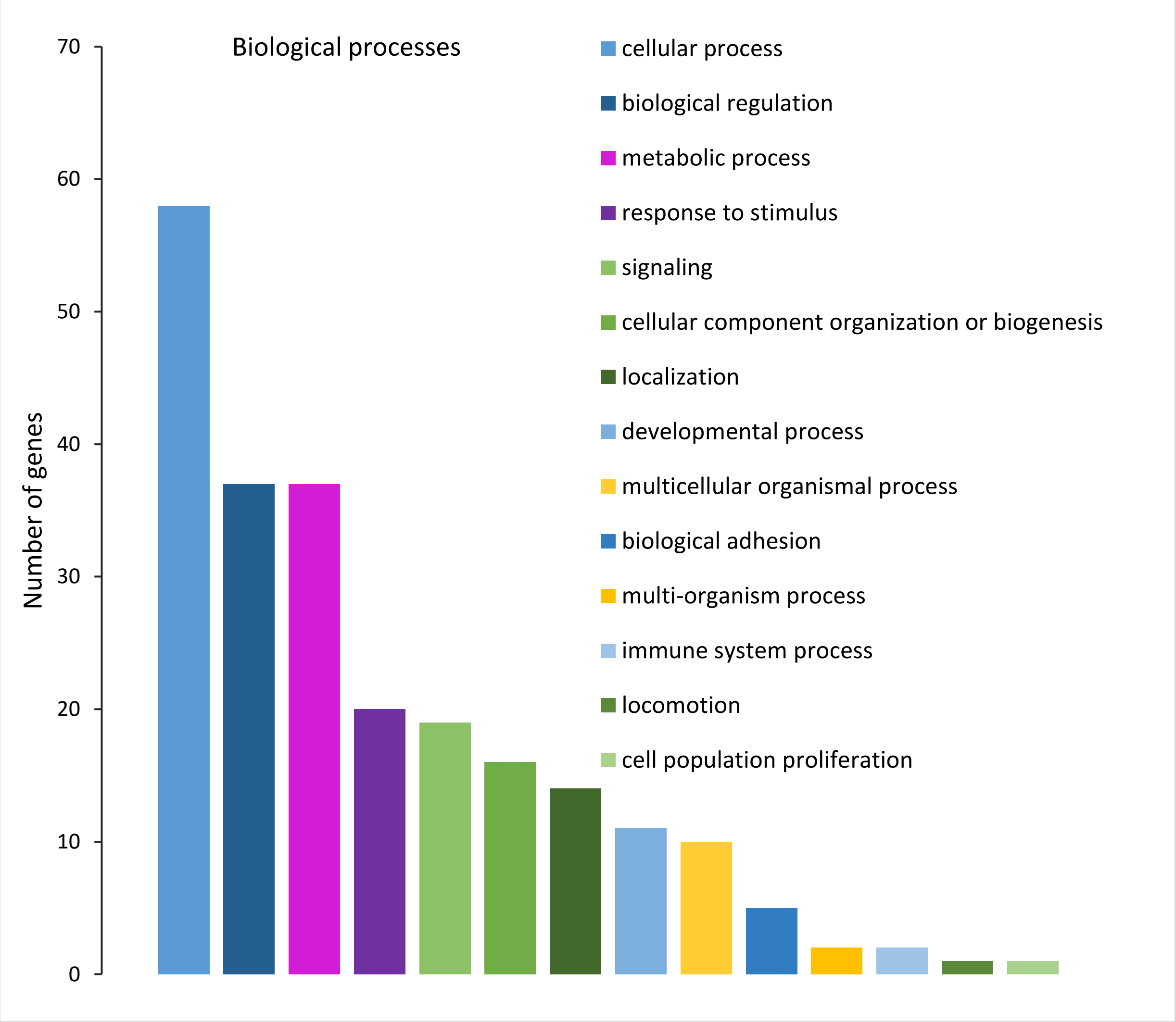
Supplementary_Fig_S5. Number of genes involved in biological pathway networks relating to different disease classes out of the 164 genes associated with the 244 markers identified for their correlation with chronological age in blood.



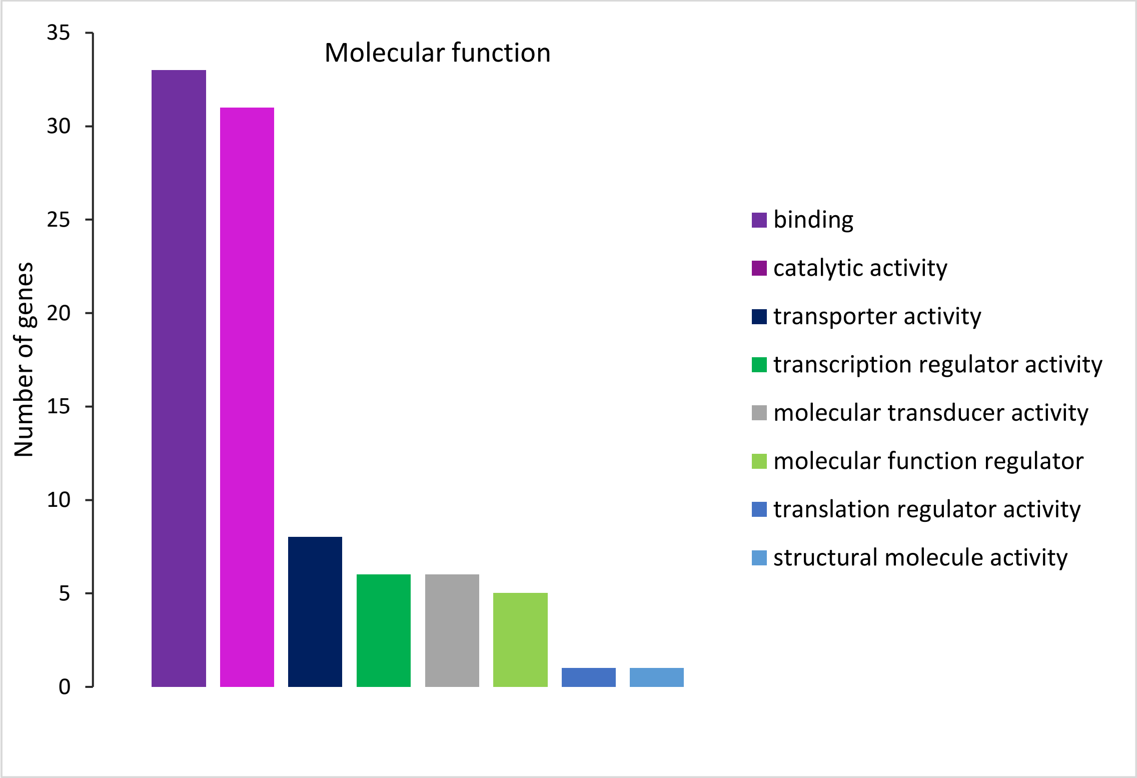
Supplementary_Fig_S6. Number of genes involved in KEGG pathway networks out of the 164 genes associated with the 244 markers identified for their correlation with chronological age in blood.



Supplementary_Fig_S7. Number of genes involved in the different biological processes for the 164 genes associated with the 244 markers identified for their correlation with chronological age in blood. In relation to the highest correlation groups (cellular process, biological regulation and metabolic process), the majority of genes associated with cellular processes (34/58 genes, 59%) were linked to proteins contributing to cellular metabolic and biosynthetic processes with groups of 16-19 genes were also associated with cell communication, cellular response to stimulus, signal transduction and cellular component organisation processes. Genes involved in biological regulation also showed a strong link to metabolic processes (19 genes associated involved in the regulation of cellular metabolic processes), signal transduction (17 genes) and the regulation of cellular communication (9 genes). Finally, associations with the metabolism of different compounds such as organic substances (35 genes) and nitrogen compounds (30 genes) were identified for the genes involved in metabolic processes.



Supplementary_Fig_S8. Number of genes the associated proteins showing activity in the various molecular functions. This graph relates to the 164 genes relating to the 244 markers identified for their correlation with chronological age in blood. In relation to the highest correlation groups (binding and catalytic activity), binding activity related heavily to protein binding (19 genes) as well as binding of organic cyclic compounds (10 genes), heterocyclic compounds (10 genes) and ions (8 genes). Catalytic activity related to hydrolase (13 genes) and transferase (12 genes) activity as well as activity affecting proteins (11 genes), such as protein kinase, peptidase and ubiquitin-like protein transferase activity.


Molecular function
Number of genes
■ binding
■ catalytic activity
■ transporter activity
■ transcription regulator activity
■ molecular transducer activity
■ molecular function regulator
■ translation regulator activity
■ structural molecule activity