DEPARTMENT OF PURE & APPLIED CHEMISTRY

University of Strathclyde Science

Forensic age estimation using DNA methylation analysis

By

Hussain Alsaleh

Centre for Forensic Science

Department of Pure and Applied Chemistry

University of Strathclyde

A thesis presented in fulfilment of the requirements for the degree of **Doctor of Philosophy**

2019

Centre for Forensic Science

Department of Pure and Applied Chemistry

University of Strathclyde

PhD Thesis

Hussain Alsaleh

2019

Signed:  Date: 26/09/2019

Abstract

Over the last decade, researchers have identified age-related (AR) DNA methylation (DNAm) markers, which have outperformed all other known AR biomarkers in estimating the chronological age of individuals with high accuracy, using various tissue types. Their accuracy in age estimation has led to them being suggested as a source of intelligence for forensic investigations, to determine the age of unidentified donors of biological samples left at crime scenes. Initially, in this research, different statistical methods have been tested in order to demonstrate which one of them is optimum for the identification of AR CpG sites. The selected method was then used to identify saliva specific AR CpG markers using DNAm profiles from saliva retrieved from an online genomic repository and assayed on the Illumina HumanMethylation450 BeadChip microarray. These AR CpG markers were used to build a saliva-specific age prediction model that was tested *in silico* on an independent saliva testing data set. They were shown to perform well in terms of age prediction, and consequently, they were further validated by targeted bisulfite sequencing of additional saliva samples, using the Illumina MiSeq® platform. Subsequently, a large cohort of 754 DNAm profiles from blood samples assayed on the newly launched Illumina MethylationEPIC® BeadChip were downloaded from an online genomic repository, in order to be tested for the first time for age association. Novel AR CpG sites were identified from both the newly added probes on this chip and from probes that were found on older platforms, however, the prediction accuracy of the blood-specific age prediction model did not improve compared to models built from the older Illumina HumanMethylation platforms. Finally, a multi-tissue age prediction model that is able to predict the age across the tissues was constructed. This multi-tissue age prediction model will have potential applications in forensic science in assisting investigations to predict chronological age across different biological samples, regardless of the tissue(s) those samples are derived from.

Acknowledgments

Firstly, I would like to thank my supervisors Dr. Penelope R. Haddrill, and Dr. Lorraine Gibson for their encouragement, support and guidance throughout the course of my PhD studies.

I would like to thank the Ministry of Interior of Kuwait, for sponsoring my studies and allowing me to fulfil the ambition of studying for a PhD. Without their assistance, this project would simply not have been possible.

I give thanks also to Nicola McCallum, and Daniel Halligan for their guidance and contributions towards my research during the first year of my PhD. I'm also grateful to Dr. Kirsty Ross for helping me in recruiting participants for my study, which made the project to go smoothly and quickly.

Finally, the biggest thanks go to my family and particularly my wife for her endless support and patience, without you, I would not be the person I am today.

Publications and presentations

Published articles:

H. Alsaleh, N.A. McCallum, D.L. Halligan, P.R. Haddrill. (2017). A multi-tissue age prediction model based on DNA methylation analysis. Forensic Science International: Genetics Supplement Series, 6, e62–e64. https://doi.org/10.1016/j.fsigss.2017.09.056

H. Alsaleh, P.R. Haddrill. (2019). Identifying blood-specific age-related DNA methylation markers on the Infinium MethylationEPIC® BeadChip. Forensic Science International, 303, 109944. https://doi.org/10.1016/j.forsciint.2019.109944

Submitted articles:

DNA methylation-based age prediction from saliva samples using next-generation sequencing on the Illumina MiSeq® platform. Submitted to Forensic Science International: Genetics (under review).

Poster presentations:

A multi-tissue age prediction model based on DNA methylation analysis. The 27th Congress of the International Society for Forensic Genetics, Seoul, South Korea (August 2017).

Identifying blood-specific age-related DNA methylation markers on the Infinium MethylationEPIC® BeadChip. The 28th Congress of the International Society for Forensic Genetics, Czech Republic, Prague (September 2019).

Oral presentation:

Age estimation using DNA methylation analysis. GCC forensic science conference, Abu Dhabi, Emirates (November 2017).

Table of contents

List of figures

List of tables

List of key abbreviations

AR:	Age related
BMIQ:	Beta- mixture quantile dilation normalization
CpG:	Cytosine-guanine
DNAm:	DNA methylation
DNMTs:	DNA methyltransferases
GWAS:	Genome-wide association study
HM27K:	Infinium HumanMethylation27
HM450K:	Infinium HumanMethylation450
MALDI-TOF:	Matrix-assisted laser desorption/ionization time-of-flight
MSP:	Methylation-specific PCR
MZ:	Monozygotic twins
SBE:	Single-base extension
SNPs:	Single nucleotide polymorphisms
SWAN:	Subset-quantile within array normalization
tDMRs:	Tissue-specific differentially methylated regions
TE	Tris and Ethylenediamine tetra-acetic acid (EDTA)

Chapter 1: Introduction

The development of DNA profiling for human identification has had a huge impact on the field of forensic science. It has been acknowledged as one of the greatest forensic advancements for its ability to establish a link between the criminal and the crime scene, and to allow the identification of the criminal. However, when recovered DNA evidence has no direct match with a suspect, and no indirect hit with a reference profile on the National DNA Database, this valuable evidence will be useless [1]. As such, tremendous effort has been made over the last ten years developing methods that can extract information about the externally visible characteristics of an unidentified person from their DNA. This has been successfully accomplished with the advancement of genome-wide association studies (GWAS), which test the statistical association between externally visible characteristics and up to a million genetic markers, which are usually single nucleotide polymorphisms (SNPs). SNP markers are DNA sequence variants that occur when a single nucleotide at a specific genomic site differs between individuals and can be significantly associated with a specific phenotype. For instance, predicting the hair and eye colour, as well as the ethnic background of the sample's donor can currently be achieved with high accuracy using specific SNP markers [2-4].

However, finding an association between the alleles present at SNP markers and human characteristics such as the age of the sample donor is not currently possible. This is due to the nature of the DNA sequence, which remains relatively static throughout a person's life. Furthermore, identifying the tissue source by using SNP markers is also not possible, as the genome of all cells in the body comprises an identical genetic sequence. The potential to address these issues has arisen after the discovery of a new layer of information that the DNA molecule carries, through the chemical modifications of its nucleic proteins and nucleotides, which can be used for tissue identification, and predicting an

individual's chronological age [5-8]. The chemical modifications of the DNA and its chromatin, which alter the phenotypic expression of the cell, are classified under the overall term of epigenetics. Therefore, in the last few years there has been a growing interest in the analysis of epigenetics in forensic science, in order to overcome the limitations of using conventional genetic markers for predicting human characteristics.

1.1 Epigenetics

The term epigenetics was first introduced in the early 1940s by the British biologist Conrad Waddington [9]. However, only in the past ten years have significantly rapid advances in the field of epigenetics been seen [10]. These rapid advances are attributed to the significant new developments in molecular biology technologies [11]. The definition of epigenetics was introduced by Waddington (1942) as "the branch of biology which studies the causal interactions between genes and their products which bring the phenotype into being during development" [10]. This definition has undergone subsequent modifications with the advancement of molecular biology research, which has changed our understanding of epigenetics [9]. More recently, epigenetics was defined by Riggs et al. (1996) as "the study of mitotically and/or meiotically heritable changes in gene function that cannot be explained by changes in DNA sequence" [12]. In other words, the epigenome is an additional layer of information that does not strictly depend on the DNA sequence [13]. This has clarified a phenomenon that has confounded scientists for years, endeavouring to explain how a complex organism comprising of different cells and tissues with the same genetic makeup can come from a single fertilized egg.

The molecular basis of epigenetics mainly involves controlling transcriptional activities in the cell through the activation and silencing of specific genes, which is reflected in the biological phenotype. The molecular mechanism

that controls gene expression uses various reversible modifications to the genome [9]. These epigenetic modifications include: DNA methylation (DNAm), regulatory RNAs, covalent modification of histone tails, and ATP-dependent remodelling of the nucleosome core (Figure 1.1). All these modifications play an essential role in regulating specific gene expression without altering the DNA sequence [14-16]. Therefore, this explains how individuals can possess a wide range of phenotypes but the same underlying genotype [17].

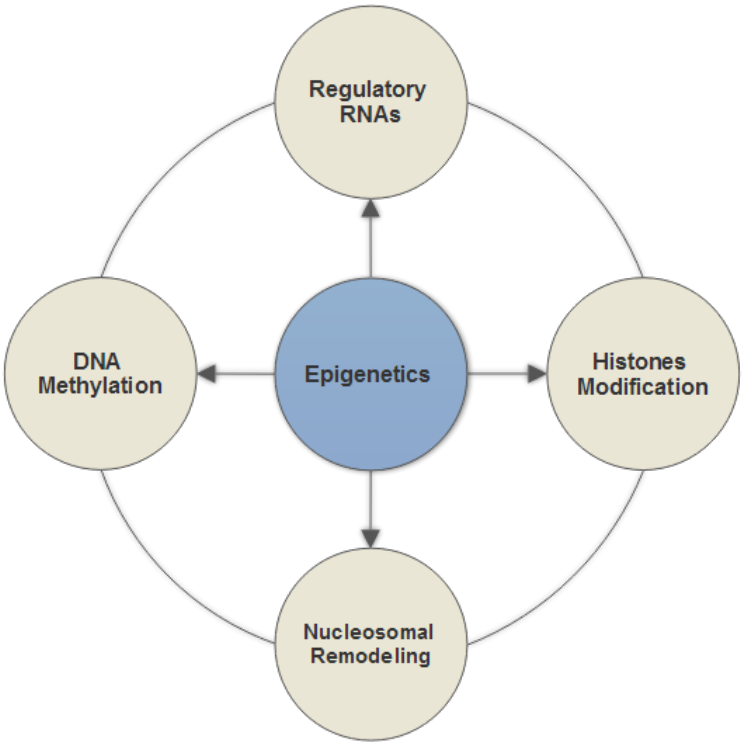

Figure 1.1 Epigenetic mechanisms used by the cell to regulate gene activity independently of the DNA sequence.

The epigenetic status of somatic cell lines is preserved during cell division, and thus any epigenetic defects are transmitted into newly divided somatic cells throughout the life of an organism [17]. In the germ line, some epigenetic biomarkers such as DNAm undergo epigenetic reprogramming by specific gene

regulatory networks in order to 'reset' the epigenome [18,19]. Researchers have speculated that this prevents inheritance of aberrant DNAm patterns that might adversely influence gene expression in the next generation. In contrast, chromatin modifications are considered to be epigenetically heritable biomarkers, as they can be transmitted from parents to offspring [20]. Epigenetic modifications will begin to take place during early foetal development in the uterus and then continue during an individual's lifetime as a response to environmental influences and various factors such as diet and smoking [15,17]. Despite an extensive body of research dedicated to understanding epigenetic modifications and their molecular mechanism, their regulatory role is still not fully understood [21,22]. Nonetheless, there is evidence of the vital role of epigenetics in genomic imprinting, gene silencing, cell programming during embryogenesis, X-chromosome inactivation and carcinogenesis [12,21,23]. In addition, it has been found that cell differentiation in many eukaryotes is mediated by DNAm [21]. Thus, recent stem cell research has been focusing on the effects of epigenetics during embryogenesis, which may help scientists to understand, prevent or treat developmental defects that can occur [24].

1.2 DNA methylation

1.2.1 The molecular basis of DNA methylation

DNA methylation is one of the most important epigenetic mediators playing a key role in the regulation of gene expression in the human genome. As the name suggests, DNAm involves methylation of the 5' position of cytosine residues, yielding 5-methylcytosine. The methyl group ($-CH_3$) is frequently, but not exclusively, added to the cytosine residues found in cytosine-guanine dinucleotide sequences. These repetitive linear CG dinucleotide sequences are called CpG sites, which are found along the genome in the 5' to 3' direction. There are ~ 28 million CpG sites in the human genome and approximately 75% of these CpG

sites are methylated. The methylation reaction is a reversible covalent modification initiated by the protein family of DNA methyltransferases (DNMTs), which use S-adenosyl-L-methionine (SAMe) as the source of a methyl group (-CH_3), yielding S-adenosyl homocysteine (SAH) as a waste product (Figure 1.2). The disruption of any component of the DNAm pathway has been found to be associated with most, if not all, types of cancers [25].

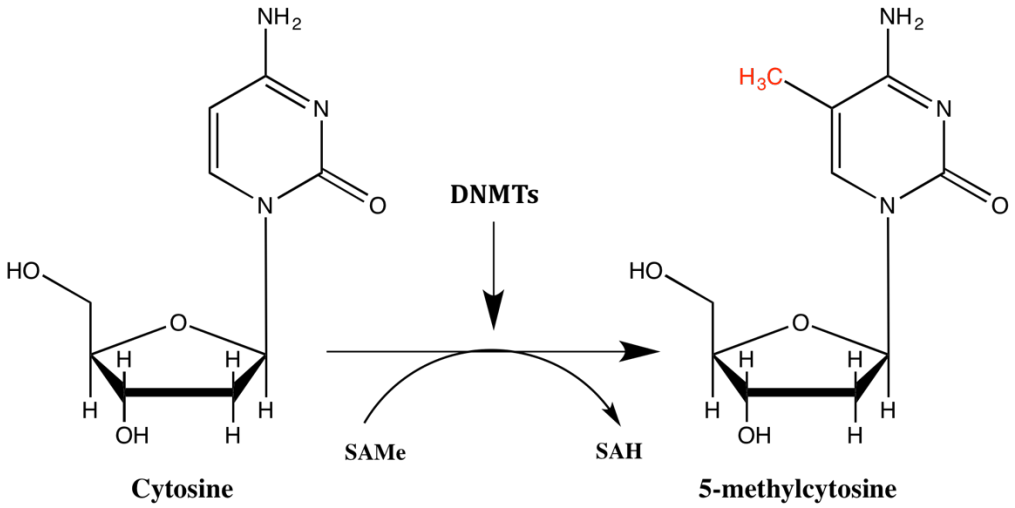

Figure 1.2 DNA methyltrasferases (DNMTs) enzymes attach a methyl group to the 5th atom of the 6-atom ring of cytosine residues using S-adenosyl-L-methionine (SAMe), rendering 5-methylcytosine and S-adenosyl homocysteine (SAH). Illustration generated with ChemDraw software v15.0.

In mammalian cells, there are three active DNMTs known as DNMT1, DNMT3A, and DNMT3B, and between them, these DNMTs have two major enzymatic activities in the cell. DNMT1 is responsible for maintaining the methylation pattern throughout the genome during cell division, and embryonic lethality has been seen in mice with mutated DNMT1 [26]. On the other hand, DNMT3A and DNMT3B are responsible for *de novo* DNAm, which plays a crucial role in cellular differentiation during embryonic development [16] (Figure 1.3). As such, DNMTs are found at high levels during embryogenesis in contrast to adult tissues [27]. In addition to their direct interaction with CpG sites, DNMTs can

combine with other gene expression repressors such as methyl CpG binding protein 2 (MeCP2) and histone deacetylase, which play a role in mediating gene silencing [28]. Studies have also shown a significant association between aberrant expression of DNMTs and tumour progression. For instance, DNMT1 was found in high levels in various types of cancers such as lung hepatocellular, acute and chronic myelogenous leukaemia, colorectal, gastric, and breast cancers [29-33]. Similarly, DNMT3A, and DNMT3B with mutated promoters, have been linked to several haematological malignancies, and high cancer risk, respectively [25].

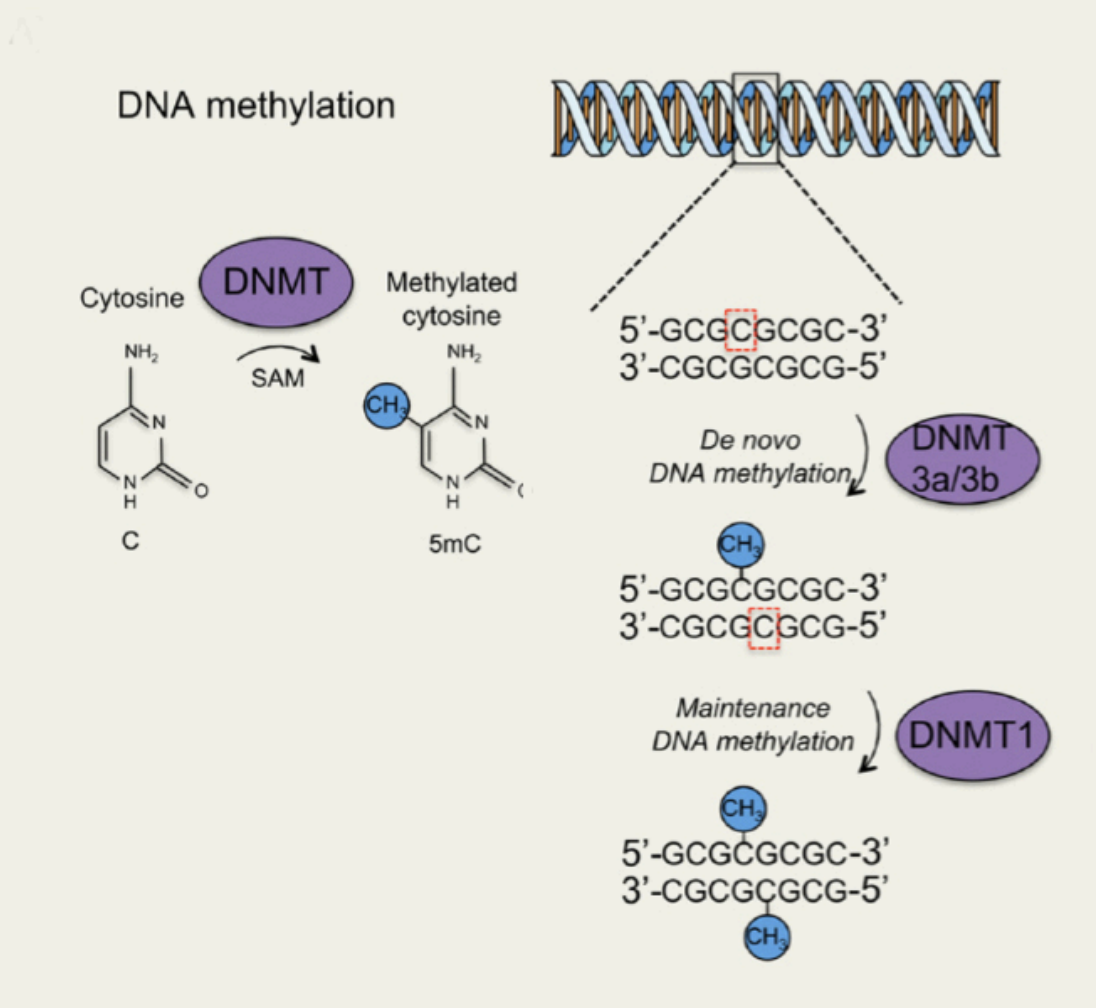

Figure 1.3 Schematic model depicting the role of mammalian DNA methyltransferases (DNMTs) in both DNAm and maintenance of already methylated CpG sites (Source: Heyward & Sweatt 2015).

1.2.2 DNA demethylation

As mentioned in the previous section (1.2.2), DNAm process is a reversible covalent reaction. Removing the methyl group (-CH_3) from 5-methylcytosine residues (or demethylation) along the genome, can be done through enzymatic reactions controlled by the biomolecular machinery of the cell, and/or stochastic loss during repeated rounds of cell divisions, due to the absence of methylation

of the newly synthesized DNA strands. The process of enzymatic demethylation is known as active demethylation, whereas demethylation caused by stochastic effects is called passive demethylation.

Various enzymes and pathways have been proposed to function in DNA demethylation, although their roles are often specific to the individual biological system being examined [34]. Oxidation-mediated demethylation pathways have recently been found to be most abundant in mammalian cells. The most prominent intermediate in this oxidation reaction is 5-hydroxymethylcytosine (5hmC), which is particularly abundant in cells from early embryos and the nervous system [35]. The protein family that is responsible for this oxidation reaction is the ten-eleven trans-location (TET) protein family.

Three TET proteins exist in vertebrates (TET1, TET2, TET3), with TET1 found in embryonic stem cells, TET2 in haematopoietic cells and TET3 in oocytes and zygotes [36,37]. Other oxidation intermediates have been found in the genomic DNA of various cells, namely 5-formylcytosine (5fC) and 5-carboxylcytosine (5caC) [38,39]. These intermediates are produced by further oxidation of 5hmC by TET proteins. Three mechanisms have been suggested for TET-mediated demethylation. The first scenario is that derivatives produced by TET oxidation will be lost during cell replications [40]. The second scenario is that the oxidation derivatives are removed by DNA glycosylases such as thymine DNA glycosylase (TDG), followed by base excision repair (BER). Finally, demethylation by BER could also be triggered by deaminases of the AID/APOBEC family [35]. Deamination of 5mC and 5hmC will create T:G and 5hmU:G mismatches respectively, which directly triggers TDG and MBD4 glycosylases followed by

BER [41]. Figure 1.4 summarises the chemical pathways of cytosine demethylation.

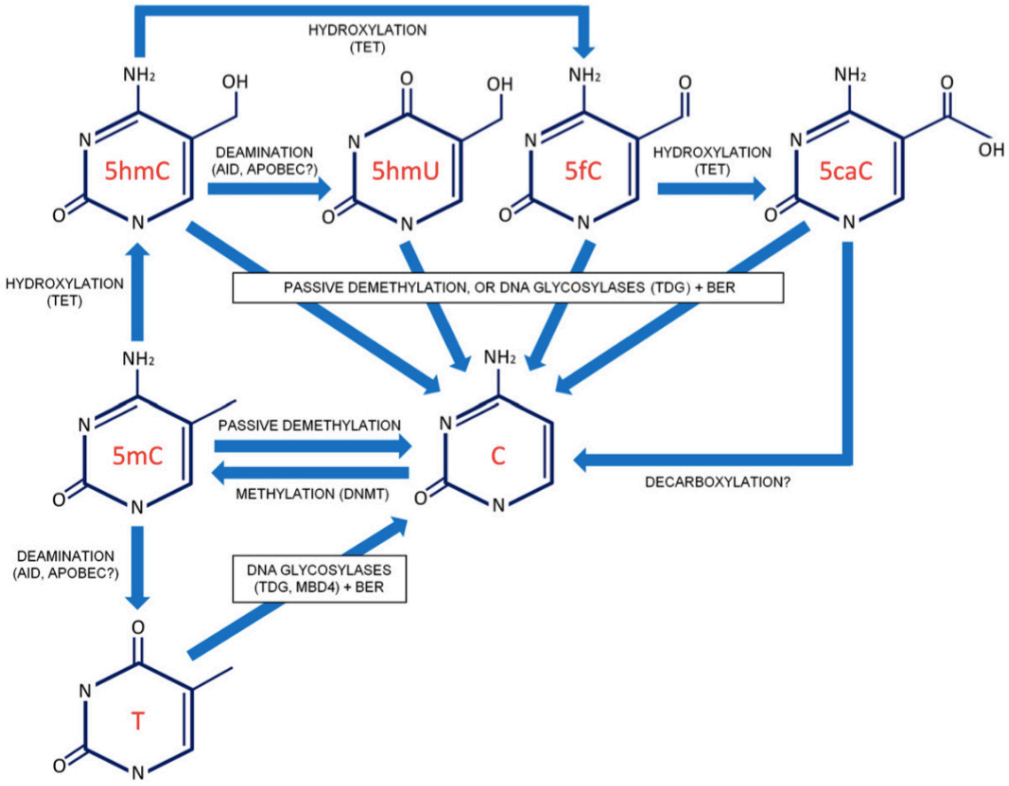

Figure 1.4 Chemical pathways of 5-methylcytosine demethylation. (Source: [35]).

1.2.3 Physiological effects of DNA methylation

There is evidence that methylation of DNA is a crucial step for normal development, and any impairment will lead to apoptosis or growth arrest in normal cell lines [9]. On a strand of DNA, CpG dinucleotides can be found in clusters, termed CpG islands, which are linked with the promoter and exonic regions of approximately 40% of mammalian genes [42]. In addition, CpG dinucleotides are disproportionately rare in the genome compared to the other dinucleotides except around the CpG islands that are linked with promoters' regions of the genes.

Therefore, they are considered landmarks in the genome for identifying the location of genes. The methylated CpG sites are recognised by specific binding proteins associated with histone deacetylase and chromatin remodelling complexes that stabilise the condensed structure of chromatin. Stabilising the condensed structure of the chromatin causes tight compaction, which makes the area less accessible by transcription factors. Therefore, the expression of genes surrounded by the condensed chromatin structure will be repressed [43,44]. This is why the structure of the chromatin around the gene promoter controls the transcriptional activity of the gene. Moreover, methylation of CpG sites at the recognition site of a number of transcription factors is sufficient to block binding of DNA transcription factors, thus inhibiting gene expression altogether (Figure 1.5) [45]. This was confirmed in a study conducted by Lokk et al. [46], who found an inverse correlation between DNAm and gene expression in 17 human somatic tissues.

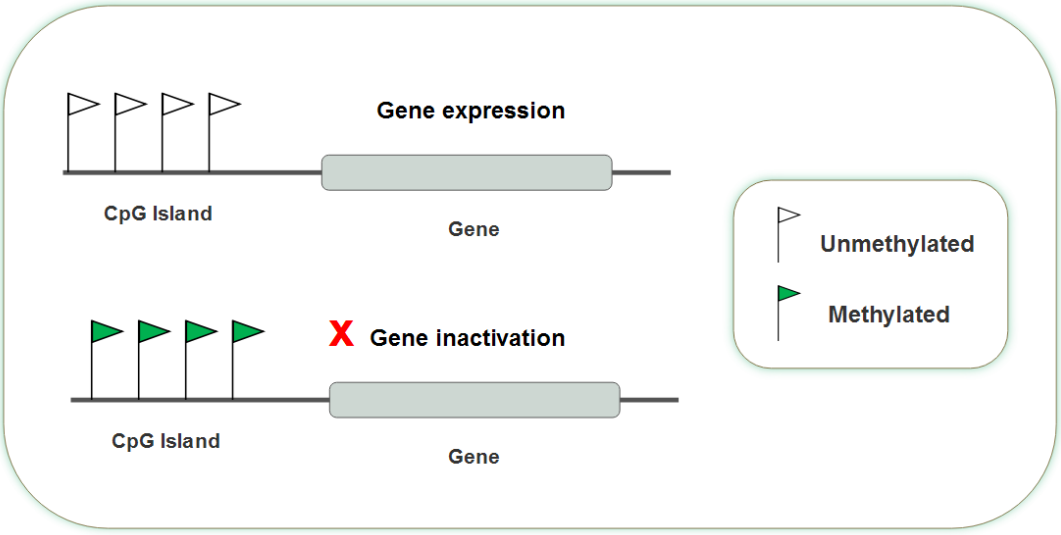

Figure 1.5 DNAm within CpG islands associated with gene promoter regions inhibits gene expression.

The mammalian cell exploits this repression effect of DNAm on gene expression for several genomic phenomena. One well-studied phenomenon is X-chromosome inactivation, which is used to compensate for the difference in gene dosage of X chromosomes between the sexes. A study conducted by Weber et al. [47] using chromosome-wide methylation analysis, discovered that CpG islands associated with promoters showed hypermethylation on the inactive X-chromosome when compared to the active X-chromosome. This significant difference in the methylation state of CpG islands between active and inactive X-chromosomes suggests that DNAm plays an important role in silencing of genes located on the inactive X-chromosome. The fact that cells rely on DNAm for X-chromosome inactivation further explains why the number of CpG islands linked to promoters of genes on the X-chromosome exceeds those linked to genes found on the autosomal chromosomes [48].

Another important genomic mechanism that is found to be associated with DNAm is genomic imprinting. This allows monoallelic expression of a subset of genes and is controlled by two epigenetic modifications, DNAm and histone modification [49]. It has been shown that the DNAm patterns that determine the imprinting state of genes is inherited from parental germ cells [50]. In addition, DNAm has also been discovered to be responsible for the differential expression of genes in tissues. Although early studies proposed no role of DNAm in regulating genes in a tissue-specific manner, this has been disputed by research conducted by Song et al. [51] who found tissue-specific differentially methylated regions (tDMRs) that regulate genes in specific tissue types. These tDMRs were either in a positively- or negatively-methylated form, and they can be found on CpG-rich, CpG-poor sequences, and within the 5 promoter regions of the tissue specific genes [52].

1.2.4 DNA methylation and the environment

Tremendous effort has been made by scientists to answer the question of whether any molecular mechanism exists that makes a fixed DNA sequence capable of adapting and communicating with the external environment. This endeavour has led to support for the hypothesis that the epigenome serves as an interface between the genome and the surrounding environment. External stimuli can be transferred from the dynamic environment to the genome by manipulating this second layer of information that is superimposed on the DNA sequence [53]. This manipulation is achieved using various epigenetic modifications such as DNAm, histone modifications, and chromatin remodelling. This in turn changes the epigenomic program that controls the gene expression profile in responsive cells or tissues [54]. However, in certain contexts, this alteration in genome function has been shown to be associated with a variety of physical and mental conditions (Figure 1.6).

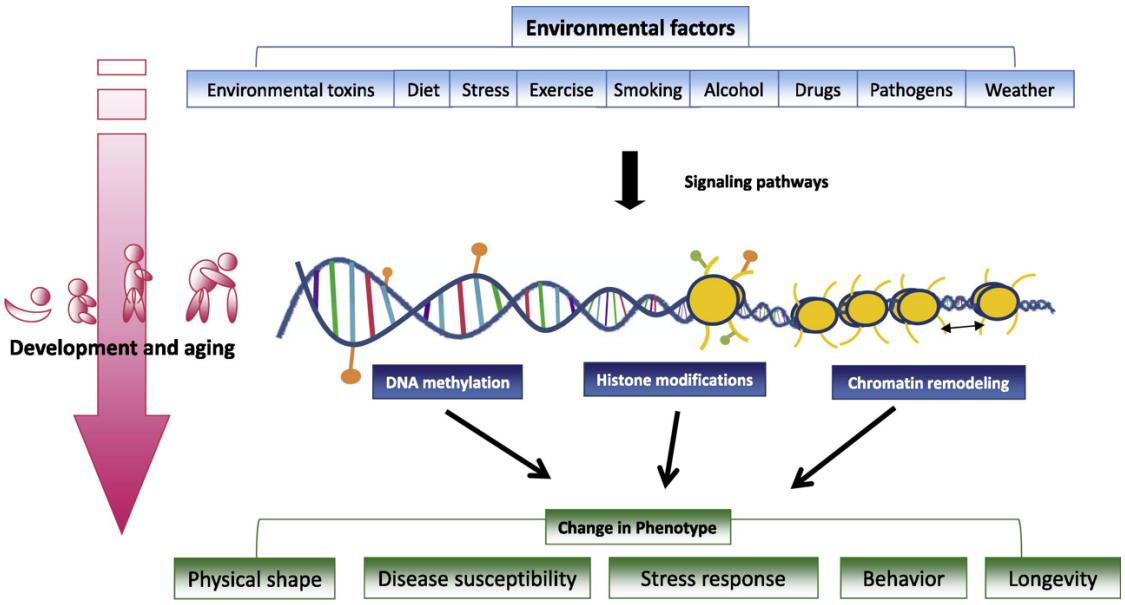

Figure 1.6 Epigenetic modification can act as an interface between environmental factors and the static DNA sequence, resulting in phenotypic changes (Source: Tammen et al. 2013).

The environmental factors in this context can be defined as factors that come from the external environment and are not initiated from properties intrinsic to the individual. These factors can be in the form of diet, smoking, stress, pathogen infection, alcohol consumption, exercise, etc. [55]. The best experimental model that can used to demonstrate the effect of these environmental factors on the epigenome is monozygotic twins (MZ). Since, MZ twins carry identical genomes, any alteration in phenotypic traits found between them could be attributed to non-DNA sequence-based factors, which are mainly environmental factors. This twin study design has allowed various researchers to identify epigenetic differences that were not only associated with phenotypic differences, but also with certain types of diseases [56]. For instance, Fraga et al. [57] showed that MZ twins exhibit distinctive epigenetic patterns, especially in those twins who were older, had different lifestyles and who had spent less of their lives together. This reflects how significant environmental stimuli can be in translating a common genotype into a different phenotype [57]. In addition, it provides an explanation as to why MZ twin pairs exhibit differing susceptibility to diseases [58].

In early life, DNAm is very susceptible to stimulation from the environment. The reason for this is that epigenetic patterns are generated at embryogenesis during cellular differentiation, which is a highly programmed and organized process. For example, it has been found that increasing supplementation of folic acid in maternal diet during gestation increases DNAm of critical genes, and that this persists into adulthood [59]. It is noteworthy that environmental factors affecting the epigenome are not limited to the input of chemical or biological materials. Studies have shown that even social factors can cause significant epigenetic reprogramming in the brain of offspring [60]. For instance, individuals who experienced early childhood adversity and committed suicide showed decreased hippocampal glucocorticoid receptor (GR) expression and increased hypothalamic–pituitary–adrenal gland (HPA) activity [60]. Therefore, any

adversity in early life and during childhood could have a lasting impact on the epigenome.

Other well studied external factors that have a powerful impact on the epigenome are cigarette smoking and alcohol consumption. Epigenome-wide studies (EWAS) have shown that smokers, as opposed to non-smokers, have two genes, namely *F2RL3* (encodes a protease-activated receptor), and *AHRR* (encodes aryl hydrocarbon receptor repressor) that are significantly hypomethylated in three types of tissues; lungs, peripheral lymphocytes, and in whole blood [61-63]. Among these tissues, lungs are the most epigenetically affected by smoking, exhibiting a 34% greater hypomethylation effect compared to the other tissues [62,63]. This smoking related-hypomethylation has been found to alleviate the toxic effect of polycyclic aromatic hydrocarbons found in cigarette smoke [64]. Decreasing DNAm of the *AHRR* gene leads to increased expression of the protein it encodes, which mediates the detoxification of environmental pollutants such as those found in cigarettes [65]. Furthermore, maternal smoking during pregnancy has also been seen to cause methylation changes in the umbilical cord blood and placental DNA, at the *AHRR* and *CYP1A1* (encodes Cytochrome P450 Family 1 Subfamily A Member 1) genes, which are risk factors for diseases in adulthood [66,67].

The effect of sustained heavy use of ethanol on the human body has been well studied and found cause a high risk of various types of diseases, including cardiovascular diseases, hypertension, cancer, and liver diseases [68]. Alcohol consumption can cause these types of diseases either through disrupting a number of biochemical pathways, and/or by disrupting normal DNAm patterns in cells and tissues, which in turn disrupts normal gene expression that the tissues require in order to function properly [69]. Various EWAS studies have shown that the blood of alcohol dependent (AD) individuals exhibit global hypermethylation, and differential methylation levels at *dehydrogenase 1A*, *ADH7*, *ADH3B2*,

CYP2A13, plus five additional loci (*C8orf4*, *HCRTR1, FLJ38379, HSA277841, and TSC2*) [70-72]. In addition, a study conducted by Zhao et al. [73] using EWAS for AD-discordant siblings reported 865 hypomethylated and 716 hypermethylated loci in the AD sibling, with the most significant hypomethylation and hypermethylation being found at the *SSTR4*, and *GABRP* genes respectively [73]. This epigenetic effect of ethanol intake has been clinically exploited as a useful epigenetic biomarker for the diagnosis and treatment of alcohol-related diseases [74].

1.2.5 The inheritance of DNA methylation

Acquiring epigenetic patterns from the previous generation can be either through transgenerational or intergenerational transmission. The former occurs when the offspring inherits epigenetic changes from the parents, whereas intergenerational transmission results when the stimulation for the epigenetic alteration occurs during pregnancy and both the mother and child are simultaneously exposed to the external stimuli [75]. Epigenetic reprogramming is therefore a key mechanism in preventing the inheritance of aberrant DNAm patterns that might adversely influence gene expression in the offspring [18,19].

Two major stages in epigenetic DNAm reprogramming occur during the mammalian life cycle. The first occurs during gametogenesis, and the second directly after fertilization (during embryogenesis) [75]. The reprogramming process involves genome-wide erasure (demethylation) of the DNAm pattern, followed by selective methylation of a novel set of CpG markers that are required for both cellular activity and cellular differentiation [76]. However, despite this epigenetic resetting system, it has been shown that DNAm in some genomic regions escapes this process and persists, getting transferred into the next generation.

Escaping the genome-wide demethylation and/or selective methylation process during reprogramming can be controlled by the cell, or may occur at random [20]. For instance, genomic regions that are protected from demethylation and left hypermethylated by the cell can be found in the subtelomeric regions, pericentromeric satellite repeats and in single copy loci [77,78]. Moreover, regions associated with retrotransposable elements, and imprinting-associated DMRs with sex-specific methylation patterns are also protected from the demethylation machinery [77,79]. To date, it has not been fully elucidated how these genomic regions can escape the epigenetic reprogramming steps. If CpG sites with environmentally-acquired DNAm patterns (from smoking and/or alcohol etc.) withstand these waves of demethylation/methylation steps during gametogenesis and embryogenesis, they will find their way into the subsequent generation [75].

1.2.6 Detection and quantification of DNA methylation

Determining the relationship between DNAm levels at CpG sites across the genome and different types of covariates was not possible without developing methods that can quantitatively measure DNAm levels. There are now a number of methods that can be used to detect and/or quantify DNAm levels. Selecting the appropriate method among these will depend on various factors such as the research goals, the nature of the samples under study, the scale of the assessment (genome-wide or gene-specific), and more.

Although methylated cytosine is chemically different from unmethylated cytosine, the sequencing and fragment analysis methods above cannot differentiate between them. To overcome this issue, scientists developed three pre-treatments that can be implemented before analysis in order to make methylated cytosine detectable by the methods most frequently found in the molecular genetic laboratories [80,81]. These pre-treatments are:

1. Restriction enzyme digestion.

2. Affinity enrichment by monoclonal antibodies.
3. Sodium bisulfite conversion.

The first method exploits the fact that some restriction enzymes are sensitive to the presence of a methyl group on cytosine residues and thus different digestion patterns would result from methylated and unmethylated loci. This method has been used to detect tDMRs and aberrantly methylated genomic regions in diseases such as cancer [51,82]. In the affinity enrichment method, methylation is detected by monoclonal antibodies and then quantified by array-based hybridisation techniques, for example using the Methylated Immunoprecipitation (MeDIP) chip or by sequencing via the MeDIP-seq method [81].

A key advancement in DNAm analysis has been the development of a chemical treatment called sodium bisulfite modification, which converts unmethylated cytosines to uracil, leaving methylated cytosines unchanged [83]. The majority of the methods used for analysing DNAm levels are based on this approach [84]. The oxidation of cytosine by sodium bisulfite can either be initiated with nucleosides (ribo- or deoxyribo-) or oligonucleotides as a substrate, but the reaction is highly specific to single-stranded DNA. As such, it can be used in single-stranded DNA studies due to its ability to differentiate between single- and double-stranded DNA regions. As shown in Figure 1.7, the first step in the bisulfite modification reaction is converting cytosine residues into a cytosine-sulphonate derivative. This is done by adding a sulphonate group, which subsequently leads to an irreversible spontaneous deamination reaction resulting in a uracil-sulphonate derivative. The sulphonate group is then removed by alkali treatment with sodium hydroxide, yielding uracil (Figure 1.7).

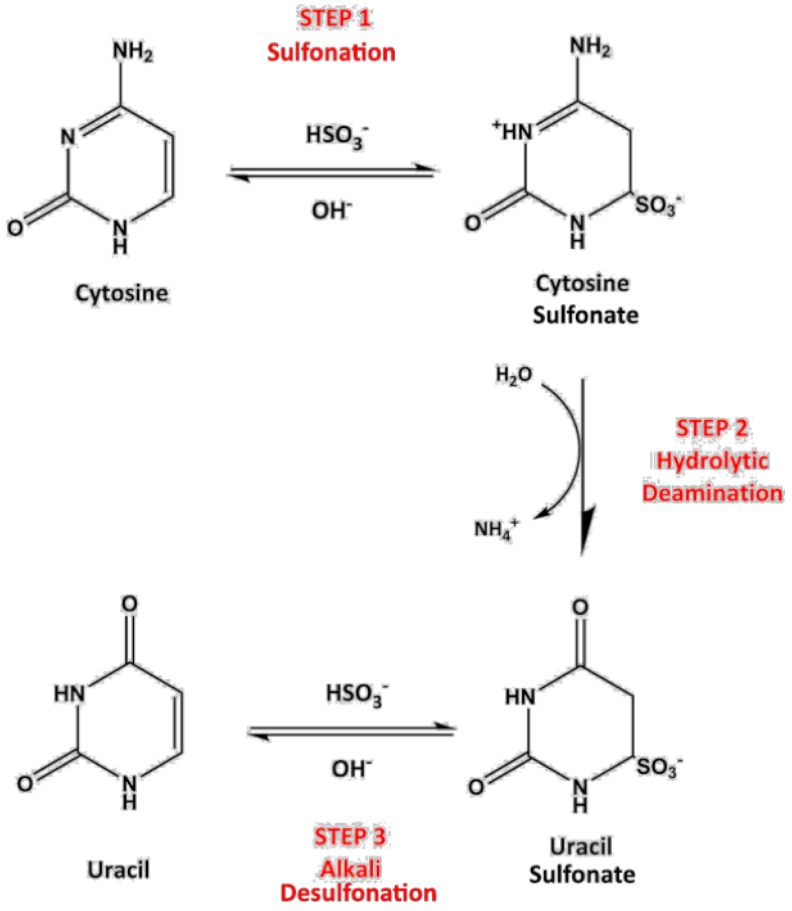

Figure 1.7 Sodium Bisulfite Treatment. (Diagram was produced using ChemDraw Software).

The formation of cytosine sulphonate and uracil sulfonate derivatives are reversible reactions controlled by three factors; pH, temperature, and bisulfite concentration. In the first step, the sulfonation reaction equilibrium is favoured towards cytosine-sulfonate by low pH (below 7) and reversal towards cytosine by high pH. A high concentration of sodium bisulfite is required in the second step during the deamination reaction, which is also initiated by high temperature and low pH, rendering uracil sulfonate. Finally, the bisulfite adduct is removed from uracil-sulfonate by sodium hydroxide (high pH), giving uracil.

At this stage, the methylation status of any genomic DNA sequence can be

detected using the polymerase chain reaction (PCR). During PCR, *Taq* polymerase amplifies unmethylated cytosine as thymine, whereas 5-methyl-cytosine is amplified as cytosine. Therefore, bisulfite treatment converts the chemical modification (DNAm) into DNA sequences that can be detected and quantified using any quantitative genotyping methods (Figure 1.8) [85].

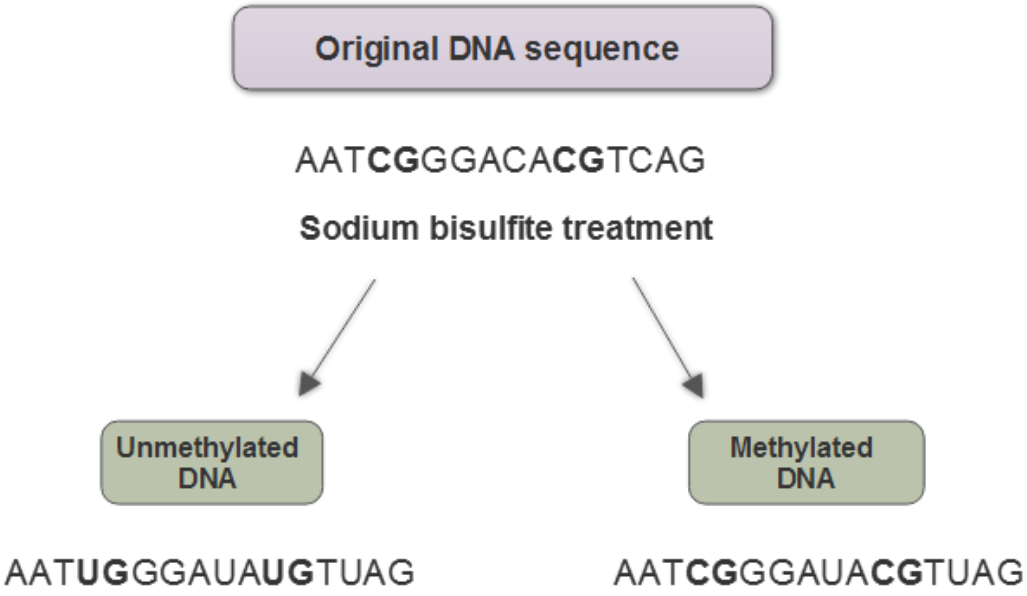

Figure 1.8 The effect of sodium bisulfite treatment on non-methylated and methylated DNA sequences. Only non-methylated cytosines are converted into uracil, while the methylated residues remain unchanged.

Generally, the goal of most of DNAm studies is either to discover new CpG sites or to quantify the DNAm level at locus-specific CpG site(s). There are various DNAm methods that are based on sodium bisulfite treatment that can be used for these purposes. Detection of new methylation markers can be done using epigenome-wide approaches, which involve scanning the whole genome for CpG sites, and can also be directed to specific regions in the genome such as specific genes and/or CpG islands [86]. However, in order to quantify DNAm, the CpG markers under study must already be known, with PCR primers or probes

available to query the methylation level at the CpG sites under study. There are numerous bisulfite-based quantification assays, but only the methods used for forensic applications will be described here.

1.2.6.1 EpiTect® MethyLight PCR

EpiTect® MethyLight is a PCR-based method that can differentiate between methylated and unmethylated CpG sites. This is achieved by designing primers that specifically bind and amplify CpG sites in bisulfite-modified DNA. As sodium bisulfite treatment renders differences in DNA sequence between methylated and unmethylated cytosine, primers can be specifically designed to distinguish and recognise these alterations in nucleotide sequences. That is, the primers are designed to anneal and amplify only methylated loci, and the methylation level of CpG sites under study is quantified by measuring the amplified products using real-time PCR. The reliability of the results obtained with this assay type is thus dependent on primer design, so newly designed primers should be fully validated before they are used for analysis of DNAm patterns [87].

1.2.6.2 Allele-specific bisulfite sequencing

Using this method, methylated and unmethylated alleles are amplified post-sodium bisulfite treatment. The methylated and unmethylated CpG sites are co-amplified by specifically designed primers. Following this, the amplified products are introduced into cloning vectors and transfected into competent cells. After cloning, plasmid DNA is then extracted and sequenced using the dideoxynucleotide chain-termination method [45]. The sequence data obtained represents the methylation status of a single allele, which is particularly useful in studying genomic imprinting. Despite the fact that this approach is widely used for the characterisation of allele-specific methylation, it is expensive and labour intensive, especially when studying a large number of loci.

1.2.6.3 EpiTYPER assay

EpiTYPER is a high-resolution DNAm profiling technique, which is performed on Agena Bioscience's MassARRAY System. This system utilizes Matrix-assisted laser desorption/ionization time-of-flight (MALDI-TOF) mass spectrometry for quantifying the methylation level at CpG sites under study. The first step in using EpiTYPER is designing primers to amplify the CpG sites in a selected genomic region. This is can be done by using an automated online software called EpiDesigner, which delivers an easy-to-read graphical interpretation of the amplicons designed over the target regions under study, as well as annotating distinct CpG sites covered by the assay [88]. After PCR amplification of the bisulfite treated DNA, any unincorporated dNTPs are removed by treating the reaction mix with shrimp alkaline phosphatase (SAP). Then, an *in vitro* RNA transcription reaction is initiated using a T7-promoter tag, which is found in the reverse primers designed by EpiDesigner. The RNA transcription step is essential in order to preserve the bisulfite-induced sequence changes. The RNA transcripts then go through a uracil-specific cleavage using RNase A [89]. The rationale behind this step is that RNase A will produce two fragments that are identical in length, however the methylated fragment will be heavier in mass than the unmethylated fragment by 16 Da, with the presence of each additional methyl group (Figure 1.9). The resulting fragments can be differentiated using MALDI-TOF mass spectrometry based upon their mass. It has been shown that the results produced by the method are highly reproducible, when compared to other methods [90].

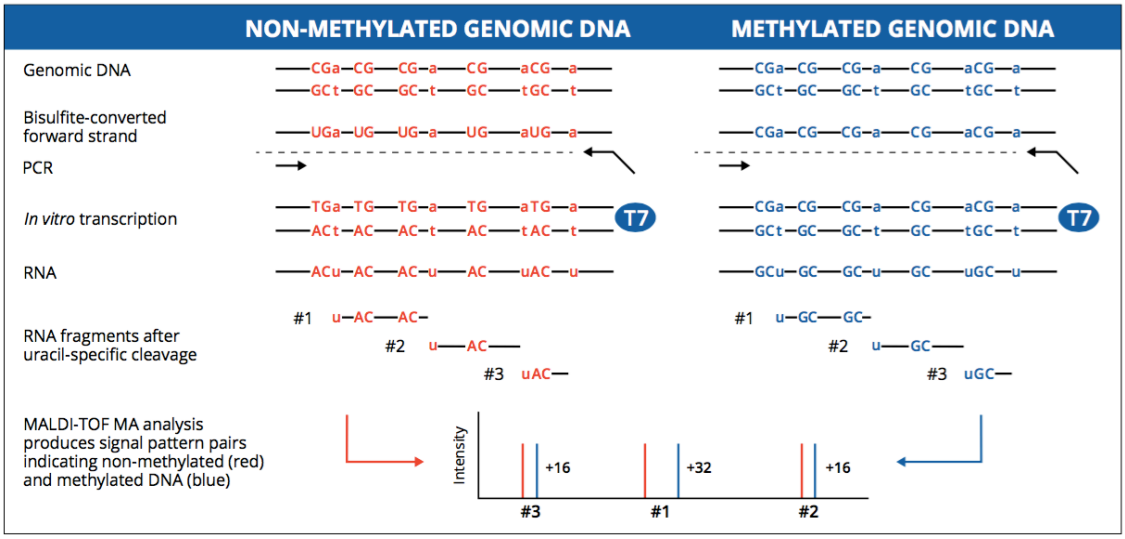

Figure 1.9 **Overview of EpiTYPER Assay.** CpG sites are amplified using primers tagged with a T7 promoter sequence for RNA transcription. The RNA transcripts are cleaved and then analysed by MALDI-TOF MS. Shifts between strands by 16 Da or 32 Da mass will indicate methylation at one or two CpG sites respectively. The methylation level can be estimated by calculating peak area ratio of corresponding mass signals. (Source: Ehrlich et al. 2005)

1.2.6.4 SNaPshot assay

This assay is based on a single-base extension (SBE) reaction, which quantitatively detects methylated and unmethylated cytosine at specific CpG sites using bisulphite-treated and untreated samples [91]. Similar to other methods, SNaPshot starts with bisulfite DNA conversion and PCR clean up. Following this, during the SBE reaction step, primers are designed so that they anneal to sequences upstream of the CpG site being queried, terminating immediately at the 5' end. In the presence of dye-labelled dideoxynucleosides (ddNTPs), DNA polymerase will extend the growing chain by a single nucleotide, and then the reaction will be terminated. The termination process is due to the lack of a 3' OH group on the dideoxynucleoside, which is required for further chain elongation. Each of the four ddNTPs (ddATP, ddTTP, ddCTP and ddGTP) is labelled with a differently coloured fluorescent dye tag, allowing identification of the dideoxynucleoside incorporated. In reality, for SNaPshot typing of CpG sites only

two of these bases are used, as the base present at the position in question is either a C (methylated) or a T (unmethylated), so either a ddGTP or a ddATP is incorporated [92]. The reaction amplicons are then analysed by capillary electrophoresis for visualization, and the dye colour used to determine the identity of the base at the CpG site [93]. Furthermore, SnaPshot can be used to quantify the DNAm level by dividing the peak height of the C/G nucleotide (from a sodium bisulfite untreated sample) by the peak heights of the C/G nucleotide plus the T/A nucleotide (from a sodium bisulfite treated DNA sample) (Figure 1.10) [92].

SNaPshot has been introduced to assay methylation of CpG sites due to its sensitivity in separating CpG loci that differ by a single base pair. The sensitivity and robustness of the assay is improved by increasing the amount of the starting bisulphite-converted genomic DNA using an amplification step before the SBE reaction, followed by a step to clean-up excess primers and dNTPs using exonuclease and shrimp alkaline phosphatase. In a single reaction, SNaPshot can interrogate up to 10 CpG sites from different amplicons. This multiplexing capability is particularly important in increasing throughput of sample processing and data analysis, while reducing sample consumption [94].

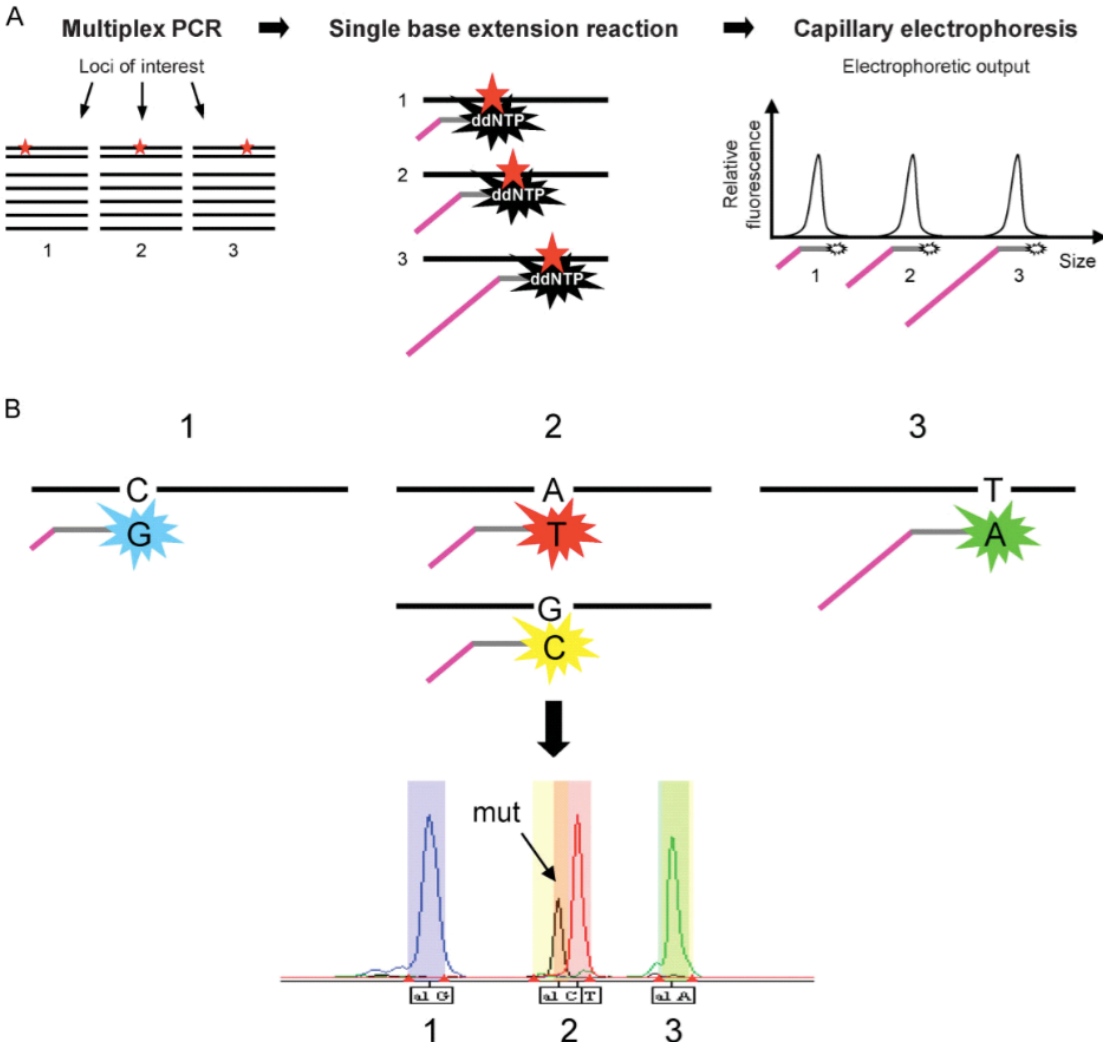

Figure 1.10 **Overview of SnaPshot assay**. (A) After sodium bisulfite modification, PCR amplification takes place using primers flanking the regions containing CpG markers. Then, the reaction mixture is cleaned to remove unconsumed dNTPs and primers prior to sequencing the methylated/unmethylated cytosine at CpG sites using a single-base extension reaction (SBE). (B) Finally, the sequencing results are analyzed using capillary electrophoresis. The presence of methylated cytosine will appear in the electropherogram in a distinctive color and the level of methylation will be represented by the peak height ratio. Source: (Thermo Fisher Scientific 2014).

1.2.6.5 Illumina Infinium assay

The Illumina Infinium assay is a high-throughput profiling technology that has been implemented on various genomic platforms, with applications ranging

from copy number variation (CNV) detection, SNP genotyping, RNA analysis, to quantifying DNAm levels [95]. The Infinium assay is based on microarrays known as BeadChips which have much higher densities of oligonucleotide probes than traditional spotted microarrays [96]. As the name suggests, BeadChips consist of microscopic silica beads assembled on micrometre-scale wells that are randomly distributed on the chip. Each of these nano-silica beads is covalently coated with hundreds to millions of copies of oligonucleotides that target specific CpG sites in the genome [97]. These oligonucleotides span 50 bases and consist of two parts; the 'address' that uniquely identifies it and the 'probe' which targets the genomic query site (Figure 1.11).

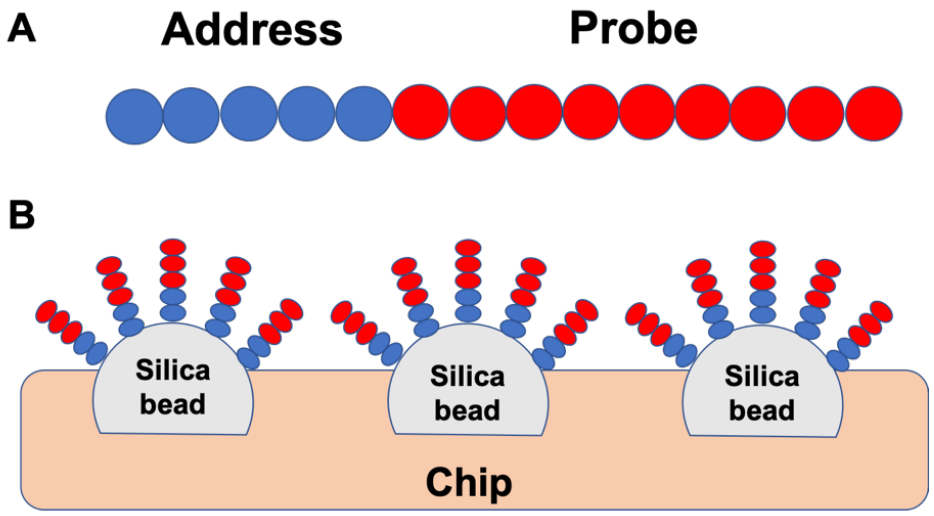

Figure 1.11 Illustrates **A** how the oligonucleotide is divided into two parts, address and probe, and **B** how oligonucleotides are attached to the silica beads that are attached to the microarray chip.

Due to the fact that the beads with the oligonucleotides are randomly distributed on the chip, determining their locations and the types of probes they are carrying is important in order to correctly conduct the analysis. This is done by a decoding step which consists of a series of hybridisation reactions with decoder oligonucleotides, complementary to the address sequences, that are

labelled with a fluorophore and captured by the CCD camera. The image captured at the end of the decoding procedure will result in a map that identifies the beads and the type of probes they carry at each well on the microarray chip. Thus, each microarray chip has a specific decoder map, and this is provided by Illumina, Inc.

1.2.6.6 Infinium HumanMethylation BeadChips

With high demand from researchers exploring the epigenome-wide DNAm profiles (so-called methylome) of cells and tissues, Illumina, Inc. developed microarray chips based on the Infinium BeadChip technology called Infinium HumanMethylation BeadChip. To date, there are three different versions, all of them measuring DNAm using target-specific probes designed to interrogate individual CpG sites of bisulfite-converted genomic DNA with single base resolution. Thus, the Infinium HumanMethylation platforms allow direct access to the largest number of genomic DNA target sites without requiring any site-specific PCR or methylation sensitive restriction enzyme treatment to capture the methylated DNA [98].

The first platform developed by Illumina, Inc. in 2008 was the Infinium HumanMethylation27 (HM27K), which quantifies methylation levels at 27,578 CpG sites in the human genome that are located within the proximal promoter regions (1 kb upstream and 500 bp downstream) of 14,495 genes including known cancer-related genes [86,161]. Since its development, HM27K has played a pivotal role in epigenome-wide association studies and has resulted in major findings about the relationship between DNAm levels at different CpG sites and various covariates such as age, tissue type, diseases, environmental factors and more. Therefore, after the successful release of the HM27K, Illumina, Inc. developed the platform further by adding more probes, targeting more than 450,000 CpG sites covering 99% of RefSeq genes and 96% of CpG islands, with additional coverage in regions a short distance from CpG islands known as island shores and the regions flanking them (176,112 CpG sites) [98]. This updated

platform is called Infinium HumanMethylation450K BeadChip (HM450K), which was released in 2011. Both Infinium platforms (HM27K and HM450K) can be used to analyse 12 samples simultaneously on a chip (arranged in six rows and two columns) and the methylation profiles generated are highly reproducible with other bisulfite-base sequencing technologies (with an average R^2 of 0.95) [98].

As described in the previous sections, there are ~28 million CpG sites in the human genome, and thus it is essential for our understanding about human development and etiology to study DNAm levels at these genomic sites. To meet this need, Illumina, Inc. developed a new BeadChip platform called MethylationEPIC (EPIC), which has 865,918 target-specific probes. In addition to targeting 90% of the CpG sites on the HM450K, the additional 350,000 probes target CpG sites in regulatory enhancers identified by the FANTOM5 and ENCODE projects [99,100]. This addition of new probes, especially those designed to target the enhancer regions, will provide a new opportunity for researchers to further understand the role of DNAm in human development and diseases [101]. Table 1.1 below summarises the differences between the three Infinium HumanMethylation BeadChip platforms.

Table 1.1 Comparison between the three Infinium BeadChip HumanMethylation platforms.

Feature	HM27K	HM450K	EPIC
Probes	27,578	485,577	865,918
Coverage	14,495 genes	- ~ 90% HM27K probes - FANTOM 4 promoters - RefSeq - miRNA promoters - non CpG island sites - DNase hypersensitive sites	- > 90% HM450K probes - FANTOM 5 promoters - RefSeq - ENCODE - miRNA promoters - non CpG island sites - DNase hypersensitive sites
Infinium assay	Infinium I	Infinium I, II	Infinium I, II
Input DNA	1μg	500ng	250ng

As they are user-friendly, have a streamlined workflow, and they need low amounts of DNA input, the Infinium HumanMethylation platforms became a key tool for many epigenome-wide DNAm studies. In addition, different consortiums such as the International Cancer Genome Consortium (ICGC) and the International Human Epigenome Consortium (IHEC) rely on the Infinium BeadChips to generate epigenome profiles from their reference samples [101]. Furthermore, there are more than 7500 epigenome profiles from over 200 different cancer types found in The Cancer Genome Atlas (TCGA) online repository, and more than 1000 epigenome profiles for mother-offspring pairs in the online genomic repository of one of the largest epidemiological studies, ARIES (Accessible Resources for Integrated Epigenomic Studies), which have been assayed on the HM450K [102,103].

The first step in the procedures for Infinium HumanMethylation BeadChip analysis is genomic DNA extraction, bisulfite conversion, followed by whole genome amplification and fragmentation. The fragmented DNA is then denatured to form single stranded DNA and suspended on the microarray chip for hybridisation to the probes. A washing step is then carried out to wash away any unhybridised and non-specifically hybridised oligonucleotides. Then, in the presence of biotin-labelled ddCTP and ddGTP, and 2,4-dinitrophenol (DNP)-labelled ddATP and ddUTP, the hybridised oligonucleotides undergo single-base extension by a polymerase enzyme. After the SBE step, the array is fluorescently stained and the intensities of the methylated and unmethylated bases are captured and stored in a high-resolution image and in intensity data files (idat). The workflow is summarised in Figure 1.12.

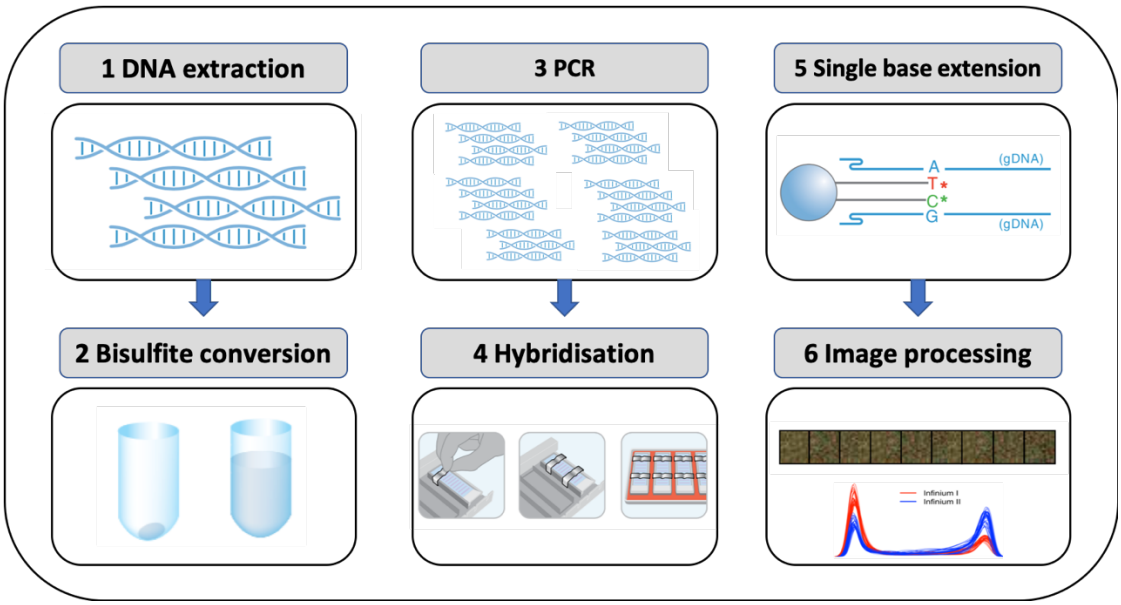

Figure 1.12 Diagram illustrating the general workflow of Illumina Infinium® BeadChips assay.

1.2.6.6.1 Probe types

There are two types of oligonucleotide probes on the Infinium HumanMethylation BeadChip platforms, based on two different chemistries, termed Infinium I and Infinium II.

- **Infinium I**

In Infinium I, there are two probes to interrogate each CpG locus, one to query the unmethylated form of the CpG site and the other to query the methylated form of the CpG site (Figure 1.13A). The 3` terminus of each probe is designed to match either the protected cytosine (methylated) or the thymine (unmethylated) base resulting from bisulfite conversion. In case of hybridisation with either the methylated or unmethylated probe, single base extension will incorporate one of four fluorescently labelled ddNTPs. Therefore, the signal intensities of the labelled ddNTPs incorporated into the methylated and unmethylated probes will reflect the methylation level at the specific CpG site. The Infinium HM27K chip contains only Infinium I probes.

- **Infinium II**

In contrast, Infinium II chemistry employs only one probe per locus that queries both methylated and unmethylated forms of the CpG locus (Figure 1.13B). The 3' terminus of the probe complements the base directly upstream of the query site. Analogous to the principle of the SNaPshot assay described in Section 1.2.6.4 above, single base extension results in the addition of a labelled G (green) or A (red) base, complementary to either the methylated C or unmethylated T [98]. Therefore, there will be two distinctive colour readouts, one colour for each allele; green for methylated sites and red for unmethylated sites [104].

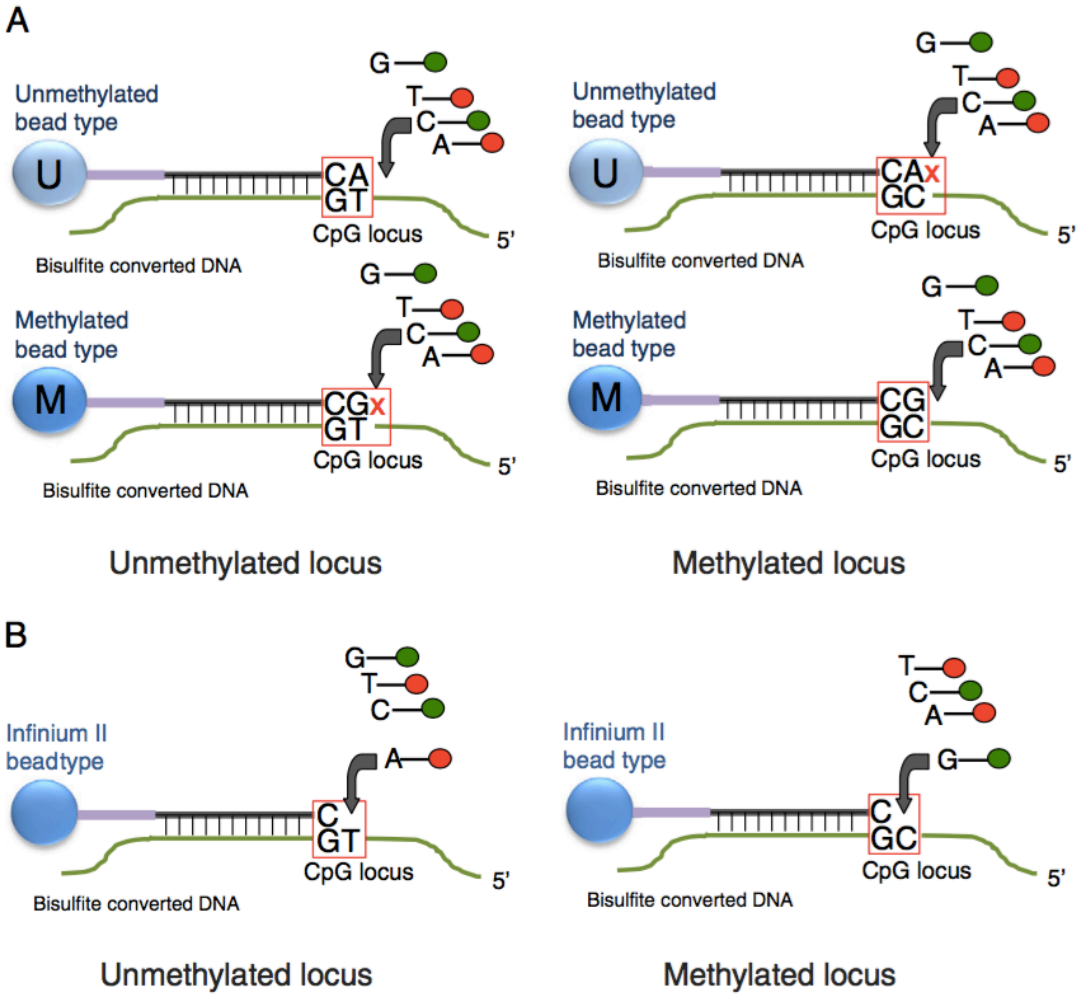

Figure 1.13 Infinium I and II Methylation Assays, applied to both methylated and unmethylated CpG loci. **A.** Infinium I chemistry uses two probes to interrogate each locus. **B.** Infinium II chemistry uses one probe type to interrogate each locus. (Source: Bibikova et al. 2011)

As the bisulfite-converted DNA fragments pass over the bead chip probes, each probe will bind with complementary sequences in the sample DNA, binding one base upstream of the locus of interest. The natural competition among the four bases in the single base extension reaction minimizes bias, allowing DNA polymerase to extend the probe with the correct base that matches the target DNA. At each locus, the intensity of the two possible fluorescent signals emitted

from the incorporated ddNTP are captured by a CCD camera; ddGTP for the C (methylated) allele, or ddATP for the T (unmethylated) allele.

1.2.6.6.2 Measuring DNAm level

The DNAm level (measured via the Beta value) is calculated by dividing the intensity of the methylated allele (*M*) over the total intensities of both methylated (*M*) and unmethylated (*U*) alleles, plus a fudge factor (ϵ), which ensures a positive denominator. Illumina, Inc. recommends ϵ =100, which is two orders of magnitude smaller than the observed intensities.

$$Beta\ (\beta) = \frac{M}{U + M + \ \varepsilon'} \tag{1.1}$$

The Beta value has an intuitive biological interpretation, that is, 0 equates to an unmethylated CpG site and 1 a completely methylated CpG site [104]. However, the main limitation of the Beta value is that it exhibits severe heteroscedasticity outside the middle methylation range, which imposes serious challenges in statistical models such as regression models [105]. Moreover, because it is not normally distributed, it violates the normality assumption used by many statistical methods, including t-test and regression analyses [105]. To overcome these limitations, Du et al. (2010) proposed an element-wise transformation of the Beta value to give an M value:

$$M = logit2\ \frac{M}{U + M + \ \varepsilon'} = \ \log 2 \frac{M}{U + \ \varepsilon'} \tag{1.2}$$

The reason why this is referred to as an M value is because it has been widely used in mRNA expression microarray analysis. It has been demonstrated by Du et al. (2010) that the M value is statistically valid and has better detection power and higher true positive rate (TPR) in the low and high methylation ranges due to its homoscedasticity [105]. However, one of the disadvantages of the M

value is that it does not have an intuitive biological interpretation. As such, in epigenome-wide DNAm analysis it is preferred to use M values for differential methylation analysis and then include the Beta value in final reports so the reader can biologically interpret the results [105,106].

1.2.6.6.3 Normalisation

Due to the presence of two probe designs (Infinium I and II) corresponding to two different chemical assays, the signal intensities between them are non-comparable and exhibit widely different distributions [107]. The DNAm levels of Infinium II probes are less reproducible and if the same methylation level measured using both probe types, Infinium II probe would give a lower value [107]. Therefore, it is necessary to apply normalisation to methylation data generated using Infinium II probes to render them comparable with Infinium I probes and reduce their technical bias, before conducting any downstream statistical analyses [108]. There is more than one normalisation method available to perform such a correction: unclosing a peak-based correction [107], subset-quantile within array normalization (SWAN) [109], subset quantile normalization [85], and beta-mixture quantile dilation normalisation (BMIQ) [108]. However, it has been demonstrated by Marabita et al. [110] that BMIQ is that the best algorithm for reducing probe design bias compared to other normalisation methods.

The BMIQ method fits a three-state (unmethylated-U, hemimethylated-H, fully methylated-M) beta mixture model to Infinium I and II probes separately. Then it uses state-membership probabilities to reassign the quantiles of unmethylated-U and fully methylated-M values for the Infinium II probes in order to determine the normalised values according to the Infinium I distribution. However, the state-membership probabilities of the hemimethylated-H values of Infinium II cannot be reassigned to the hemimethylated values of Infinium I, as they are not well described by a beta-distribution [107]. This issue can be solved

by using methylation-dependent dilation transformation, which preserves the monotonicity and continuity of the data from the endpoints that are defined by the maximum and minimum Beta values of Infinium II probes [108].

1.2.6.7 Targeted bisulfite sequencing

Assessing the methylation level of targeted genomic regions (i.e. a promoter region of a single gene or a CpG island) can also be done using different next-generation sequencing platforms, such as the Illumina MiSeq® system [111]. Detecting methylated cytosine residues in sequenced regions is facilitated by bisulfite treatment, which converts any unmethylated cytosine into uracil, while methylated cytosine remains intact. The methylation level in the sample can therefore be estimated by calculating the number of reads reporting a C (methylated), divided by the total number of reads reporting a C or a T (uracil converts to thymine during post-bisulfite PCR amplification). The workflow for carrying out targeted bisulfite sequencing begins with bisulfite treatment of DNA, followed by PCR amplification of the targeted region, library construction and sequencing of the amplicons. Targeted regions are amplified using specifically designed primers that target the region of interest. While highly useful, the limitations to the method include low quantitative accuracy, short read length, and low sample throughput [112].

1.3 Application of DNA methylation in forensics

Recent developments in DNAm detection and quantification technologies have led scientists to discover new applications of epigenetics in the field of forensic science. These new applications focus on solving some of the limitations associated with STR profiling technology. For instance, predicting characteristics that are changing throughout a person's life such as age, circumstances leading to death, or pathological states, is not possible with STR profiling. Furthermore,

STR patterns are identical in all tissues, which makes it impossible to determine the tissue type from which a biological sample originated. However, differential methylation of CpG markers has been discovered across chronological ages, pathological states, and tissue types, and thus DNAm has become a major area of interest within the field of forensic science [15].

1.3.1 Tissue identification

Determining the type of the body fluid recovered from a crime scene, especially in sexual assaults, may provide important information about how the crime events occurred and how they are linked to the perpetrator [113]. This has been done to date using two techniques, either protein- or RNA-based methods. Both techniques are based on the detection of proteins or RNA molecules that are specifically expressed in certain tissues. There are three protein-based methods that are widely used in forensic science for the detection of semen, blood and saliva. These methods are acid phosphatase (AP) and prostate specific antigen (PSA), which are used for detecting seminal fluid, haemoglobin tests for blood identification, and amylase tests to identify saliva [114,115]. However, it has been demonstrated that the sensitivity of these tests decreases with time, to the point the results can no longer be considered reliable [116]. Moreover, the mode of detection is exclusively based on a colour change, which is hard to detect in the case of minute amounts of a forensic specimen or environmentally degraded samples, thus false negatives can be introduced, in addition to a high degree of subjectivity. In addition, false positives can also be obtained with specific compounds that interfere with these tests. As such, forensic scientists have endeavoured to find alternative techniques that are more sensitive and reliable, such as the RNA based methods.

RNA based techniques have been utilised for tissue identification in forensic science. The DNA sequence does not provide any tissue specific information, however studying the differential expression of messenger RNA

(mRNA) and microRNA (miRNA) in different tissues can be used to identify the origin of body fluids [117]. RNA can be co-extracted from body fluid stains along with genomic DNA [118]. The main technique to identify differentially expressed genes in tissues is genome-wide expression profiling using microarray technology. Using this approach, specific mRNA markers have been discovered for the identification of different forensically relevant tissues including blood, saliva, semen, vaginal secretions and menstrual blood. These can be simultaneously analysed in forensic stains using multiplex reverse transcription endpoint PCR (RT-PCR) and quantitative PCR (RT-qPCR) assays [119]. Identifying the origin of mixed body fluids in a single reaction will also save the limited amount of genomic DNA in the forensic samples. Despite the fact that RNA is a vulnerable molecule and is thought to be very prone to degradation by ubiquitous ribonuclease (RNase) enzymes, a number of studies have shown that some samples stored at room temperature for more than one year contain mRNA suitable for RT-PCR [120]. However, the inherent instability of the molecule is a key disadvantage to the use of RNA for body fluid identification in forensic science, causing issues with the interpretation of negative results. Finding a tissue identification method based on DNA should significantly outperform RNA-based techniques due to the relatively higher long-term stability of DNA compared to RNA.

1.3.1.1 Tissue identification using DNA methylation analysis

During embryogenesis, CpG sites are differentially methylated in a tissue-specific manner, which is a vital process for cell differentiation [121]. Thus, each tissue has a distinctive DNAm profile, which has been successfully used for tissue identification [93,122,123]. Finding tissue identifiers that are directly linked to DNA can be very useful, especially in cases were DNA is the only evidence that has been retained from the crime scene [15]. Therefore, this will help accurately

identify body fluids in a non-destructive manner, protecting and preserving the DNA evidence.

Frumkin et al. (2011) were the first to demonstrate the possibility of identifying forensically relevant tissues using hypermethylated and hypomethylated CpG markers. Subsequent to this, An et al. (2012) successfully developed a tissue identification assay for blood, semen, saliva, vaginal fluids, and menstrual blood, using four tissue-specific differentially methylated regions (tDMRs) in four genes (*DACT1, USP49, PRMT2,* and *PFN3*). Moreover, they demonstrated that tDMRs are stable over time; regardless of the deposition time of the body fluid, the DNAm patterns of tDMRs remain stable [113]. This is consistent with previous research, which identified a set of specific DNAm biomarkers for bladder, colon, oesophagus, liver, lung, pancreas, stomach, brain, heart, kidney and spleen tissues that were collected from human autopsy [52]. Therefore, DNAm profiling in the near future is likely to play a key role in forensic cases, owing to its characteristic stability compared to other approaches to identify tissue type in a forensic context [6].

1.3.2 Age estimation

1.3.2.1 Current and historical methods for age estimation

In forensic science, there are multiple approaches to estimating the age of an unknown individual. For instance, in cases involving a deceased individual, morphological characteristics of the body have been used through radiological examination of skeletal and dental development in order to determine an age category [124]. The dental based method uses the developmental stages of the third molar(s) as evaluated using panoramic radiography, with a linear regression model to estimate age. The prediction accuracy of the dental based age prediction models is very high (two years from the actual chronological age) [125]. However,

this method is only reliable between 15 to 23 years old, at which point dental development stops, making this method impractical in cases involving adults [125]. Therefore, forensic scientists have redirected their attention towards molecular methods such as radiocarbon analysis and racemization of aspartic acid, which provide a better prediction accuracy compared to morphological based techniques.

Age can be estimated at a molecular level using the racemization of aspartic acid in dentine or tooth enamel. This is one of the oldest methods that has been used to predict age in forensic cases [124]. The rationale behind this method is that amino acids in cells and tissues are normally present in L-form, but during the course of aging they gradually transform by racemization to D-form enantiomers. In order to completely transform all L-amino acids in the human body to D amino acids by racemization, it would take 100,000 years at 25 °C [126]. The rate of racemization is controlled by various factors such as pH, temperature, and humidity. Aspartic acid is used for age prediction because of its fast racemization rate and the high correlation between the ratio of L-Asp and D-Asp with age. However, in tissues with a high metabolic rate such as liver, and brain, D-Asp is undetectable, whereas D-Asp can be measured in tissues showing low metabolic rate, thus providing better age prediction. Tissues with a low metabolic rate such as the dentine and enamel of teeth are therefore used for age estimation analysis [124]. Despite the high prediction accuracy that can be achieved using racemization, inconsistencies exist among papers, which report large differences in the level of accuracy [124,126].

Another molecular based method that has been recently introduced in forensic science that offers more accurate age estimation compared to racemization method, is radiocarbon or carbon-14 (^{14}C) analysis [127]. Unlike racemization analysis, radiocarbon dating predicts the birth date instead of chronological age of an individual, as ^{14}C decay continues even after death.

However, their chronological age can be determined when the deceased's date of death is known. The method is based on determining the year of tissue formation based on its $^{14}C/^{12}C$ ratio. ^{14}C is naturally formed by cosmic ray interaction with nitrogen-14 and its presence in organic materials has remained stable over the past several thousand years [128]. After the oxidation of ^{14}C atoms to CO_2, it enters the terrestrial biosphere through assimilation into plant biomass via the process of photosynthesis. Then it gets incorporated into living systems by consumption of plants or organisms that consume plants [129]. The detonation of nuclear weapons during the period of the Cold War (1955-1963) has nearly doubled the ratio of $^{14}C/^{12}C$ in the atmosphere. However, after the nuclear test ban treaty in 1963, the ^{14}C level has decreased linearly with time due to mixing with large marine and terrestrial carbon reservoirs. Interestingly, this steady decrease in ^{14}C has created bomb curve that can be used as an isotopic chronometer of the past 60 years [128].

Therefore, forensic scientists have used ^{14}C dating to predict the birth date of unknown individuals by using the upper limit of enamel formation in teeth, which contains the $^{14}C/^{12}C$ ratio that reflects the atmospheric ratio at the time it was formed. The prediction accuracy of this technique has been shown to be outstanding, with an overall absolute error of ±1 years [124]. However, the key limitation of radiocarbon dating is that it can be used to estimate the birth date if someone was alive after the period of the nuclear detonation, but will give false estimations of age if born before this period [128]. Moreover, in cases where the only evidence that can be obtained from the crime scene is DNA or body fluids, neither radiocarbon dating, nor racemization of aspartic acid are useful for age prediction.

1.3.2.2 Age estimation using DNA methylation analysis

Another potential application of DNAm is in the estimation of an individual's chronological age from their DNA, using age-specific CpG markers. This has been

established based on numerous studies, which have shown that global hypomethylation of the genome and regional hypermethylation of specific genes occurs during aging of cells and tissues [130-133]. Most of these studies were involved in studying chronic age-associated pathologies such as cancer, atherosclerosis, Alzheimer's disease, autoimmunity and macular degeneration [134-136]. For instance, in cancer patients, researchers have shown a global decrease in DNAm of the genome and significant hypermethylation in regulatory regions of the tumour-suppressor genes [28]. The main driver for this global decrease in DNAm is not well understood, but it could be explained by a decline in cellular DNMT1 with age [137]. It has been proposed that the molecular changes in DNAm patterns in cells and tissues during aging could be exploited to estimate the age of individuals for forensic application [15,124].

The main experimental approach to studying the relationship between age and the overall change in DNAm profile was to determine the overall change in the ratio of 5-methylcytosine and cytosine with age through enzymatic hydrolysis of genomic DNA followed by high-resolution separation of DNA fractions [138]. However, the development of array-based methods of screening the genome for changes in methylation patterns has allowed scientists to identify and characterise AR CpG sites and their association with genes [139]. A study conducted by Day et al. (2013) confirmed the presence of common and tissue specific AR CpG sites in humans, which previously had been identified in various rodent tissues [133,140]. Tissue-specific CpG sites are frequently found within non-CpG islands and mostly exhibit decreasing methylation with age (negative CpG sites). In contrast, the majority of AR CpG sites that are common across tissues are positively methylated with age (positive CpG sites) and positioned within CpG islands. Interestingly, the negative CpG sites are generally found close to tissue-specific transcriptional repressor binding sites in the genome [141]. Consequently, genes that are near negative CpG sites had higher expression levels than those near positive AR CpG sites. Since the effect of aging has been shown to be

associated with increased methylation, thus negative methylation of tissue-specific CpG sites may be protected against common age-dependent methylation, in order to maintain optimum tissue-specific gene expression [142].

Recently, there has been an increasing amount of literature using this significant correlation between the level of DNAm at different CpG sites with age, in order to predict a person's age [7,143,144]. The predicted age, which is referred to as DNAm age, can be used to address several questions in aging research, as well in forensic science [106]. Most of the AR CpG markers reported have been identified using the two genome-wide DNAm platforms, Illumina HumanMethylation27 (HM27K) and HumanMethylation450 (HM450K) BeadChips [136,143-145]. From the high-dimensional data on DNAm values generated from these platforms, AR markers are identified by testing the correlation of each marker with age either using Pearson's or Spearman's correlation coefficients [145,146]. Then, the age prediction model is built from these highly correlated markers by using either multivariate linear regression or quadratic regression in order to identify the DNAm age [8]. However, the number of highly correlated markers is usually large, and they cannot all be used to build the prediction model. For this reason, stepwise regression is used in order to construct models with all possible combinations of CpG sites and then evaluate their performance in order to select the best model with the lowest Bayesian information criterion (BIC) value. In contrast, penalised regression methods such elastic net regression, have the ability to select the most predictive markers from a pool of AR markers and construct prediction models without the need for reducing or filtering the data from the uncorrelated markers [106,136].

The methylation levels at the AR CpG markers assayed on the Illumina HM27K and HM450K chips are highly reproducible in other assays such as SNaPshot, pyrosequencing, EpiTYPER, and EpiTect Methyl II, which have been used to develop age-prediction assays by many researchers [92,145,147-150]. A

considerable number of studies have identified tissue-specific AR CpG markers in various tissues such as blood, semen, saliva, and teeth, and have developed age prediction assays that can be used for forensic purposes. For instance, based on blood samples, Weidner et al. (2014), developed a pyrosequencing-based assay to interrogate DNAm levels at three AR CpG markers residing in the *ASPA*, *EDARADD* and *PDE4C* genes, which were previously discovered on the Illumina HM27K platform and have an age prediction accuracy of 5.43 years [145] (Figure 1.14A). Moreover, a study conducted by Zbieć-Piekarska et al. (2015) developed a blood-based pyrosequencing assay that is based on seven AR CpG sites located in the promoter region of the *ELOVL2* gene, which was discovered on the Illumina HM450K platform. This assay produced a prediction accuracy of 5.03 years on training samples (Figure 1.14B) and ±7 years on an independent validation set of 124 blood samples.

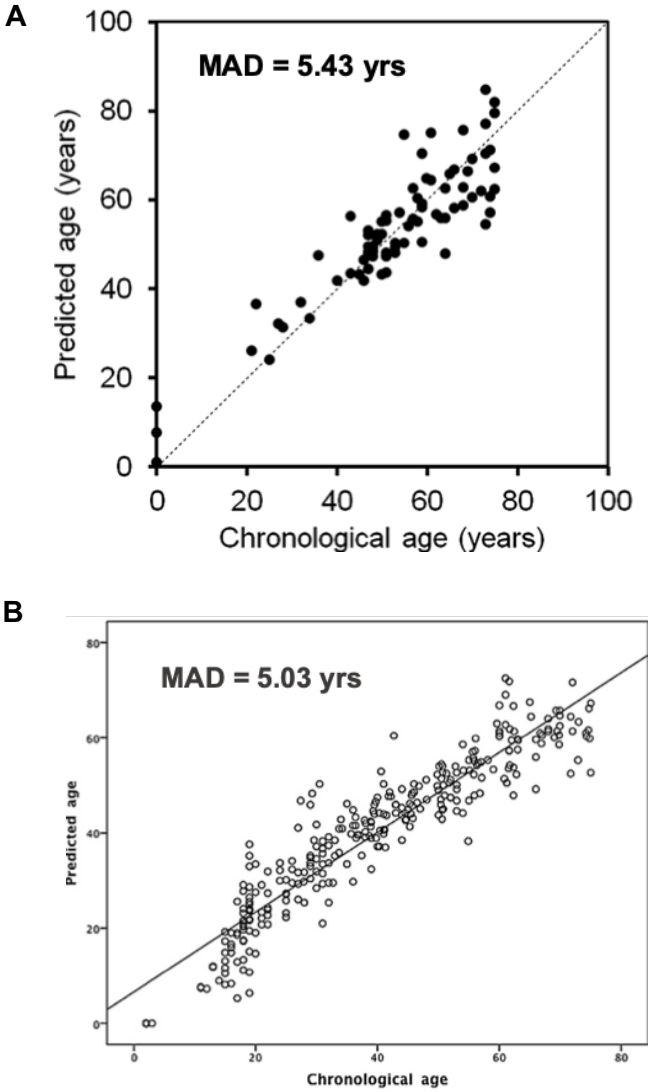

Figure 1.14 Predicted versus actual chronological ages of individuals used for building age-prediction models. Prediction accuracy calculated by mean absolute deviation (MAD) for (**A)** a blood-based model with three CpG sites and an accuracy of 5.43 years (MAD) [145]. (**B)** a blood-based model with seven CpG sites and 5.03 years (MAD) accuracy [148].

Due to differences in age ranges, methods, and statistical techniques, not all studies who examined the same type of tissue share the same set of AR CpG markers [146,149]. For example, most papers studying blood-specific AR markers identified different sets of CpG sites [143]. However, there are some consistent

AR CpG markers that are found between studies that reside in the *ELOVL2* gene, which has been linked to the photoaging response in human skin [151]. Interestingly, a large and growing body of literature has suggested that DNAm at a single CpG site in the *ELOVL2* gene promoter explains more than 70% of variation in human age, making it a very promising age predictor for various types of tissues (Figure 1.15) [136,146,152,153].

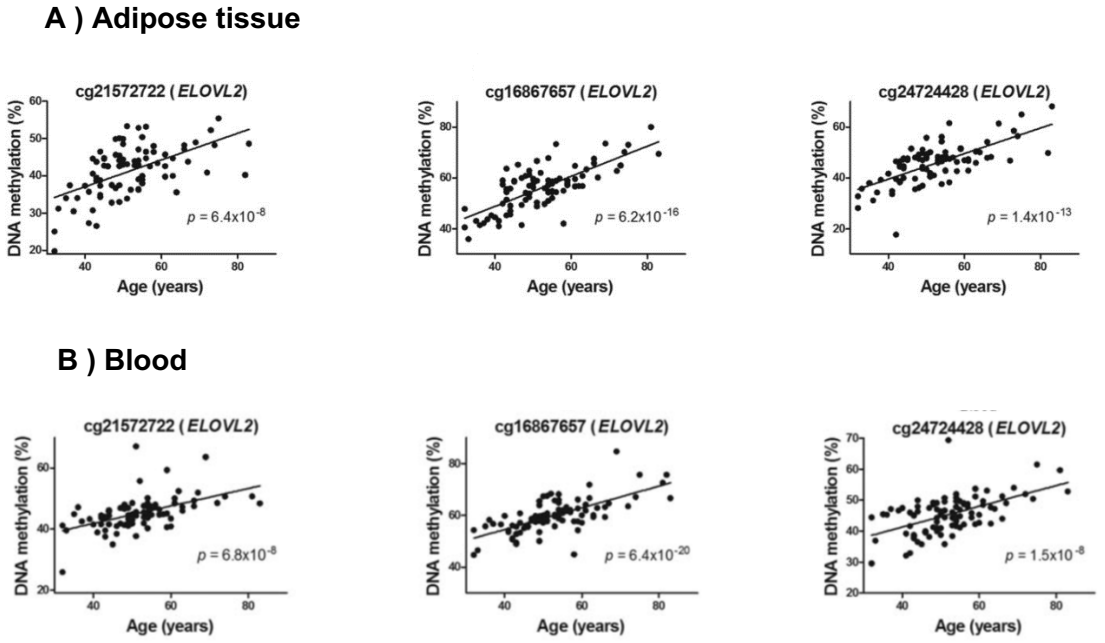

Figure 1.15 Correlation between DNAm level and chronological age at three CpG sites located within the *ELOV2L* gene. The "cg" numbers are Illumina's ID for the CpG sites and *p* is the *P*-value from a Spearman's correlation test. Data shown for **A** adipose tissue and **B** blood. (source [154]).

Tissue specific genes do not only contain tissue-specific AR CpG markers, but also AR CpG markers that are common with other tissues. Bekaert et al. (2015) identified AR CpG sites specific for teeth in the same genes that contain blood-specific AR CpG markers namely *ASPA, ITGA2B, PDE4C, EDAR-ADD,* and *ELOVL2* [149]. Moreover, other genes have been found to contain both saliva- and blood-specific AR CpG markers, namely *GRIA, NPTX2* and

EDARADD [155,156]. However, using the methylation level of one set of tissue-specific AR markers to predict the chronological age for other tissues produces very low prediction accuracy. Thus, researchers have endeavoured to identify universal AR CpG markers that can be used across multiple tissues to predict chronological age.

In the literature, two prominent studies looked for universal AR CpG markers across a wide range of tissues and built multi-tissue prediction models. By using DNAm profiles of 130 samples from five different tissues (dermis, epidermis, cervical smear, T-cells and monocytes) that were assayed on an Illumina HM27K BeadChip array, Koch and Wagner (2011) identified four universal AR CpG sites that are found in the vicinity of four genes, *NPTX2, TRIM58, GRIA2* and *KCNQ1DN*. These four universal markers were implemented in a prediction model and tested on 766 independent validation samples from different tissues (peripheral blood, cord blood, saliva, and breast samples), which provided a prediction accuracy of 11.7 years [94]. This prediction accuracy was significantly improved when a study conducted by Horvath (2013) used DNAm profiles of 8,000 samples encompassing 51 healthy tissues and cell types, also assayed on the Illumina HM27K BeadChip array. Horvath used elastic net regression, which automatically selected 353 universal AR CpG markers to construct a multi-tissue prediction model, which gave a prediction accuracy of 3.6 years on an independent validation set [106].

All of the publications found in the forensic literature are focused on the identification of CpGs correlated with age in a single body fluid/tissue type. There is very limited research focused on the identification of universal age-related CpG biomarkers that are common to all of the forensically-relevant tissues, including blood, saliva, semen, vaginal secretions, and menstrual blood. This is very important in terms of estimating the age of a biological sample that is recovered from a crime scene, where its identity (e.g. blood, saliva, semen, etc.) may be

unknown. Despite the fact that Koch and Wagner (2011) proposed 19 universal AR CpG markers across tissues and built a multi-tissue age prediction model, their model-building dataset was based on five tissues (dermis, epidermis, cervical smear, T-cells and monocytes) assayed on Infinium HM27K BeadChips, none of which was from a forensically-relevant tissue type. In addition, its age prediction accuracy was very low when tested on multiple tissues, with an average difference between the predicted and chronological age of 11 years, which is not useful when applied in forensic science. In attempt to overcome some of the issues in the Koch and Wagner (2011) study, Horvath (2013) used 51 different tissues and cell types to build a multi-tissue age prediction model. Although the prediction accuracy of this model was very high (3.6 years), when it was tested on semen samples in an independent study conducted by Lee et al. (2015) it produced a very poor prediction accuracy of 13.3 years. Furthermore, due to the large number of CpG markers (353 markers) that Horvath's model contains, it is not possible to develop a DNAm assay that includes all these markers and yet can be used for forensic purposes. Such an assay would be highly technical and expensive to be used for forensic purposes. In addition, even if there were an assay that could incorporate this large number of markers into a single test, the likelihood that forensic evidence would produce any results on such an assay would be very low due to the nature of these samples, which are usually low in quantity and quality. Highly technical and expensive

1.4 Research objectives

DNAm analysis presents an opportunity to take forensic genetics to a new level by answering questions that conventional DNA profiling techniques cannot answer. The potential applications of DNAm analysis in forensic science include:

1. Determination of the parental origin of alleles.
2. Authentication of DNA samples.

3. Discrimination between monozygotic twins.
4. Determination of the cause and circumstances of death.
5. Age estimation.
6. Body fluid identification.

Of these applications, this thesis is focused on age estimation using DNAm analysis and, in particular, on the following themes: identifying the optimum statistical methods for discovering AR DNAm markers, using these methods to build a saliva-specific age prediction model, identifying blood-specific AR DNAm markers using the newest Illumina microarray platform, and finally building a multi-tissue age prediction model for forensic applications that is capable of predicting the age of individuals regardless of the type of tissue being used.

To detect AR DNAm markers, researchers have used different statistical methods for testing the association between DNAm level and chronological age. However, there is no consensus as to which of these statistical methods is most efficient in identifying AR CpG markers. This will be explored in Chapter 3, which aims to establish a standard set of procedures that are optimum for selecting AR DNAm markers from high dimensional data generated using microarray platforms, in order to build highly accurate age prediction models. In the same Chapter, the selected statistical methods were used to enhance age prediction accuracy from saliva samples. This was achieved bioinformatically *in silico,* that is DNAm profiles from saliva samples assayed on the Illumina HumanMethylation450 (HM450K) BeadChip were downloaded from an online epigenetic data repository and then statistically analysed to identify AR markers for building a saliva-specific age prediction model. This model was then validated *in silico* using another independent set of DNAm profiles from saliva samples, assayed on the HM450K.

In Chapter 4, the saliva-specific AR DNAm markers identified in Chapter 3 were further validated by targeted bisulfite sequencing using the Illumina MiSeq®

platform. Their surrounding genomic regions were sequenced in order to discover additional CpG sites that may have a stronger association with age, which could be used to further enhance the accuracy of age prediction from saliva. The performance of the saliva-specific age prediction model constructed was compared with the best saliva-specific age prediction model reported in the literature, which was created by Hong et al. [157]. The reason for choosing saliva is that it constitutes a major source of DNA from various types of evidence collected at crime scenes, such as cigarette butts, chewing gum, toothbrushes, drinking/eating items, and also common in sexual offences. In addition, saliva sampling is non-invasive and convenient for medical screening and other diagnostic applications. For this reason, researchers have shown an increased interest in identifying AR DNAm markers for saliva samples.

Recently, a new array, the Illumina MethylationEPIC® (EPIC) BeadChip was introduced, containing over 860,000 probes. In Chapter 5 a comprehensive evaluation of blood-specific AR DNAm markers found on the EPIC BeadChip was performed, and their associated genes identified, which will provide new insights for researchers in various epigenetic and genetic disciplines. Enhancing the accuracy of DNAm based age-prediction models by searching for new AR CpG sites on the EPIC BeadChip with better age prediction accuracy will aid forensic investigations in criminal cases where biological samples of unknown origin have been recovered. For this reason, an age prediction model was constructed using the probes on the EPIC BeadChip, the performance of which was tested in comparison to models constructed using older Illumina microarray platforms.

Although DNAm markers have outperformed all other known AR biomarkers, such as telomere length and mitochondrial dysfunction, in terms of the accuracy with which they estimate biological age, AR DNAm markers are tissue-specific, and if used on other tissues will predict age with a large margin of error. Identifying a universal set of AR DNAm markers that can be used across

tissues to predict an individual's chronological age will therefore reduce estimation error as well as bypassing the necessity to first identify tissue type, a step that often exposes valuable DNA evidence to chemical destruction. Thus, the main aim of Chapter 6 was to identify a set of universal AR DNAm markers that are common across tissues and able to predict chronological age from body fluids that are frequently found at crime scenes (blood, saliva, semen, menstrual blood, and vaginal secretions).

Chapter 2: Materials & methods

2.1 Overview

Analysing DNA methylation (DNAm) level at specific age related (AR) CpG sites in a biological specimen recovered from a crime scene can potentially be used to estimate the chronological age of the unknown individual(s) who deposited the sample. However, identifying these AR CpG sites that can be used for this purpose requires managing and analysing the DNAm data sets through both bioinformatic protocols and statistical analyses. This section describes the bioinformatic pipelines and the statistical methods that have been used to identify the AR CpG sites from samples assayed on Illumina HumanMethylation BeadChips, starting from the first step of retrieving the data sets from the genomic repository, to the final step of building and testing the age prediction model. This protocol pertains to all data presented in Chapter 3, 5, and 6.

2.2 R software

In this thesis, R software [158], was used for downloading and processing of the Illumina HumanMethylation BeadChip data sets, and for all statistical analyses. R is a language-based environment used for statistical computing such as linear and nonlinear modelling, classical statistical tests, time-series analysis, classification, clustering, and for constructing a wide variety of graphics. Two terms will be frequently used throughout the thesis, which are function, and package. Function is a command line script that can execute a series of algorithms in R software, and package is a collection of functions created by a certain developer and stored together in code-based software that runs on R software. All the packages used in this thesis were downloaded from *Bioconductor* [159], which is a special electronic repository containing a broad

range of powerful statistical and graphical packages for the analysis of genomic data.

2.3 Genomic repositories and data sets

Identifying AR DNAm markers for building age prediction models requires a large sample size with a wide range of chronological ages. As described in Section 1.2.6.6, there are more than 117,000 epigenome profiles from a wide range of tissues and chronological ages that have been deposited into online genomic repositories, which can be exploited to answer various research questions in forensic science. For this reason, the Illumina HumanMethylation27 (HM27K), HumanMethylation450 (HM450K), and MethylationEPIC (EPIC) data sets used in this thesis were retrieved from three different genomic repositories: Gene Expression Omnibus (GEO) (http://www.ncbi.nlm.nih.gov/geo/) [160] run by the National Centre for Biotechnology Information (NCBI), the Cancer Genome Atlas (TCGA) (http://cancergenome.nih.gov/) [161] run by the National Cancer Institute, and ArrayExpress (https://www.ebi.ac.uk/arrayexpress/) [162] run by the European Bioinformatics Institute (EMBL-EBI). The downloaded data sets comprise of three components: an expression matrix, metadata (also known as phenotype data), and the feature annotation data frame. The expression matrix contains the DNAm values for each sample, and the metadata contains information about the experiment and samples, such as experimental platform, sample genders, disease status, ages, and ethnic origin. Finally, the feature annotation data frame contains information regarding the set of probes on the platform, such as the chromosomal location, associated gene(s) and/or associated promoter(s). In addition to the feature annotation data, there are always special annotation packages in R that can be download, which contain updated information regarding the probes.

These three components come in a folder called ExpressionSet, which can be directly downloaded into R software using R packages (Figure 2.1). The

columns in the expression matrix represent the samples, and the rows represent the probes, whereas in the metadata data frame the samples are in rows and the covariates (such as age, sex, disease status, etc.) in columns. Finally, the annotation data frame contains the probe information such as chromosome number, chromosomal coordinates, and gene association in columns, and probe names in rows. Each genomic repository has its own R package that can be used to download the data set from its own database. For example, ExpressionSets from the GEO, TCGA, and ArrayExpress databases are downloaded using *GEOquery, TCGAbiolinks, ArrayExpress* packages, respectively.

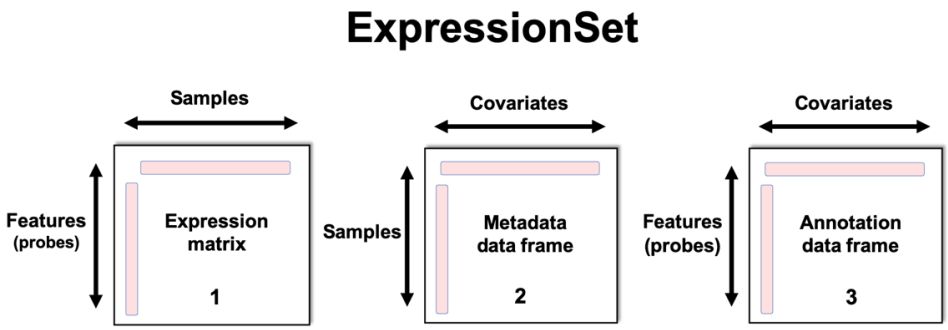

Figure 2.1 The diagram shows the three components that are embedded in an ExpressionSet when retrieved from genomic databases such as GEO, TCGA, and ArrayExpress directly into R software.

2.4 Processing Illumina HumanMethylation data

The downloaded Illumina BeadChip data (HM27K, HM450K, or EPIC) can either contain expression matrices with raw signal intensity values from the microarray instrument, or with Beta values already calculated, representing the DNAm levels at each probe. If required, the Beta values can be calculated from the raw signal intensities using equation 1.1. The processing steps of the Illumina HumanMethylation data are summarised in Figure 2.1, and are described further below.

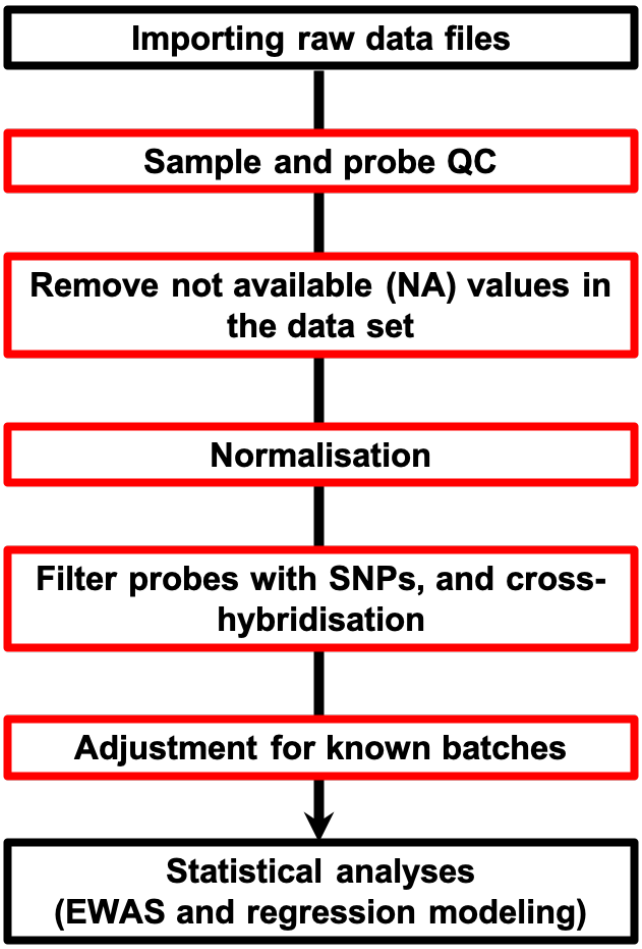

Figure 2.1 The analysis pipeline for Illumina HumanMethylation BeadChip data.

2.4.1 Sample and probe quality control (QC)

Managing and processing the probes and samples was done using a series of custom-written scripts in R software. Probes with a detection *P*-value ≥ 0.05, which is the probability that the signal intensity of the probe comes from background noise rather than from a true biological signal, were removed prior to analysis. In addition, replicates with large discrepancies in methylation levels were removed, as were samples with missing values in more than 50% of the probes. Furthermore, the samples assayed on the Illumina HumanMethylation BeadChips platforms should have a bimodal distribution of methylation Beta values, as shown

in Figure 2.2. Thus, samples deviating from this pattern indicates samples are corrupted or are otherwise outliers and should be removed prior to analysis. Density graphs were plotted using the *density* function in R software. In addition, outlier samples can also be identified using Single Value Decomposition, and/or cluster analysis (described in Section 2.4.4).

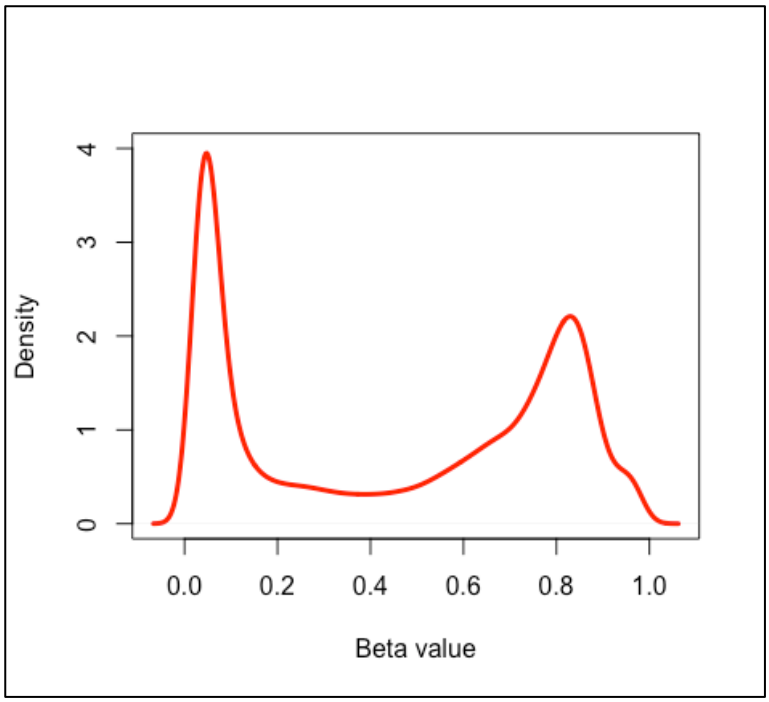

Figure 2.2 Density plot showing the bimodal distribution of the methylation Beta values.

2.4.2 Normalisation

As discussed in Section 1.2.6.6.3, it is necessary to normalise the data generated from Illumina HumanMethylation BeadChips in order to render the methylation values (Beta values) of Infinium II probes comparable with Infinium I probes, before conducting any downstream statistical analyses [108]. First of all, each downloaded data set was checked to determine whether it had been already normalised or not. This can be checked by viewing the density plot of the DNAm

levels, where normalised probes, will have Beta value distributions from Infinium I and II probes that are aligned with each other. For data sets with unnormalised probes, the Beta Mixture Quantile (BMIQ) method was performed, which is available in several R packages, including *ChAMP* [163], *RnBeads* [164], and *WateRmelon* [165]; the latter was used here for normalisation. The parameters used in the BMIQ command line were those recommend by the packages' authors. After normalisation, Beta value distributions were again checked by examining density plots, to see if the central peaks of the two probe designs were aligned with each other, indicating that the data had been normalised [108].

2.4.3 Probe filtering

Previous findings have indicated that there are a number of non-specific autosomal and sex-linked probes on the Illumina HumanMethylation BeadChips should be removed prior to the statistical analyses. The probe filtration was conducted using information in an annotation file developed by Illumina, and updated by Price et al. [166], detailing the probes to be removed. The annotation file was imported in R software and probes were removed accordingly. In addition, any probe containing a known SNP marker was removed from the data sets, also using the annotation file. Furthermore, to avoid gender bias in the prediction model developed, all probes targeting CpG sites located on sex chromosomes were also removed before downstream statistical analyses.

2.4.4 Detecting batch effects and outliers using singular value decomposition and cluster analysis

In both genome- and epigenome-wide association studies (EWAS), the values (gene expression and DNAm levels, respectively) are significantly affected by certain unwanted factors (also known as covariates) such as batch, sex, cell type, smoking, and, in some studies, chronological age. It is important to account for these covariates as they may cause confounding effects in EWAS [167]. One way to discover the presence of covariate effects is to determine whether there is

an association between samples that possess a given unwanted covariate. For instance, in the case of data on DNAm levels at two CpG sites (two variables) for a set of samples, by plotting a 2D scatter plot (each variable on one-axis), samples that have similar values and possess the unwanted factor (have the same batch, cell type, sex etc.) will be close to each other on the scatter plot. In this case, the researcher would know the association between the samples is due to the covariate rather than to the factors under study. The use of scatter plots can be extended to look at three CpG sites by plotting their values using a 3D plot, and the association between the samples can still easily be examined. However, when the number of CpG sites gets to more than 300,000, it is too difficult to visualise any association between samples, and thus this high-dimensional data needs to be assessed in a different way.

One of the methods used to visualise the association between samples in high-dimensional data is Singular Value Decomposition (SVD) [168]. The basic concept of SVD is similar to the principal component analysis, which reduces the number of variables into a small number of abstract variables known as singular vectors. These singular vectors are ranked from 1 to the number of variables in the data, however the first three singular vectors explain the most variation in the data and reflect the underlying structure of the data in terms of the relationships between the samples. By plotting the values of the 1st and 2nd singular vectors in a 2D scatter plot, or a 3D plot using the 3rd singular vector, the association between samples will appear on the scatter plot. SVD analysis can also be used to identify outlier samples, which can be seen as samples clustering away from their original sample type.

The SVD analysis was carried out using a built-in function in R software called *svd*. The function returns the ranked singular values in a three-object list (*v*). For the purpose of examining whether each sample type will form a cluster based on tissue type, rather than other factors that may represent hidden

confounding variables in the data, the 1st and 2nd singular vectors for the samples were plotted against each other in a 2D scatter plot using the first and second columns in the *v* list.

Cluster analysis was also used to discover whether there was any relationship between DNAm level for samples with covariates other than their tissue type, such as batch, or sex. In the cluster analysis the samples will be clustered or separated from each other based on the similarity/dissimilarity between samples using their DNAm patterns, which was calculated using the Euclidean distance between each sample, as shown in equation 2.1. The samples would normally cluster to their sample type. However, in case of batch or sex effect, samples with the same batch or sex would cluster to each other.

$$d(x, y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2} \qquad (2.1)$$

Where *x* is the DNAm level at a given CpG site in sample1, and *y* is the DNAm level in sample2. In R software, the *dist* function was used to calculate the Euclidean distance between samples based on their DNAm profiles. Subsequently, the hierarchical clusters were formed using *hclust* function in R and illustrated by a dendrogram which was plotted using *as.dendrogram* function in R.

2.4.5 Estimating and adjusting for cell type composition

As described in Section 1.3.1.1, there are CpG sites that are differentially methylated between tissues and cell types, therefore any change in the cell type composition in heterogenous tissues, such as blood, will also change the DNAm levels at some CpG sites. Since it has been demonstrated that the cellular constituents of blood change with aging, if the cell type composition is not accounted for, any AR CpG sites that are identified might be false positives. That

is, the change in DNAm level at these CpG sites is due to the change in cell type composition rather than aging. If this is case, the cell type composition will be a confounding variable. In order to prevent any confounding effect from the cell type composition, any AR CpG sites identified in blood as part of this thesis were tested for any confounding effects as a result of cell type composition. This was done by calculating the change in the coefficient for the age term in regression equations before and after including cell type composition as a variable. If the change in the coefficient for age was within 5%, it was concluded that the cell type had no confounding effect and should be ignored [169]. Otherwise, the cell type was determined to be a confounder and therefore would need to be included in the regression equation to correct for this effect [106,170,171]. Blood cell composition in samples was estimated based on their DNAm profiles, using a regression calibration algorithm (model) created by Housemen et al. [172], which is implemented in the *estimateCellCounts* function in the *minfi* package [173]. This function estimates the proportion of the six blood-cell types in each sample: CD8T, CD4T, natural killer cell, B cell, monocyte and granulocyte.

2.5 Identifying AR CpG sites and constructing age prediction models

Direct construction of a prediction model using a high dimensional data such as HM27K, HM450K, or EPIC data with 27,000, 450,000, and 850,000 probes, respectively, is not possible. Therefore, there are four major steps that should be implemented in order to construct a prediction model from high dimensional data set:

1- Variable reduction
2- Variable selection
3- Building the model
4- Testing the model

The variable reduction step is performed in order to substantially reduce the number of CpG sites into a manageable dataset that can be used in the downstream statistical analyses. This is done by identifying the CpG sites that are AR using correlation/regression tests and excluding non-AR CpG sites. Then, after reducing the number of CpG sites and ending up with a manageable number of AR CpG sites, the next step would be variable selection, during which the best subset of those AR CpG sites is selected for use in the building of an age prediction model. In the final step, the age prediction model is constructed using one of two regression modelling systems. Figure 2.3 summarises the steps in the construction of age prediction models from an expression matrix containing samples with DNAm values.

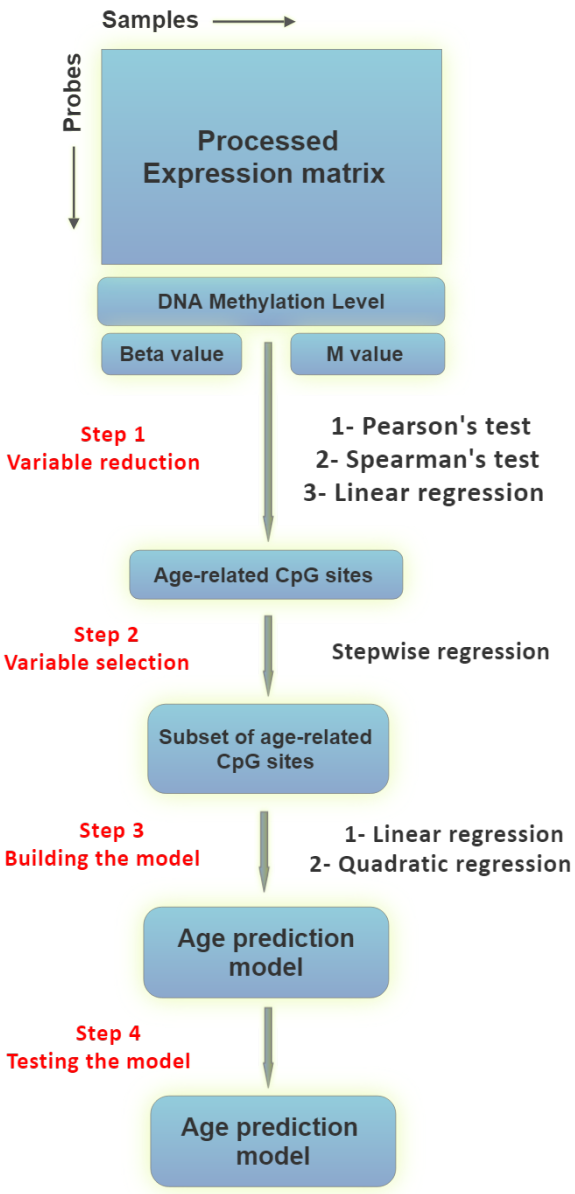

Figure 2.3 Schematic diagram illustrating the main steps in constructing age prediction models from high dimensional data.

2.5.1 Variable reduction

The number of CpG sites in the expression matrix can be reduced by identifying which of the sites show a relationship with age, using one of the three following statistical tests:

2.5.1.1 Pearson's correlation test

Pearson's correlation test measures the relationship between two variables by fitting the best straight line through their points on a scatter plot [174]. The sign (±) of the slope value of the fitted line indicates whether the relationship is positive or negative. For instance, in the case of a positive slope, as the points of variable A increase the points of variable B increase, and vice-versa in the case of a negative slope. There are two requirements needed in order to implement a Pearson's correlation test; the variables being analysed should be normally distributed, and the relationship between the two variables should be linear rather than monotonic [175]. The result of the correlation test is indicated by the correlation coefficient (r), which is an index of how close the points on the scatter plot fit the best-fitting straight line. The correlation coefficient can be calculated using the following equation:

$$r = \frac{\sum XY - \frac{\sum X \sum Y}{N}}{\sqrt{(\sum X^2 - \frac{(\sum X)^2}{N})\,(\sum Y^2 - \frac{(\sum Y)^2}{N}}} \qquad (2.2)$$

Where, *X* is the values of X, *Y* is the values of Y, and N is the number of the pairwise combinations of points in the data. The numerical value of the correlation coefficient ranges from 0-1 with either a positive or negative sign, as described above. A correlation coefficient close to 0 means there is no correlation between the variables, whereas a value >0.5 indicates that there is a correlation between the two variables under study. During this project, the Pearson's correlation test was conducted using custom scripts written in R, to measure the correlation between DNAm level (using the Beta value) at each CpG site and the chronological age. The correlation coefficient cut-off value for selecting CpG sites as being AR was for them to have an absolute (abs) r ≥ 0.5 as recommended by various studies [8,94,145].

2.5.1.2 Spearman's rank correlation (rho)

There are cases where Pearson's correlation test cannot be implanted, such as when the relationship between the variables under study is non-linear, and/or when the distribution of their scores are markedly asymmetrical and do not approximate a normal distribution. In this case, Spearman's rank correlation or Spearman's rho can be used. It measures the monotonic relationship between variables and does not assume normal distribution of the variables [175]. Spearman's rho is calculated using the same equation as Pearson's correlation coefficient (2.2), however, the values of both variables are ranked from smallest to largest [174]. That is, the smallest value for variable *X* is given rank 1, the second smallest value for variable *X* is given rank 2, and so forth. Spearman's correlation coefficient (rho) is then calculated using the following equation:

$$rho = \frac{\sum X_r Y_r - \frac{\sum X_r \sum Y_r}{N}}{\sqrt{(\sum X_r^2 - \frac{(\sum X_r)^2}{N})\,(\sum Y_r^2 - \frac{(\sum Y_r)^2}{N}}} \qquad (2.3)$$

Where X_r is the rank values of X, Y_r is the rank values of *Y*, and N is the number of the pairwise combinations of points in the data. A Spearman's rank correlation test was carried out between DNAm level at each CpG site and the chronological age of the donor, using a series of custom scripts written in R software. The cut-off value for selecting CpG sites as being AR was abs rho ≥ 0.5.

2.5.1.3 Simple linear regression

Numerically, a simple linear regression coefficient (R^2) is the square value of the correlation coefficient between two variables, which also describes the relationship between those variables, but in terms of how accurately the *X* variable on the x-axis can predict the value of the *Y* variable on the y-axis [174]. The *X* variable is called the "predictor", "explanatory" or "independent" variable, while

the *Y* variable is the "dependent", "response" or "outcome" variable. The main difference between correlation tests and linear regression tests is that the former quantifies the degree to which two variables are related, by computing the correlation coefficient (r) and it does not fit a line though the data points. However, simple linear regression finds the best fitting line through the points of the two variables on the scatter plot, and then uses this line to predict any value of *Y* from the corresponding *X* value. As illustrated in Figure 2.4, practically speaking, this can be done by drawing a vertical line from any value of the *X* variable from the x-axis to the regression line, and then draw a horizontal line from the regression line to the y-axis, which will contain the predicted *Y* value.

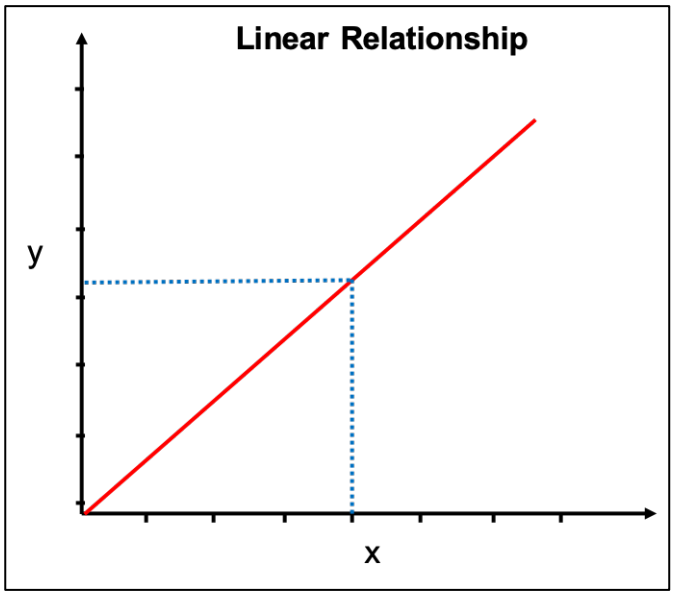

Figure 2.4 Predicting the value of a Y variable from an X variable in the case where their relationship is linear.

In this thesis, the DNAm value at a given AR CpG site will be the independent variable, and the age, which is to be predicted (DNAm age), is the dependent variable. The best fitting line can be drawn using the regression line equation:

$$Y = a + \beta X \qquad (2.4)$$

Where *Y* is the predicted value, α is the y-intercept (where the regression line intersects with the y-axis), β is the slope, *X* is any known value of *X*, and ε is the residual standard deviation. The slope (β) (also known as the "coefficient") of the regression line is given by the following formula:

$$\beta = \frac{\sum XY - \left(\frac{\sum X \sum Y}{N}\right)}{\sum X^2 - \frac{(\sum X)^2}{N}} \qquad (2.5)$$

Where *X* is the values of the independent variable on the x-axis, Y is the values of the dependent variable on the y-axis, and *N* is the number of observations. The intercept (α) of the regression line on the y-axis can be calculated as follows:

$$a = \frac{\sum Y - \beta \sum X}{N} \qquad (2.6)$$

Therefore, by determining the values of the y-intercept and the slope we can predict the values of the y variable from the *X* variable by using the regression line equation. In this project, simple linear regression analysis was used to regress chronological age of the donors of biological samples on the DNAm level at each CpG site in the data set in order to select the best CpG sites for use in accurately predicting the chronological age. The predicted Y variable would therefore be the DNAm-predicted age. The *lm* function in R, along with a custom script written in house, were used to conduct linear regression analysis between the DNAm level at each CpG site and the chronological age of the donor of the samples.

2.5.1.4 False discovery rate (FDR)

In high dimensional data sets, if the significance level of the correlation and regression tests is set to a *P*-value ≤0.05, then the set of significant results would contain 5% false positives. Thus, in genome-wide data sets, the use of this significance level will result in a large number of false positives, due to the large number of CpG sites being tested. Therefore, in order to counteract this problem (the multiple comparisons problem), the incidence of false positives should be significantly lowered by using a very low significance level, at which the rate of false positives is reduced to only 5% of the total false positives at the *P*-value ≤0.05. This false positive rate is referred to as the false discovery rate (FDR), and the new extreme *P*-value at any specific FDR is called the q-value. In this study, the appropriate q-value at which the FDR is ≤0.05 was calculated in R using the *q-value* package [176]. This was done by taking the *P*-values from each test used (Spearman, Pearson, or linear regression), storing them in a vector and passing them into the *q-value* package, along with an FDR level ≤0.05. After analyzing the *P*-values, the *q-value* package produced a vector containing only CpG markers that were significantly associated with age in the tested tissue.

2.5.2 Variable selection by stepwise regression analysis

After excluding those CpG sites that do not correlate with age, a data set remained that contained only AR CpG sites. After this, the next step was variable selection, during which the best subset of sites in the AR CpG data set for building an age prediction model were selected. This can be done using a stepwise regression analysis, which examines all possible combinations of the AR CpG markers selected in the previous step and identifies the best combination of those sites for predicting age. The criteria used to select the best set of sites is based on the Bayesian Information Criterion (BIC), which measures the efficiency of the parameterised model in terms of its ability to predict the relevant variable. Mathematically, BIC was calculated using the following equation:

$$BIC = -2 \times \log(L) + D \times \log(N) \qquad (2.8)$$

Where L is the maximised value of the likelihood function for the estimated model, N is the sample size and D is the total number of the particular combination of the markers in the model. The model with the smallest BIC value will have the best prediction accuracy, with the lowest possible number of variables.

As discussed in section 1.2.6.6.2, the Beta values produced from DNAm analysis are not normally distributed, and thus, they do not satisfy the normality assumption of regression analysis methods. For this reason, the Beta values were transformed into M values using the *Beta2M* function in the *WateRmelon* package, before conducting the stepwise regression analysis. The training data was prepared by including only the highly AR CpG markers, selected in the variable reduction step, and excluding all non-correlated CpG markers. Stepwise regression analysis was performed using the *regsubsets* function from the *leap* package in R software.

2.5.3 Model building

After selecting the best subset of AR CpG sites using the stepwise regression test, the next step is to construct the age prediction model. As described in Section 2.5.1.3, the equation (2.4) created by regression analysis describes how accurately the *X* variable (predictor) can predict the value of the *Y* variable (response). Therefore, the value of the X variable in the regression equation can be substituted by any value to predict its expected response. Thus, the regression equation can be used as a prediction model. The most important term(s) in the regression equation is the coefficient (or coefficients in the case of more than one predictor being used), which defines the relationship between the predictor(s) and the response variable. The process of defining and creating the values of the coefficients in the regression equation is known as training the model, and the samples used in this process are called training samples (or

training data set). Note that not all samples are used to train and test the model, instead the samples should be randomly split into a training and testing set. In this thesis, the *sample* function in R was used to randomly split the samples into two sets.

The number of predictors being used and the type of relationship between the predictors and responses determines whether the regression analysis (modelling system) used is linear or nonlinear. The following two modelling systems were used to build prediction models:

2.5.3.1 Multivariate linear regression

As explained in Section 2.5.1.3, the main function of regression tests is to predict the value of the Y variable using a predictor, which is the X variable. However, there are cases where Y can be predicted using more than one predictor (X_1, X_2, …. X_n). The regression method that can examine more than one predictor is multivariate linear regression. The multivariate regression equation is as follows:

$$Y\ (DNAm\ age) = a + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_n X_n \qquad (2.9)$$

Each X predictor (in this case, CpG site) in the multivariate linear equation has its own regression coefficient (β). For the age prediction model generated in this project, the specific predictors are the AR CpG sites determined in the stepwise regression test. After this, the multivariate linear regression model is built based on the training samples, which are used to calculate the coefficient values for each CpG site. In R software, the *lm* function was used to construct the age prediction model.

2.5.3.2 Quadratic nonlinear regression

While multivariate linear regression is adequate for modelling a wide range

of relationships in biological systems, many situations require nonlinear functions. There are different types of nonlinear relationships between variables that can be modelled using polynomial regressions. However, the nonlinear function focused on in this project is the monotonic relationship, which is the relationship seen when the response variable (*Y*) steadily increases (or decreases) with the independent variable (*X*) but the rate of increase (or decrease) becomes smaller and smaller, with the response variable reaching a plateau (Figure 2.5). This type of nonlinear relationship can be modelled using a type of polynomial regression known as quadratic regression, which is a type of multivariate regression. In the linear regression equation (2.4), the y variable increases (or decreases) by β units for each unit increase (or decrease) in the X variable. However, in a nonlinear relationship, estimation of the y variable can be improved by including a second squared term of the X variable in the multivariate linear regression analysis:

$$Y\ (DNAm\ age) = a + \beta X + \ \beta X^2 \qquad (2.10)$$

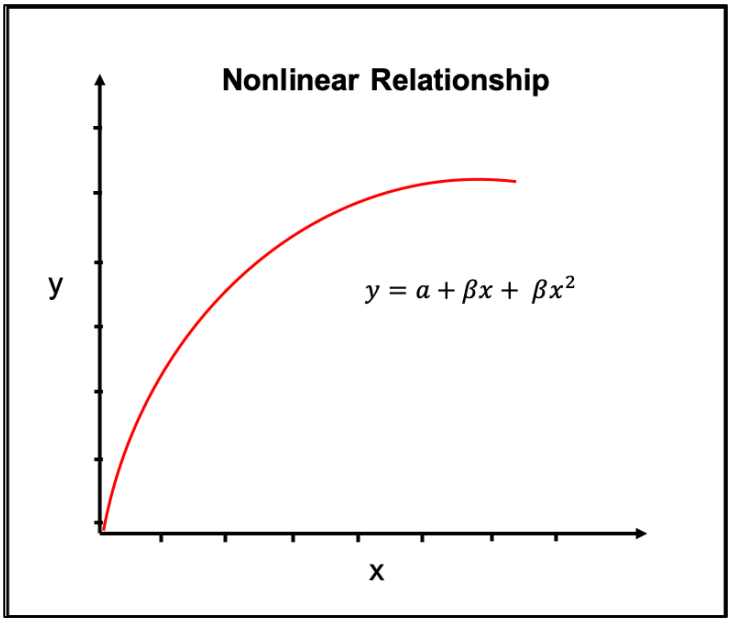

Figure 2.5 A monotonic (nonlinear) relationship between two variables can be captured by fitting a quadratic regression (red line), which is generated by adding extra squared values of X into the regression equation.

In R software, the monotonic relationship between methylation level at the AR CpG sites and chronological age can be captured by including an extra term based on squared Beta values alongside the standard Beta value term for each CpG site, using the *lm* function.

2.5.3.3 Elastic net regression

The first three of the aforementioned steps used to build age prediction models from high-dimensional data can be conducted in a single step using elastic net regression. Elastic net regression is a penalised algorithm used for variable reduction, selection, and model building in a single step, which is particularly useful in cases where the number of variables exceeds the number of samples [106]. In this study, elastic net regression was conducted using the *glmnet* package in R software, to reduce the number of CpG sites, select the best subset of AR CpG markers, and build these into a prediction model. To select the best

model, elastic net regression performs what is called “ten-fold cross-validation”, that is, the algorithm splits the training set into ten parts, one part serves as training set and the rest as validation sets, and it does this ten times (ten-fold). Each time, the average error and standard deviation are computed, and those subsets of markers that have the lowest estimation error will be selected as the best model.

Since elastic net regression is a penalised progression method, the number of CpG sites in the prediction model can be controlled by a value known as the lambda value. The lambda values that correspond to different numbers of CpG sites in the models, starting from one CpG marker to the optimum number of CpG markers, is provided by the *cross-validation* function in the *glmnet* package. The final age prediction model contains the intercept (α), and the coefficients (β_0, β_1, ..., β_n) that correspond to each CpG site, which relate to the chronological age as follows:

$$y\ (DNAm\ age) = a + \beta_1 CpG_1 + \cdots + \beta_n CpG_n \qquad (2.11)$$

Thus, the regression model can be used to predict the age value by simply substituting the Beta values of the selected CpGs into the formula.

2.5.4 Testing the model

For unbiased assessment of a constructed age prediction model, the samples used to train the model should not be used again for testing the model. Instead, independent samples should be used to evaluate the performance of the constructed model. Note that the samples in the testing data set used to validate the age prediction model should be assayed on the same assay system that was used for the training data set. Testing the performance of the model was carried out by calculating the mean absolute deviation (MAD) from the chronological age, which is the absolute difference between the predicted age and the chronological

age for an individual donor in the data set. Using the MAD value for assessing the accuracy of age prediction models has now been used in various age-related studies [106,144,149]. For further evaluation of the obtained MAD value, bootstrap analysis was carried out, which involves sampling the testing set with replacement 10,000 times and calculating the MAD between predicted and chronological age in each bootstrap cohort. From the distribution of the bootstrap observations (Figure 2.6), the 95% confidence interval around the mean of the MAD value was calculated from the bootstrap distribution. Based on this, the bootstrap analysis provides a range of MAD values that is expected in 95% of samples randomly drawn from the population. The bootstrap analysis was carried using custom scripts written in R.

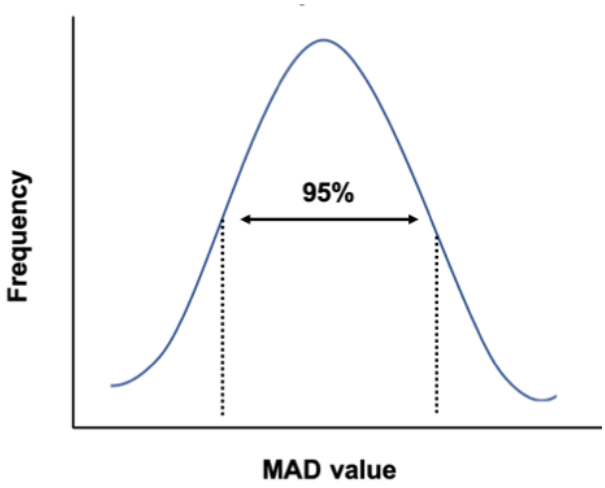

Figure 2.6 Bell curve illustrating the distribution of MAD values calculated in different bootstrap cohorts. The 95% confidence interval represents the range of MAD values recorded by 95% of bootstrap cohorts.

2.6 Next-generation sequencing on the Illumina MiSeq® platform

The following sections describe the steps carried out for validating the saliva-specific AR CpG sites that were identified in Chapter 3, by targeted bisulfite sequencing using the Illumina MiSeq® platform. The principle of using next-generation sequencing to assess DNAm level at specific genomic regions is described in Section 1.2.6.7. These sections pertain to the data presented in Chapter 4. The sample collection, DNA extraction, DNA quantification, statistical analyses were carried out at University of Strathclyde. However, assessment of the quantity and quality of the genomic DNA (gDNA), primer design, sequencing of the targeted specified regions of interest (ROI), and sequence alignments were carried out by Zymo Research Corporation.

2.6.1 DNA extraction using QIAamp® DNA Mini Kit

DNA was extracted from saliva using the QIAamp® DNA Mini Kit (Qiagen, Hilden, Germany) as described in the manufacturer's standard guidelines [177]. For each sample, 200 µL of saliva was added to a 1.5 mL microcentrifuge tube containing 20 µL proteinase K. Then, 200 µL of Buffer AL was added to the sample and mixed by pulse-vortexing for 15 seconds. This mixture was incubated at 56°C for ten minutes, and then briefly centrifuged to remove any liquid from the inside of the lid. 200 µL ethanol (96–100%) (Sigma, Gillingham, UK) was added to the sample and mixed by pulse-vortexing for 15 seconds, and then briefly centrifuged to remove any liquid from the inside of the lid. This mixture was transferred to the QIAamp Mini spin column and centrifuged at 6000 x *g* (8000 rpm) for one minute. After centrifugation, the QIAamp Mini spin column was transferred into a clean 2 mL collection tube, and the tube containing the filtrate was discarded. Then, to the QIAamp Mini spin column, 500 µL Buffer AW1 was added and centrifuged at 6000 x *g* (8000 rpm) for one minute. Again, the QIAamp Mini spin column was

transferred into a clean 2 mL collection tube, and the filtrate was discarded. 500 µL Buffer AW2 was added to the QIAamp Mini spin column and centrifuged at full speed (20,000 x *g*; 13,000 rpm) for three minutes. The filtrate from the previous step was discarded and the QIAamp Mini spin column was transferred into a new 2 mL collection tube and centrifuged at full speed for one minute. Again, the filtrate was discarded and the QIAamp Mini spin column was transferred into a clean 1.5 mL microcentrifuge tube. To the QIAamp Mini spin column, 25 µL Buffer AE was added and then incubated at room temperature (15–25 °C) for five minutes. After incubation, the mixture was centrifuged at 6000 x *g* (8000 rpm) for one minute. Another 25 µL Buffer AE was added to the QIAamp Mini spin column and then centrifuged at 6000 x *g* (8000 rpm) for one minute. The 50 µL eluted sample that contained the extracted genomic DNA (gDNA) was stored at -30 to -15°C.

2.6.2 Quantifying DNA before outsourcing the samples

The amount of the extracted DNA in each sample required by Zymo was ≥500 ng (ultraviolet absorbance: A260/A280 ratio > 1.7) in at least 20 µl, that is at least 25 ng/ µL. NanoDrop-100 Spectrophotometer (Thermo Scientific, Wilmington, DE, USA). In this study, the DNA samples were used only when the A260/A280 ratio was >1.7, with DNA concentration of ≥20 ng/50µL.

2.6.3 The quantity and quality of the extracted DNA (by Zymo)

The concentration and quality of the gDNA was measured using the Genomic DNA ScreenTape system (Agilent Technologies, Germany). The quality of the gDNA is indicated by the DNA integrity number (DIN) and the recommended quantity and quality of gDNA for the next-generation sequencing on the Illumina MiSeq® platform are ≥10ng/µL and ≥3 DIN, respectively.

2.6.4 Primer design

Primers were designed using Rosefinch software (Zymo Research's proprietary primer design tool), to target the specified regions of interest (ROI), i.e. the AR CpG sites that were identified in this study (see Appendix A: supplementary materials). Design parameters were chosen such that PCR amplicons would ideally be larger than 100 bp but smaller than 300 bp. In addition, primers were designed to avoid annealing to CpG sites at the ROI to the maximum extent possible. All primers were resuspended in TE buffer at 100 µM, then mixed and diluted to 2 µM. All primers were tested using real-time PCR with 1 ng of bisulfite-converted control DNA, in duplicate individual reactions. High-resolution melt curve analysis was performed to confirm the presence of a single specific PCR product.

2.6.5 Targeted bisulfite sequencing

Following primer validation, gDNA from the saliva samples was bisulfite converted using the EZ DNA Methylation-Lightning™ Kit (Zymo Research, CA, USA), according to the manufacturer's instructions. Library preparation and multiplex amplification of all samples using the ROI-specific primer pairs was performed using the Access Array™ System (Fluidigm, CA, USA), according to the manufacturer's instructions. The resulting amplicons were pooled for harvesting and subsequent barcoding according to the Fluidigm instrument guidelines. After barcoding, samples were purified using the ZR-96 DNA Clean & Concentrator™ Kit (Zymo Research, CA, USA), prepared for massively parallel sequencing using the Illumina MiSeq® V2 300bp Reagent Kit (Illumina, CA, USA), and then sequenced using a paired-end sequencing protocol, according to the MiSeq® manufacturer's guidelines.

2.6.6 Sequence alignments

Sequence reads were identified using standard Illumina base-calling software and then analysed using a Python-based analysis pipeline created by Zymo Research. Low quality nucleotides and adapter sequences were trimmed during QC processing. Sequence reads were aligned back to the reference genome using Bismark, which is a sequence aligner optimised for bisulfite sequence data and methylation calling [178]. Paired-end alignment was used as default, thus requiring both read 1 and read 2 to be aligned within a certain distance, otherwise both reads were discarded. Index files were constructed using the Bismark genome preparation command and the entire human reference genome. The non-directional parameter was applied while running Bismark and all other parameters were set to default. Nucleotides in primers were trimmed from amplicons during methylation calling. As described in Section 1.2.6.7, the methylation level of each sampled cytosine was estimated as the number of reads reporting a C, divided by the total number of reads reporting a C or a T.

Chapter 3: Finding the optimum statistical method for identifying age related CpG sites

3.1 Introduction

Recently, researchers in forensic genetics have shown an increased interest in the use of DNA methylation (DNAm) markers for age estimation in forensic casework. This has led to the introduction of various statistical methods that can be used to identify age related (AR) CpG sites for use in constructing age-prediction models. As described in Section 2.5, building an age prediction model from high dimensional data produced by high-throughput technologies such as the Illumina HumanMethylation27 (HM27K), HumanMethylation450 (HM450K), and MethylationEPIC® (EPIC) BeadChip platforms consists of four main steps: variable reduction, variable selection, building the model, and finally testing the model. The variable reduction step is done to reduce the number of variables (CpG sites) in the data to a manageable size for any downstream statistical analyses. The second step, variable selection, involves finding the best subset of this reduced number of CpG sites with the highest predictive accuracy. Finally, the third and fourth steps involve building the model using a regression modelling system and then testing the prediction accuracy of the constructed model.

In the variable reduction step, the majority of epigenetic age-prediction studies dealing with high dimensional data reduce the dimensionality of their data by removing CpG sites that are uncorrelated with the chronological age of the sample donor. The association between DNAm level at the CpG sites and chronological age is determined mainly using three statistical tests, namely Pearson's correlation, Spearman's rank correlation, and simple linear regression.

However, age-related studies have not all agreed on one standard test that is best for identifying age-related markers. The lack of a standard test that can be used as a benchmark is problematic, because it means that each study has to develop its own statistical methods for identifying AR markers, and this is likely to lead to different outcomes, as well as resulting in duplicated effort. In addition, there is no previous study demonstrating which of these statistical methods that can be used in the variable reduction step is optimum for identifying AR CpG sites for the purposes of building age-prediction models. Selecting a standard test of this nature would therefore improve the accuracy and reproducibility of the results of these types of studies, as well as making them more comparable.

Another parameter in DNAm data that could alter the outcomes even when using the same type of statistical test is the type of DNAm measurement being used. As mentioned in Section 1.2.6.6.2, there are two types DNAm metrics commonly used in the analysis of Illumina HumanMethylation BeadChip DNAm data, Beta and M values. One of the major limitations of the Beta value is that it does not satisfy the normality assumption of regression analyses. Despite this limitation, a number of studies have used Beta values in linear regression analyses to identify AR CpG sites [92,157,179]. This issue can be easily solved by taking the Logit transformation of Beta values, to give M values, which will satisfy the normality assumption as well as reducing the heteroscedasticity of the data at highly methylated or highly unmethylated CpG sites [157]. A review of the literature revealed that there is no previous study examining the effect of using the two different DNAm measurements (Beta and M values) on the efficiency of identifying AR CpG sites.

After identifying AR CpG markers, the next step in the process would be variable selection, that is, selecting the best subset of these AR CpG sites to build an age prediction model with the highest estimation accuracy. In the literature, this is usually done using stepwise regression analysis [157]. As described in

Section 2.5.2, this method constructs models with all possible combinations of the markers and then select the best one, with the highest predictive accuracy. One major requirement for stepwise regression analysis is that the data should be normally distributed. However, although Beta values are not normally distributed, they have been used in stepwise regression analysis in some age-prediction studies [108]. The outcomes of these studies could have been enhanced, if the most accurate statistical methods had been used, combined with the most appropriate DNAm measurement.

Based on the aforementioned, age prediction accuracy can be further enhanced by implementing the optimum statistical method for identifying the AR CpG sites, and using it with the right DNAm measurement. Finally, using the appropriate DNAm value that satisfies the algorithmic assumptions of the stepwise regression analysis could also potentially enhance selection of the best prediction model.

3.2 Aims

The main aim of this preliminary study was to identify the optimum method in the variable reduction step for selecting AR CpG sites from high dimensional data generated using the HM450K BeadChip platform. Identifying such a method at this stage would enhance the outcomes of the upcoming studies in this thesis, which will aid in building age prediction models with better prediction accuracies. The aim of the next part of this study was to identify saliva-specific AR CpG sties by using the identified optimum method.

3.3 Objectives:

- DNAm profiles were downloaded from an online genomic repository.
- Three statistical methods, namely Spearman's rank correlation, Pearson's correlation, and simple (or univariate) linear regression were implemented to identify AR CpG sites from the downloaded DNAm profiles.
- Each one of the statistical methods was tested using two DNAm measurements, namely Beta and M values.
- The method that identifies the most significant AR CpG sites was selected as the optimum statistical method.
- Saliva DNAm profiles were downloaded from an online genomic repository to identify saliva-specific AR CpG sites using the selected optimum statistical method.

3.4 Materials and methods

All the R codes used in this Chapter can be found in Appendix C1.

3.4.1 Data set

The analyses described in this study were based on a data set downloaded from the National Centre for Biotechnology Information (NCBI) Gene Expression Omnibus (GEO) database. The data set was downloaded using the *GEOquery* package, as described in Section 2.3. The accession number of the data set used was GSE59509 [93], and it consisted of 42 samples assayed on the Illumina HM450K BeadChip array. The samples were derived from 36 males and 6

females with ages ranging from 20 to 59 years old, and came from five different types of body fluids, namely blood, saliva, semen, menstrual blood, and vaginal secretions (Table 3.1).

Table 3.1 Tissue types and age distribution among the 42 samples in the GSE59509 data set

Tissue type	Age range (years)	Mean age (years)	Sample size
Whole blood	24 – 59	40	12
Semen	20 – 59	41	12
Saliva	20 – 59	38	12
Vaginal secretions	21 – 23	22	3
Menstrual blood	21 – 27	24	3
Total			**42**

3.4.2 Processing the Illumina HM450K data set

The pre-processing steps required for this type of data, which include quality assurance measures such as removing replicates with large discrepancies in methylation levels and probes with detection *P*-value >0.05 (as described in Section 2.4.1), had already been carried out by the authors before uploading their data into the online data repository [93]. Further management and processing of the DNAm data was done as described in Section 2.4.1. In particular, before conducting any downstream statistical analyses, the data set was normalised using the BMIQ method, as described in Section 2.4.2, in order to render the two types of probes (Infinium I and II) comparable with each other. After probe QC filtration, as described in Section 2.4.3, the number of CpG sites in the data set was reduced from 485,577 to 310,014 CpG probes.

3.4.3 Singular value decomposition and cluster analysis

Although the authors who uploaded the data set (GSE59509) had already adjusted the Beta values in the data set for batch effects using the *ComBat* package, SVD and cluster analysis were also used here to assess the DNAm pattern in each tissue type, and how the different tissues clustered based on their DNAm profiles [168]. SVD and cluster analysis were carried out using the *svd* and *hclust* functions in R, as described in Section 2.4.4.

3.4.4 Identifying AR CpG sites

The performance of the two correlation tests (Spearman's rank and Pearson's) and the simple linear regression method were evaluated based on their ability to detect AR CpG sites from the 310,014 CpG probes across five tissues (whole blood, semen, saliva, menstrual blood, and vaginal secretions).

3.4.4.1 Spearman's rank correlation

The Spearman's rank correlation coefficient (rho) was calculated between each CpG site in the data set and the chronological age of the donor, using a series of custom scripts written in R software, as described in Section 2.5.1.2. In order to demonstrate the effect of using different DNAm metrics, the Spearman's rank test was first run using Beta values, and second with M values. The cut-off value for selecting AR CpG sites was absolute (abs) rho ≥ 0.6, as recommended by a number of studies in the literature [8,94,145]. In order to detect true positive markers, a more stringent significance level (*P*-value) was used, at which the false discovery rate (FDR) (q-value) is ≤ 0.05. The corrected *P*-value was calculated in R using the *q-value* package, as described in Section 2.5.1.4 [176].

3.4.4.2 Pearson's correlation

As above, the Pearson's correlation test was conducted using a series of custom scripts written in R software, as described in Section 2.5.1.1, and was also

run twice, once with Beta values, and once with M values. The cut-off value for selecting positively and negatively correlated markers was abs r ≥ 0.6. Again, the same stringent significance value with an FDR equal to ≤ 0.05 was used to select the true AR CpG sites. The corrected *P*-value was calculated in R using the *q-value* package, as described in Section 2.5.1.4 [176].

3.4.4.3 Simple linear regression

Finally, simple linear regression was conducted in R software, as described in Section 2.5.1.3, using custom-written R scripts. The chronological ages of the donors of the samples were linearly regressed on the DNAm level of each CpG marker, using Beta and M values separately, and again only CpG makers that passed the stringent FDR ≤0.05 condition were considered as AR CpG markers.

3.5 Applying the identified optimum method on a saliva data set

3.5.1 Data

To assess the standard procedures that were identified for selecting AR CpG markers, a data set consisting of methylation data from 54 saliva samples was retrieved from the NCBI GEO database, as described in Section 2.3. The accession number of this data set was GSE92767, which was used in a study conducted by Hong et al. (2017) as an initial training data set to identify saliva-specific AR CpG markers that were subsequently used for age estimation. They were obtained from 54 males aged from 18 to 73 years and assayed on the HM450K BeadChip array (Figure 3.1).

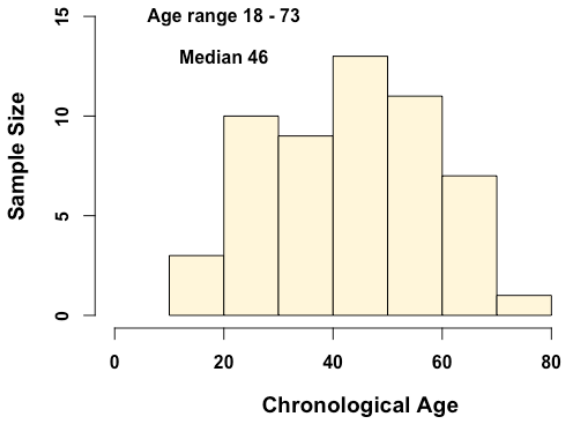

Figure 3.1 Age distribution of the donors of the 54 saliva samples in the training data set (accession number GSE92767).

Low quality probes as well as those with signal intensities less than the mean background for negative control probes (detection *P*-value ≥ 0.05) had already been removed from the data set by the authors before uploading into the online repository [157]. Managing and processing the DNAm data was done as described in Section 2.4. Before conducting any downstream statistical analyses, the data set was normalised using the BMIQ method, as described in Section 2.4.2. Figure 3.2 shows the Beta value distributions from the Infinium I and II probes before and after normalisation. After probe QC filtration, as described in Section 2.4.3, the number of CpG sites in the data set was reduced from 485,577 to 449,042 CpG probes.

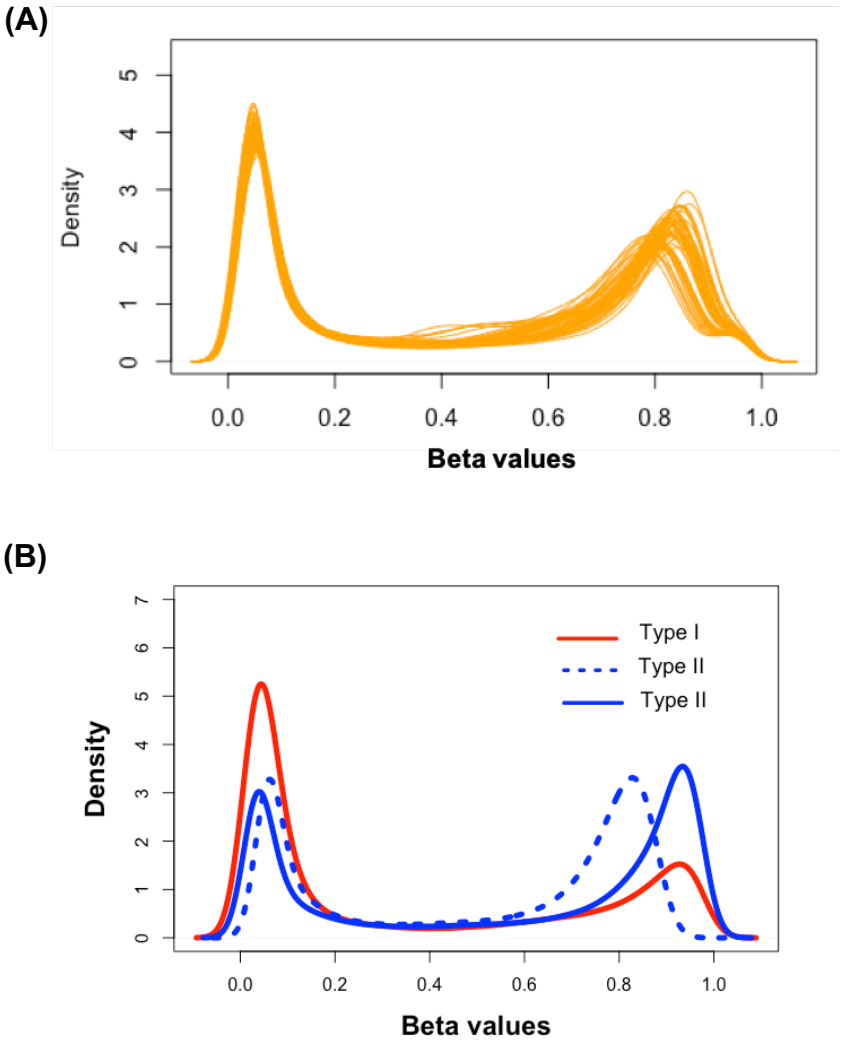

Figure 3.2 (**A**) Density plot of Beta values for 54 saliva samples (accession number GSE92767). The orange lines represent each of the samples in the data set and the height of each line represents the density of the methylation values found in each sample. (**B**) Density plot of the distribution of average Beta values for the two Infinium assay probes (type I and II). The red and blue lines represent type I and II respectively. The blue dotted line represents type II probes before BMIQ normalisation, and the blue solid line represents these probes after normalisation.

3.5.2 Identifying AR CpG sites and building the saliva model

After the optimum statistical method for identifying AR CpG sites was selected, it was implemented on the 54 saliva DNAm profiles using a series of custom scripts written in R software. After this, the selected AR CpG sites were input in stepwise regression analysis. However, instead of using Beta values in this analysis, M values were used in order to satisfy the assumptions of the stepwise regression. For the final step, the selected subset of CpG markers were used to build a multivariate model for saliva specific age prediction, as described in Section 2.5.3.1. This model will be referred to as the saliva-specific HM450K model throughout this thesis.

3.5.3 *In silico* validation of the saliva HM450K model

The performance of the constructed saliva HM450K model was assessed *in silico* on an independent testing data set of saliva samples retrieved from the NCBI GEO database. The accession number of the data set is GSE99029 and it consisted of 57 saliva samples from donors (22 male, 35 female) aged from 21 to 91, with a median age of 63 years (Figure 3.3). These samples were obtained from the study by Gopalan et al. [180] who collected them from African hunter-gatherer individuals from a population known as the Khomani San living in the South African Kalahari Desert. This population is considered to be one of the most genetically diverse populations in the world, as well as due to other population differences in terms of nutritional subsistence, ecological environment (semi-desert), and physical activity, compared to cosmopolitan populations. The aim of their study was to explore epigenetic aging across a wide range of human diversity, from populations living in distinct ecological systems.

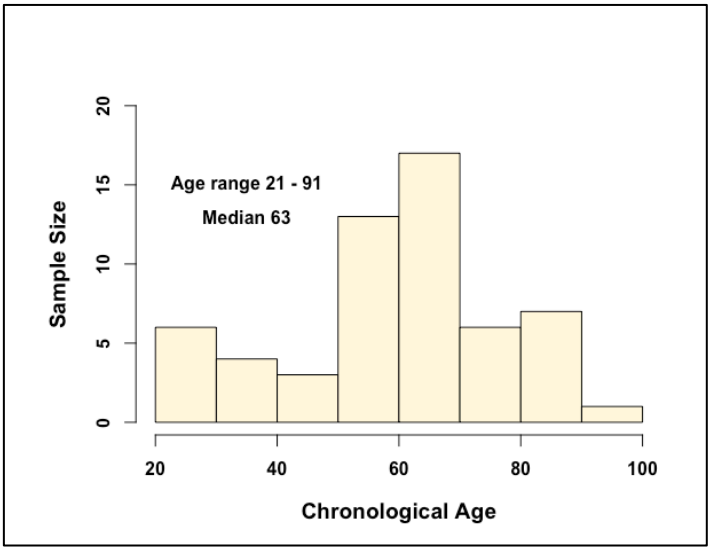

Figure 3.3 Age distribution of the donors of the 57 saliva samples in the testing data set (GSE99029).

The DNAm levels for the samples were assessed for the presence of any outliers (Figure 3.4). The data set was processed and normalised, as described in Section 2.4. Predicting the chronological ages of donors based on the saliva-specific HM450K model was conducted as described in Section 2.5.4.

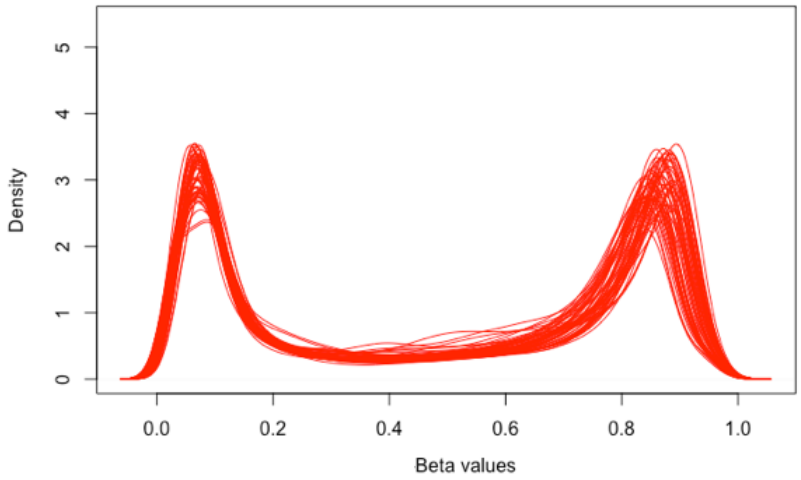

Figure 3.4 Density plot of Beta values for 57 saliva samples (GSE99029). The red lines represent each of the samples in the data set and the height of each line represents the density of the methylation values found in each sample.

3.5.4 Comparing the saliva HM450K model with Hong et al.'s model

The saliva HM450K model that was constructed using the identified optimum method was compared with another saliva-specific model that was created by Hong et al. (2017). Their model was created using the same training data set that was used to build the saliva-specific HM450K model in this study. Thus, comparing the age prediction accuracy between them would give an indication of how the identified optimum methods performed. The AR CpG markers in the Hong et al. study were identified using simple linear regression based on Beta values. They identified 61 CpG markers, four of which (cg07547549, cg14361627, cg08928145, and cg19671120, found in the *SLC12A5*, *KLF14*, *TSSK6*, and *CNGA3* genes, respectively) were selected by stepwise regression using Beta values as candidate markers for building an age prediction model. They also added three more CpG markers (cg00481951, cg12757011, and cg18384097, found in *SST*, *TBR1*, and *PTPN7*), building an

age prediction model with a total of seven CpG markers. This model was tested in their study on an independent data set of 113 saliva samples that were assayed using SNaPshot mini sequencing and predicted chronological age with a high accuracy of 3.15 years (mean absolute deviation (MAD) from the chronological age).

In this study, the Hong et al. model was built using their seven AR CpG markers (Table 3.2) and trained on the HM450K DNAm profiles of the 54 saliva samples, using the same multivariate linear regression system that was implemented in their study. Building Hong et al.'s model was conducted as described in Section 2.5.3. In order to compare the age prediction accuracy of the Hong et al. model with the model identified in this study, the Hong et al. model was also validated *in silico* on the 57 saliva samples (GSE99029) from the Khomani San population, and the MAD value calculated as described in Section 2.5.4.

Table 3.2 The seven AR CpG markers identified by Hong et al. (2017) and included in their saliva-specific age-prediction model. Genomic locations are for the human genome assembly GRCh37, also known as hg19.

Probe ID	Gene symbol	Genomic location
cg18384097	*PTPN7*	chr1:202129566
cg00481951	*SST*	chr3:187387650
cg19671120	*CNGA3*	chr2:98962974
cg14361627	*KLF14*	chr7:130419116
cg08928145	*TSSK6*	chr19:19625364
cg12757011	*TBR1*	chr2:162281111
cg07547549	*SLC12A5*	chr20:44658225

3.6 Results

3.6.1 Illumina HM450K data processing

The data set initially consisted of 42 samples coming from five different tissues. However, when the DNAm profiles were evaluated using density plots, one saliva sample (sample ID GSM1438496) showed an abnormal Beta value distribution, illustrated by the red line in Figure 3.5. This sample was therefore removed from the data set.

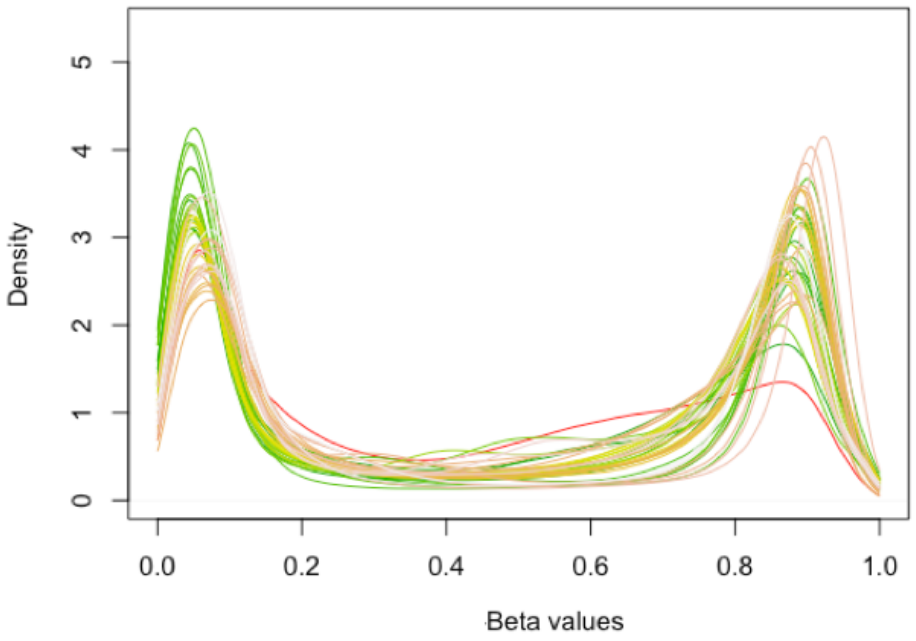

Figure 3.5 Density plot of Beta values for 42 samples in the GSE59509 data set. The coloured lines represent the samples in the data set and the height of each line represents the frequency of the methylation values found in each sample.

3.6.1.1 Normalisation

Before conducting any downstream statistical analyses, the data set was checked for normalisation. This was done by plotting a density plot of the Infinium

I and II probes in two different colours, and checking they aligned with each other. As can be seen from the density plot (Figure 3.6A), the raw Beta value distributions for the Infinium I and II probe data sets were not aligned, which means they were not normalised. The data were therefore normalised to render the Infinium II probes comparable with the Infinium I probes [108]. After performing the normalisation, the Beta value distributions for the two probe types were aligned (Figure 3.6B).

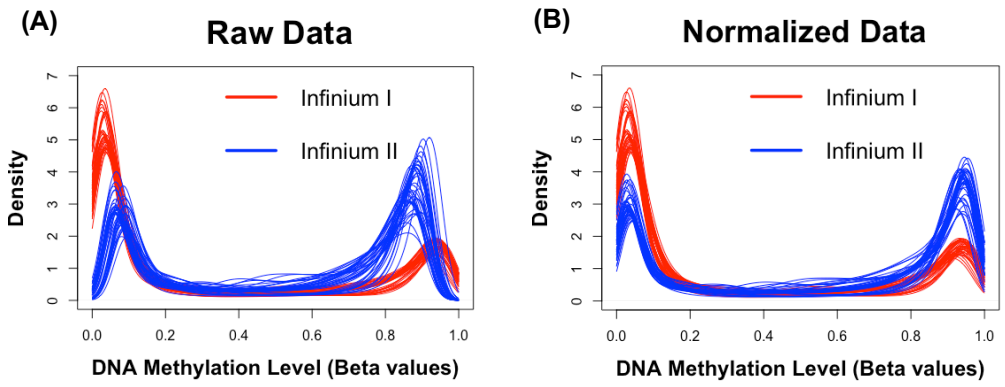

Figure 3.6 Density plots of the Beta value distributions in the GSE59509 data set for the two Infinium probe types (**A**) before BMIQ normalisation and (**B**) after normalisation.

For the purpose of exploring global DNAm patterns across the five forensically relevant tissues present in the data set, a box-plot was constructed. Figure 3.7 illustrates that the median Beta values in each tissue type are not similar, which can be explained by the fact that each tissue has its own distinctive DNAm profile. Figure 3.7 also shows that some tissues are epigenetically close to each other, whereas others are not. For instance, menstrual blood and vaginal secretions show median methylation values that are close to saliva and whole blood samples. However, semen samples exhibit very low median Beta values compared to the other tissues. As discussed in Section 1.2.3, this is likely to be

due to epigenetic reprogramming in the germ line, which occurs during spermatogenesis [18,19].

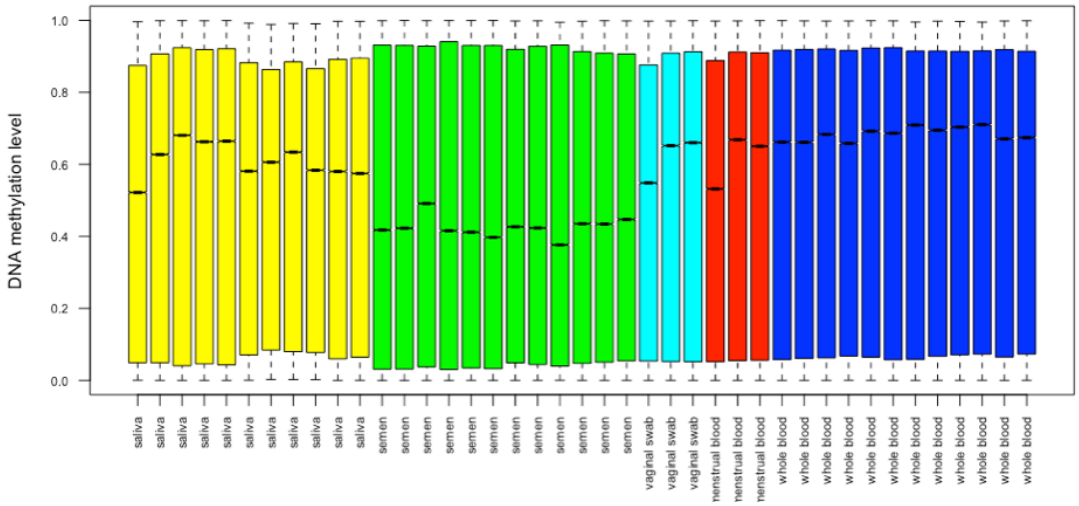

Figure 3.7 Box-plot illustrating the distribution of Beta values across the five tissues. The bottom and top of the box represent the 25th and 75th percentile (the lower and upper quartiles, respectively), and the band near the middle of the box represents the 50th percentile (the median). The lower and top whiskers represent the minimum and maximum values in the sample, respectively.

3.6.1.2 SVD and cluster analysis

The main reason for conducting SVD and cluster analyses was to demonstrate how the different tissues would cluster in relation to overall DNAm pattern. The SVD analysis showed that the samples cluster in a distribution that is based on tissue type. As shown in Figure 3.8**,** semen has formed a distinctive cluster, which reflects its distinctive distribution of Beta values, which also can be seen in the previous box plot (Figure 3.7). Saliva, menstrual blood, and vaginal secretion samples are scattered along the axes of singular values 1 and 2. This may be explained by the fact that the oral and vaginal mucosa contain very similar cell types, making their epigenetic patterns similar.

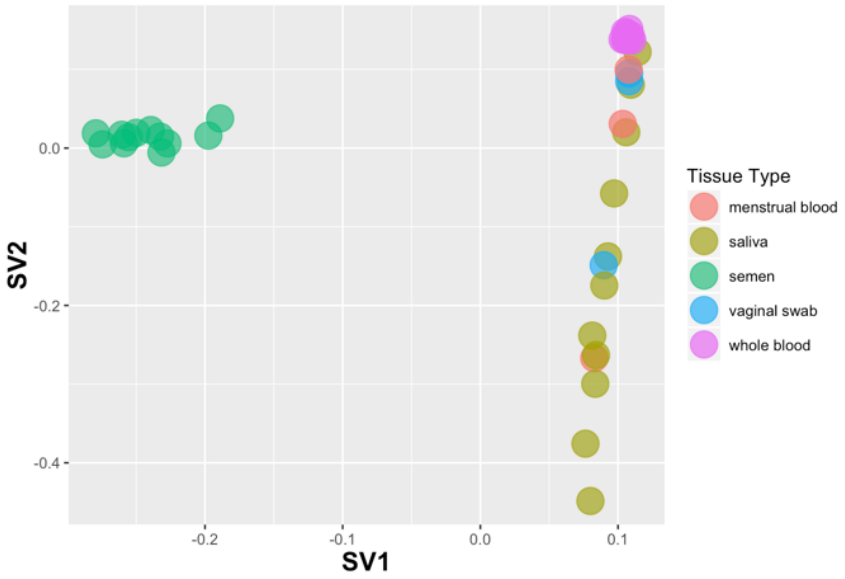

Figure 3.8 SVD plot for all 41 samples using all 310,014 CpG sites. The colours on the plots represent different tissue types.

As expected, similar results were obtained using hierarchical cluster analysis. However, the dendrogram (Figure 3.9) included an additional observation that saliva, menstrual blood and vaginal secretion samples are divided into two groups, one of which is close to the whole blood samples, the other of which is an independent group. If the relevant samples were analysed in the same batch along with the whole blood samples, this close relationship could be attributed to the batch effect. However, information provided by the author of the original study with the downloaded data indicates that the batch effect was removed before making the data available [93].

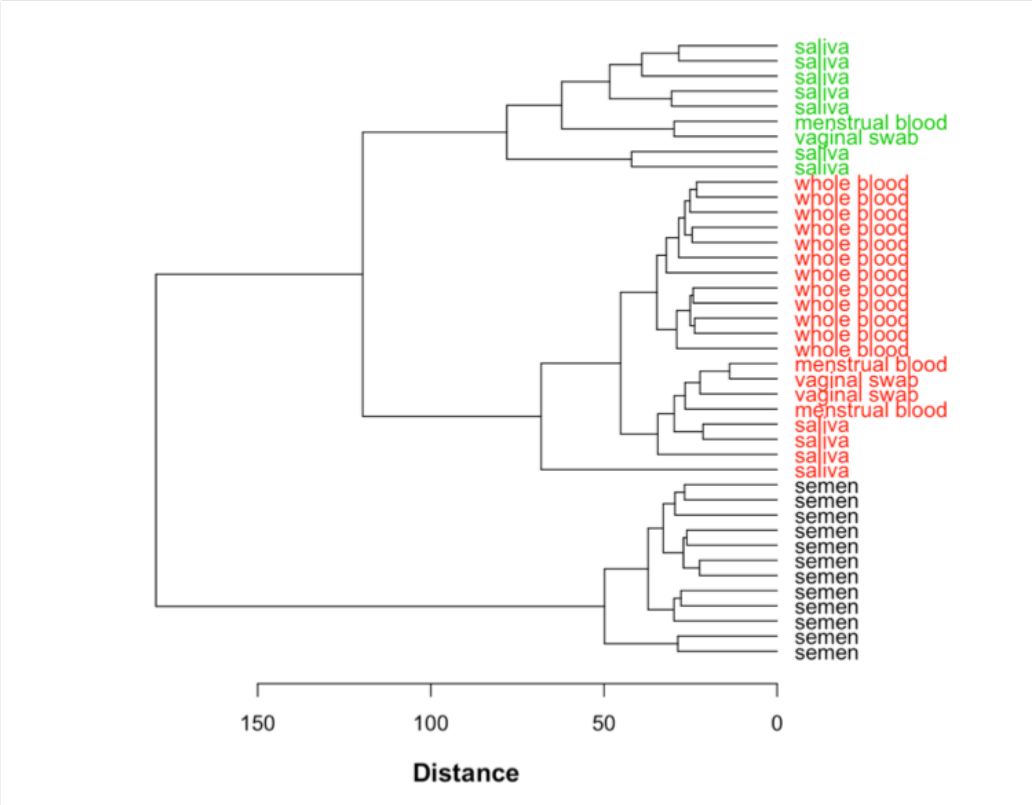

Figure 3.9 Dendrogram showing the hierarchical clustering of 41 samples using 310,014 CpG sites. The distance between tissues is based on the Euclidean distance of their DNAm values.

3.6.2 Identifying the optimum method for identifying AR CpG sites

The purpose of this study was to determine a standard test among those that are frequently used in AR studies, which would be optimum for identifying AR CpG sites to be used for building age-prediction models. Two correlation tests were investigated (Spearman's rank and Pearson's correlation tests) and one simple linear regression. Before conducting the analyses, pre-processing steps were carried out, which resulted in removing one sample outlier, and removing X/Y chromosome CpG probes, probes containing SNPs, and cross-reactive probes. After this probe imputation, the CpG sites in the data set were reduced

from 485,577 to 310,014 CpG probes in samples from five types of tissue: blood, saliva, semen, menstrual blood, and vaginal secretions.

3.6.2.1 Spearman's rank correlation

Spearman's rank correlation between the chronological ages of the donors and each of the two DNAm measurements (Beta and M values) gave the same results. The number of CpG sites that passed the abs rho ≥ 0.6 criteria was 867, and among these 747 were positively correlated (hypermethylated) and 120 were negatively correlated (hypomethylated) with age. Furthermore, after further filtration with FDR ≤ 0.05, the two DNAm measurements produced the same 31 AR CpG sites with significant age-association. Of these 31 CpG sites, 27 were positively correlated, and four were negatively correlated with age. These 31 sites were labelled as AR CpG markers. Table 3.3 shows the number of CpG sites detected under different FDR values.

Table 3.3 Cumulative number of CpG sites from Spearman's rank correlation test based on different FDR values. The same results were obtained for Beta values and M values.

FDR value	< 1e-04	< 0.001	< 0.01	< 0.025	< 0.05	< 0.1	< 1
Number of AR CpG sites	0	0	6	8	31	114	310,014

3.6.2.2 Pearson's correlation

The Pearson's correlation test between chronological age and Beta values resulted in 339 AR CpG sites with correlation coefficient abs r ≥ 0.6, which is substantially fewer than the number selected using the Spearman's rank correlation test, by 544 CpG sites. Among them, there were 278 CpG sites that were positively correlated (hypermethylated), and 61 CpG sites that were negatively correlated (hypomethylated) with age. However, after applying the FDR cut-off value (≤ 0.05), only one CpG site passed this condition (Table 3.4).

The Illumina ID of this AR CpG site is cg16875637. In contrast, the results of the Pearson's correlation test based on M values showed 867 CpG sites with abs r ≥0.6. These markers were exactly the same markers that were detected by the Spearman's rank correlation test. However, the Pearson's correlation coefficient value and the *P*-value of each CpG site were different. The results showed that the Pearson's correlation coefficients with M values were smaller than the Spearman's rank correlation coefficient values for each site. Thus, the *P*-values of the Pearson's test were larger than those generated by the Spearman's rank correlation. As a result of this, only four CpG sites passed the FDR condition (≤ 0.05) (Table 3.5). The Illumina IDs for these AR CpG sites were cg16875637, cg22971191, cg23118721, and cg27571590. The change in methylation level at each on these AR CpG sites with chronological age is illustrated in Figure 3.10.

Table 3.4 Cumulative number of CpG sites from Pearson's correlation test based on Beta values with different FDR values.

FDR value	**< 1e-04**	**< 0.001**	**< 0.01**	**< 0.025**	**< 0.05**	**< 0.1**	**< 1**
No. of AR CpG site	0	0	0	1	1	1	310,014

Table 3.5 Cumulative number of CpG sites from Pearson's correlation test based on M values with different FDR values.

FDR value	**< 1e-04**	**< 0.001**	**< 0.01**	**< 0.025**	**< 0.05**	**< 0.1**	**< 1**
No. of AR CpG sites	0	0	1	3	4	7	310,014

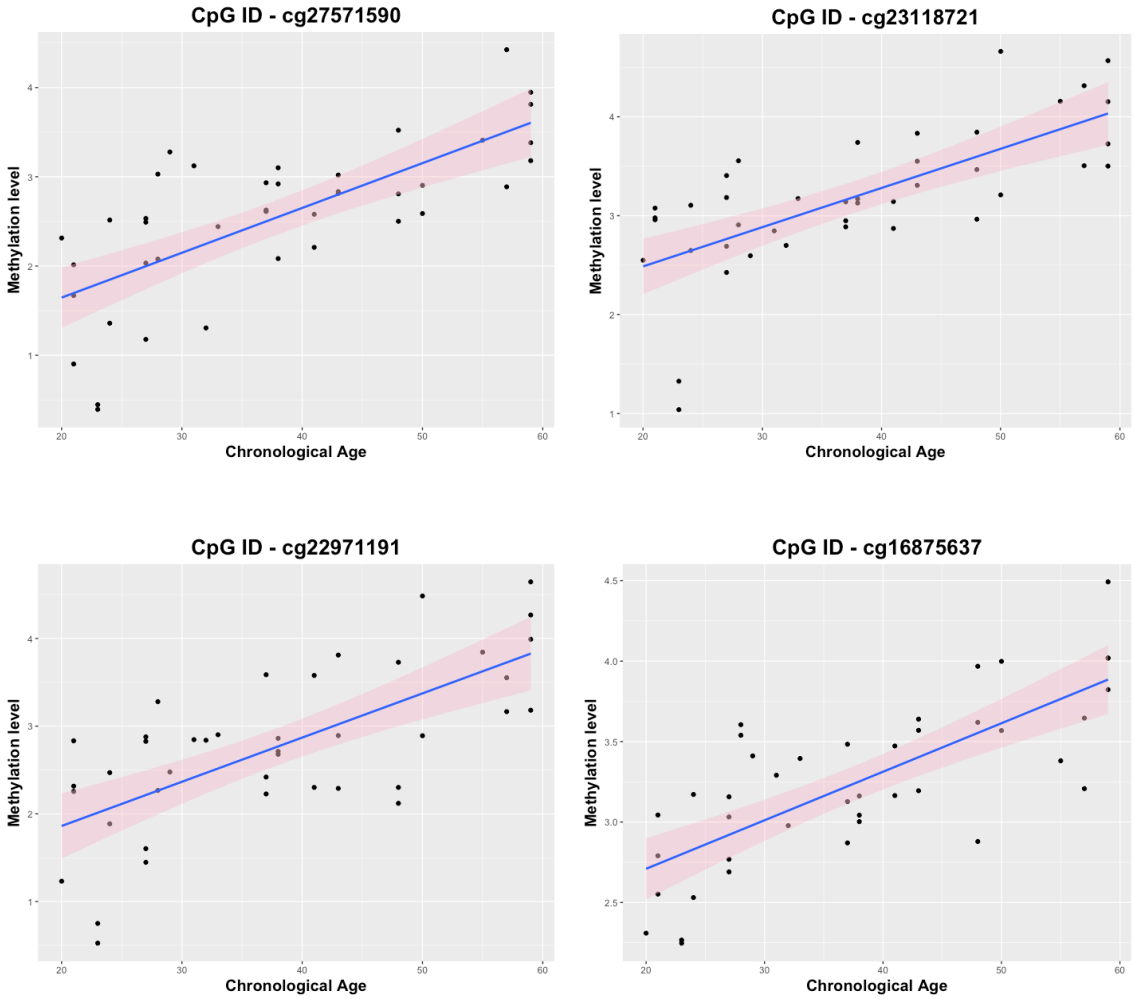

Figure 3.10 Association between DNAm level and chronological age across five tissues for four individual AR CpG sites identified by Pearson's correlation test based on M values.

3.6.2.3 Simple linear regression

Simple linear regression analysis was used to reduce the number of CpG sites by removing those with low age prediction accuracy across the tested tissues. The main theoretical difference between correlation tests and regression tests is that correlation tests describe the relationship between two variables,

whereas regression tests describe how well one variable can predict the value of the other variable. From the results (Table 3.6), it can be seen that based on Beta values, and after FDR (≤ 0.05) filtration, only one CpG site met this condition. This CpG site was the same site (cg16875637) that was detected using the Pearson's correlation test with Beta values. Similarly, the analysis was carried out using M values and identified the same four CpG sites that were identified using Pearson's correlation test with M values (cg16875637, cg22971191, cg23118721, and cg27571590) (Table 3.7). Figure 3.11 shows a Venn diagram illustrating the degree of overlap between CpG sites for the different methods used for detecting AR CpG sites.

Table 3.6 Cumulative number of CpG sites from simple linear regression based on Beta values with different cut-off FDR values.

FDR value	**< 1e-04**	**< 0.001**	**< 0.01**	**< 0.025**	**< 0.05**	**< 0.1**	**< 1**
No. of AR CpG sites	0	0	0	1	1	1	310,014

Table 3.7 Cumulative number of CpG sites from simple linear regression based on M values with different cut-off FDR values.

FDR value	**< 1e-04**	**< 0.001**	**< 0.01**	**< 0.025**	**< 0.05**	**< 0.1**	**< 1**
No. of AR CpG sites	0	0	1	3	4	7	310,014

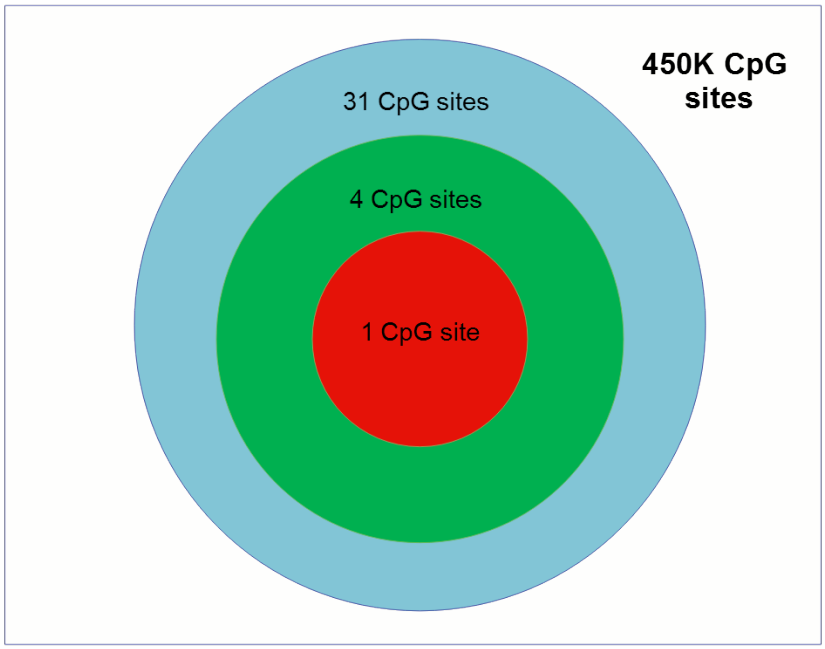

Figure 3.11 Venn diagram showing the degree of overlap between the outcomes of different methods used to identify AR CpG sites. The background represents the probes on the Illumina HM450K BeadChip, the blue colour represents the outcomes of the Spearman's rank correlation test, green represents the outcomes of both Pearson's correlation test and simple linear regression using M values, and red represents the outcomes of both Pearson's correlation and simple linear regression using Beta values.

3.6.3 Identifying saliva specific AR CpG sites using the selected optimum methods

From the results described above, it was found that using Spearman's rank correlation test detects more significant AR CpG sites compared to other methods such as Pearson's correlation and simple linear regression methods. Therefore, the Spearman's rank correlation test was used to identify saliva specific AR CpG sites in the saliva training data set (GSE92767). The Spearman's rho correlation coefficient was calculated between each CpG site in the saliva training data set and the chronological age of the donors, using a series of custom scripts written in R software. The criteria for selecting the AR CpG sites were: Spearman's abs

rho ≥ 0.6, a difference in DNAm levels between samples from younger and older aged donors of > 0.1 (based on Beta values, as recommended by Hong et al (2017)), and FDR ≤ 0.05. The total number of CpG sites that passed these criteria was 988. As shown in Figure 3.12, the overall molecular effect of aging on the methylome is positive (hypermethylation). This is demonstrated by the 'hump' at a rho value of around 0.4, which represents the density of probes that are positively correlated with age. The number of age-associated markers retrieved from the correlation test is extremely large, which cannot be handled by the stepwise regression test in R software. Therefore, the AR CpG sites were further filtered by lowering the FDR value to ≤1E-7, which resulted in 49 candidate AR CpG markers (**Error! Reference source not found.**). Among these 49 candidate markers, three (cg00481951, cg14361627, and cg07547549) were identified and included in the Hong et al. model [157].

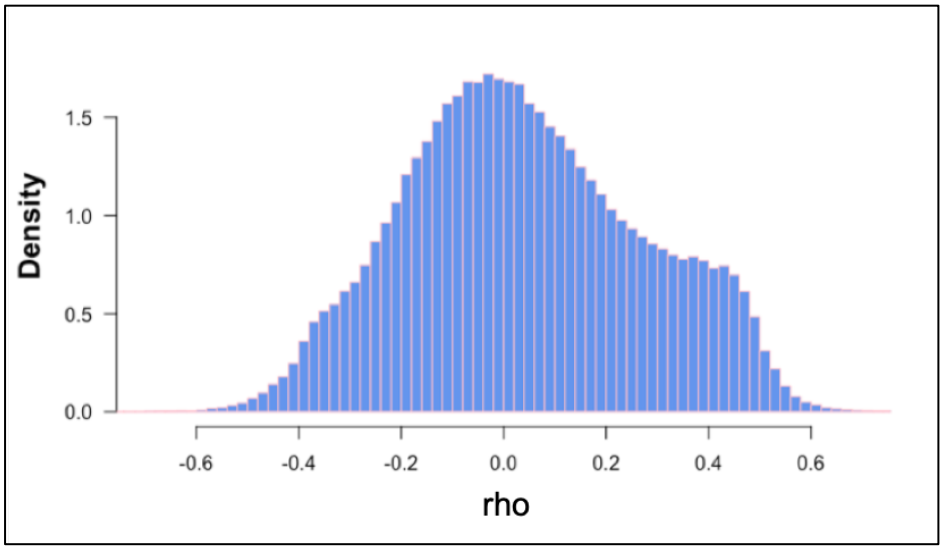

Figure 3.12 Histogram of Spearman's rank correlation coefficients (rho) obtained from the correlation test between DNAm level at each of 432,215 CpG sites (Beta value) and the chronological ages of the donors of 54 saliva samples.

Table 3.8 49 CpG sites that were significantly associated with age at FDR 0.00001% in the 54 saliva samples obtained from accession number GSE92767.

Probe Illumina ID	Chromosome	Co-ordinate	Gene	*P*-value	Spearman's coefficient
cg22736354	6	18122719	*NHLRC1*	2.30E-20	0.900
cg00481951	3	187387650	*SST*	3.04E-20	0.899
cg16867657	6	11044877	*ELOVL2*	3.24E-20	0.898
cg00094518	7	130418549	*KLF14*	6.73E-18	0.874
cg06493994	6	25652602	*SCGN*	7.68E-18	0.873
cg11084334	3	9594264	*LHFPL4*	1.14E-17	0.871
cg04875128	15	31775895	*OTUD7A*	8.74E-17	0.860
cg06782035	5	16179135	*unknown*	1.04E-16	0.859
cg20591472	1	110008990	*SYPL2*	2.37E-16	0.854
cg14361627	7	130419116	*KLF14*	4.73E-16	0.849
cg01763090	15	31775406	*OTUD7A*	7.16E-15	0.831
cg06639320	2	106015739	*FHL2*	7.68E-15	0.831
cg10804656	10	22623460	*unknown*	1.09E-14	0.828
cg08160331	11	75140865	*KLHL35*	1.71E-14	0.825
cg08097417	7	130419133	*KLF14*	2.02E-14	0.824
cg14556683	19	15342982	*EPHX3*	2.96E-14	0.821
cg00439658	17	72848669	*GRIN2C*	3.48E-14	0.820
cg13327545	10	22623548	*unknown*	4.06E-14	0.818
cg19560758	1	8086721	*ERRFI1*	4.71E-14	0.817
cg18473521	12	54448265	*HOXC4*	4.73E-14	0.817
cg07547549	20	44658225	*SLC12A5*	6.26E-14	0.815
cg25410668	1	28241577	*RPA2*	1.04E-13	0.811
cg14131273	9	135464095	*BARHL1*	1.13E-13	0.810
cg14674720	2	219827930	*unknown*	2.15E-13	0.805
cg07553761	3	160167977	*TRIM59*	2.83E-13	0.803
cg01844642	3	51989764	*GPR62*	4.02E-13	0.800
cg19802138	13	112722719	*SOX1*	4.26E-13	0.799
cg07365960	17	72848535	*GRIN2C*	4.80E-13	0.798
cg23606718	2	131513927	*FAM123C*	4.82E-13	0.798
cg08885800	1	201084119	*unknown*	9.72E-13	0.792

cg22454769	2	106015767	*FHL2*	1.45E-12	0.789
cg13954457	5	167956819	*FBLL1*	1.97E-12	0.786
cg20049415	20	21377671	*NKX2-4*	2.69E-12	0.783
cg05168491	14	38080446	*unknown*	2.71E-12	0.783
cg05213896	19	50393653	*IL4I1*	3.27E-12	0.781
cg04865692	19	50831762	*KCNC3*	3.99E-12	0.779
cg25478614	3	187387866	*SST*	4.34E-12	0.778
cg11705975	10	120354248	*PRLHR*	5.53E-12	0.776
cg24079702	2	106015771	*FHL2*	5.78E-12	0.775
cg23995914	4	10459228	*ZNF518B*	6.29E-12	0.775
cg24853724	7	28997403	*TRIL*	6.34E-12	0.775
cg25124276	10	25464008	*GPR158*	1.01E-11	0.770
cg06279276	16	67184164	*B3GNT9*	1.82E-11	0.764
cg23538901	15	46006849	*unknown*	3.16E-11	0.758
cg23142799	13	26625089	*SHISA2*	7.05E-11	0.749
cg18064714	7	20824556	*SP8*	1.04E-10	0.745
cg05694021	12	19699504	*unknown*	1.01E-11	-0.770
cg00573770	2	145278485	*ZEB2*	5.95E-13	-0.796
cg10501210	1	207997020	*unknown*	4.19E-17	-0.864

3.6.4 Building the saliva HM450K model

Following identification of the 49 candidate AR CpG markers described above, stepwise regression was implemented to select the best subset of these markers to be used in the age-prediction model. Before conducting the stepwise regression, DNAm levels were converted to M values to satisfy the assumptions of the regression test. The stepwise regression test yielded a set of nine CpG markers that had the lowest BIC value, which is reasonable in terms of the number markers that can be used for forensic analysis (Figure 3.13). The heat map in Figure 3.14 shows a consistent change (either hyper- or hypomethylation) in DNAm level at the nine CpG sites across chronological ages, without the

presence of any samples with irregular methylation patterns. Table 3.8 shows the identity of each site and the relationship between chronological age and methylation level for each of these sites is shown in Figure 3.15. Instead of using Beta values, as in the Hong et al. (2017) study, M values were used to build a multivariate linear model from these nine CpG sites. The constructed saliva-specific HM450K model explained 97.5% of the total variation in DNAm levels in the samples of 54 males in the training data set (GSE92767), with a mean absolute deviation (MAD) from the chronological age of 1.8 years (see Table 3.9 and Figure 3.16).

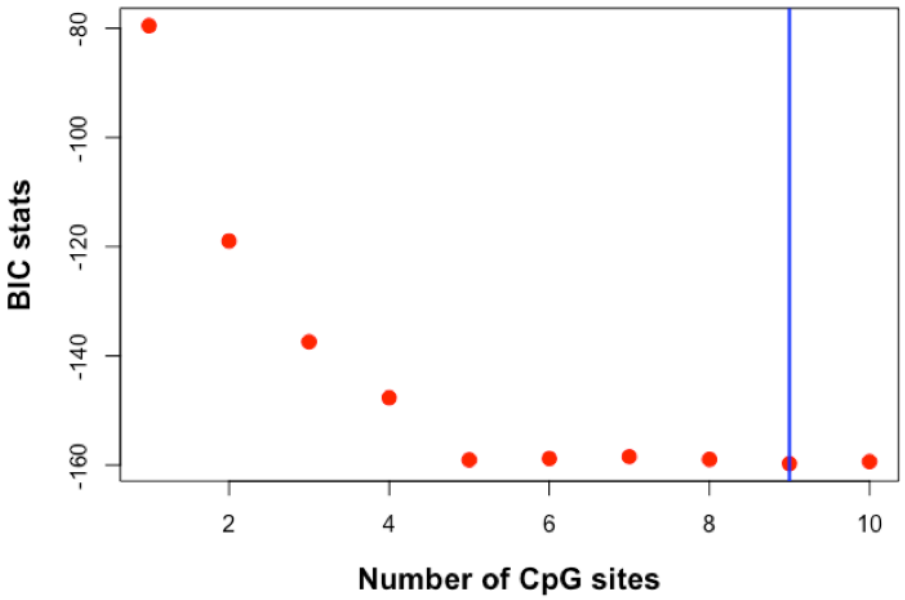

Figure 3.13 Bayesian Information Criterion (BIC) as a function of the number of markers, showing that a model with nine CpG sites has the lowest BIC value and thus the best predictive accuracy.

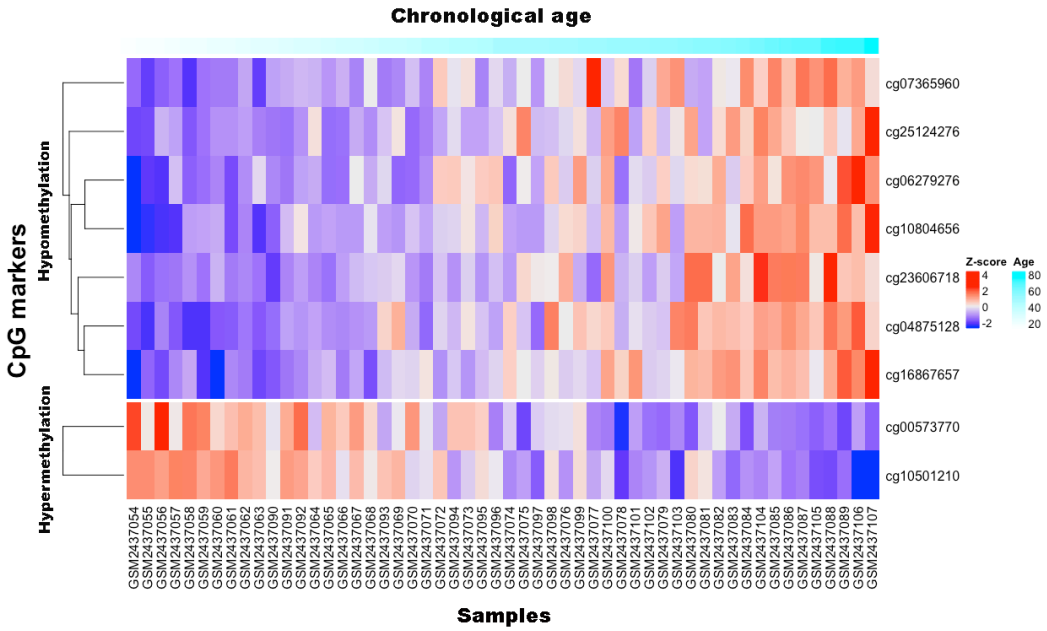

Figure 3.14 Heat map illustrating methylation levels at the nine AR CpG markers selected by stepwise regression, in samples ordered by chronological age. The methylation level is indicated by the Z-score, where red indicates a site is hypermethylated and blue is hypomethylated. Hierarchical clustering of the CpG markers is presented on the left-hand side of the heat map.

Table 3.8 Identity of the nine CpG markers selected by stepwise regression. The R^2 and P-values are from univariate linear regression analysis between each CpG site and chronological ages in the training data set.

Probe ID	Chromosome	Gene	R^2	*P*-value
cg16867657	6	*ELOVL2*	0.77	2.7×10^{-18}
cg10501210	1	*Unknown*	0.77	2.4×10^{-18}
cg10804656	10	*Unknown*	0.72	4.7×10^{-16}
cg04875128	15	*OTUD7A*	0.71	1.1×10^{-15}
cg06279276	16	*B3GNT9*	0.63	9.8×10^{-13}
cg00573770	2	*ZEB2*	0.60	5.6×10^{-12}
cg07365960	17	*GRIN2C*	0.60	7.3×10^{-12}
cg23606718	2	*FAM123C*	0.59	1.1×10^{-11}
cg25124276	10	*GPR158*	0.59	1×10^{-11}

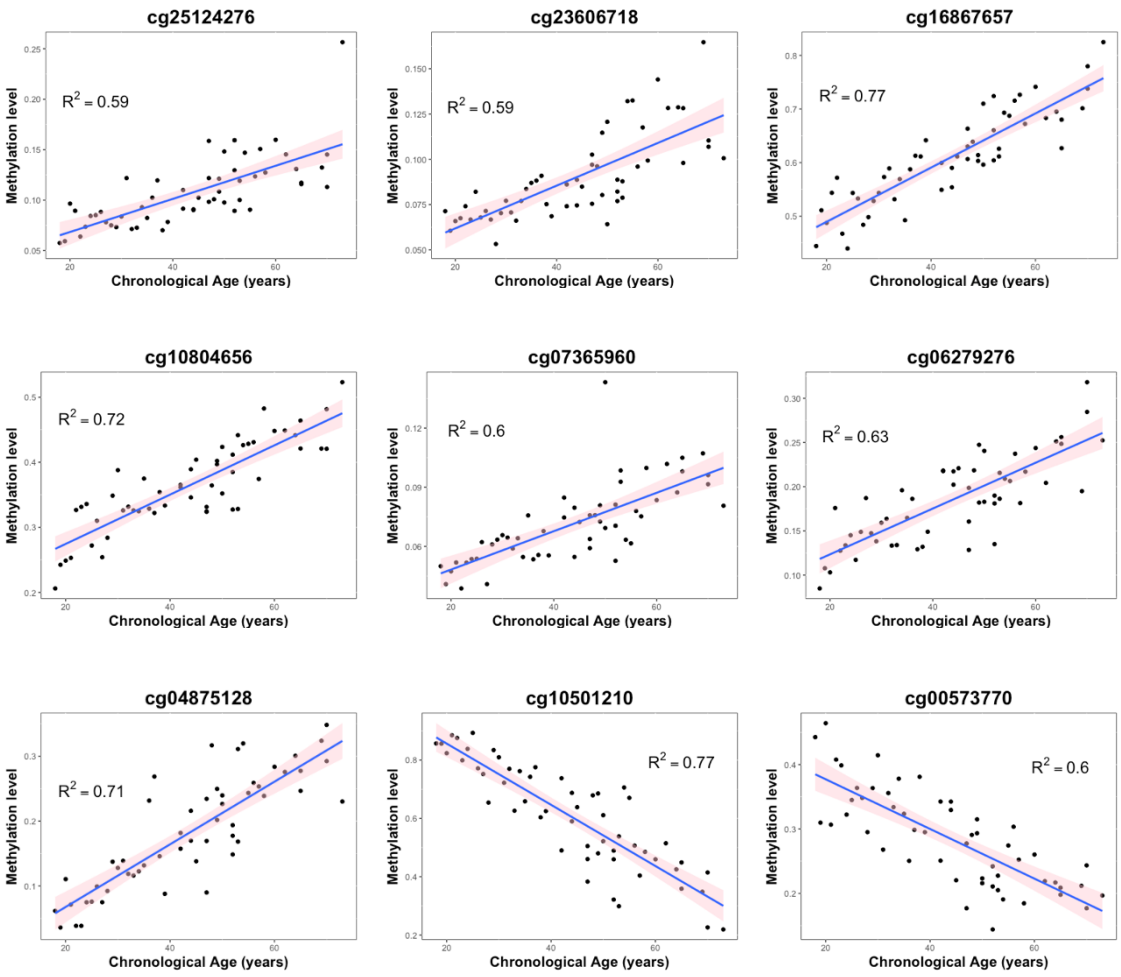

Figure 3.15 Simple linear regression analysis between DNAm level at the nine CpG markers obtained from 54 saliva samples (GSE92767) assayed on the Infinium HM450K.

Table 3.9 Multivariate linear regression statistics for the age-prediction model containing nine AR CpG sites.

CpG site ID	*P*-value	R^2	MAD
(Intercept)	1.2×10^{-10}	0.975	1.8
cg00573770	0.011		
cg04875128	0.021		
cg06279276	0.021		
cg07365960	0.049		
cg10501210	9.1×10^{-9}		
cg10804656	0.02		
cg16867657	5.3×10^{-5}		
cg23606718	6.9×10^{-4}		
cg25124276	0.05		

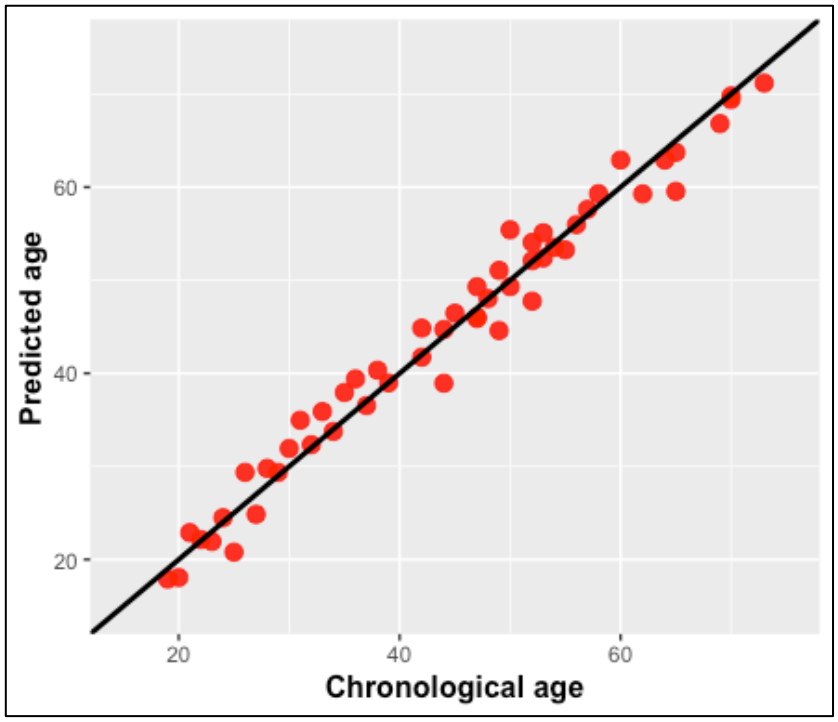

Figure 3.16 Chronological age against predicted age obtained from the multivariate linear regression model, based on the training data set.

3.6.5 *In silico* validation of the saliva HM450K model

The constructed saliva-specific HM450K model was tested on an independent data set consisting of 57 saliva samples from donors (22 males, 35 females) aged from 21 to 91, collected from the Khomani San population, from the South African Kalahari Desert. The Khomani San individuals are genetically diverse and living in a distinctive ecological system compared to individuals coming from cosmopolitan populations. One of the samples in the data set (GSM2630630) had a missing DNAm value for one of the AR CpG markers (cg06279276). Due to the fact that the prediction analysis could not be carried out with a missing marker, the sample was removed from the testing data set. The prediction accuracy of the saliva-specific HM450K model was 5.1 years (MAD) (Figure 3.17). Due to the fact that the training data set (GSE92767) contains only male individuals, and to avoid sex bias in the prediction of age, male and female samples in the testing data set were separated and their MAD values were assessed separately, to see if the difference between them was significant. The t-test showed that there was a non-significant (*P*-value = 0.2) difference in the prediction accuracy for males (MAD = 4.20 years, r = 0.93) compared to females (MAD = 5.58, r = 0.95).

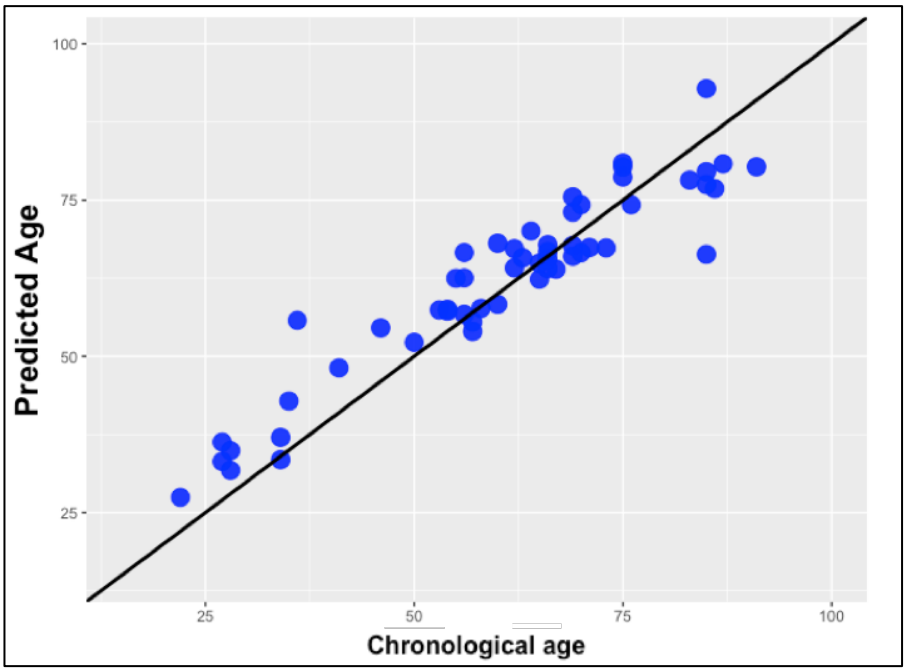

Figure 3.17 Chronological age against predicted age obtained from the multivariate linear regression model, based on the testing data set.

3.6.6 Comparing the saliva HM450K model with the Hong et al.'s model

The performance of the model containing nine CpG markers (the saliva-specific HM450K model) built in this study was compared with the model containing seven CpG markers built by Hong et al. (2017). Based on the Khomani San data set, the Hong et al. model predicted age with an accuracy equal to 8.3 years (MAD) (Figure 3.18), which is more than the model described here by 3.2 years. This difference is statistically significant, as confirmed by an analysis of variance (ANOVA) test, which gave a *P*-value equal to 2.2×10^{-16}.

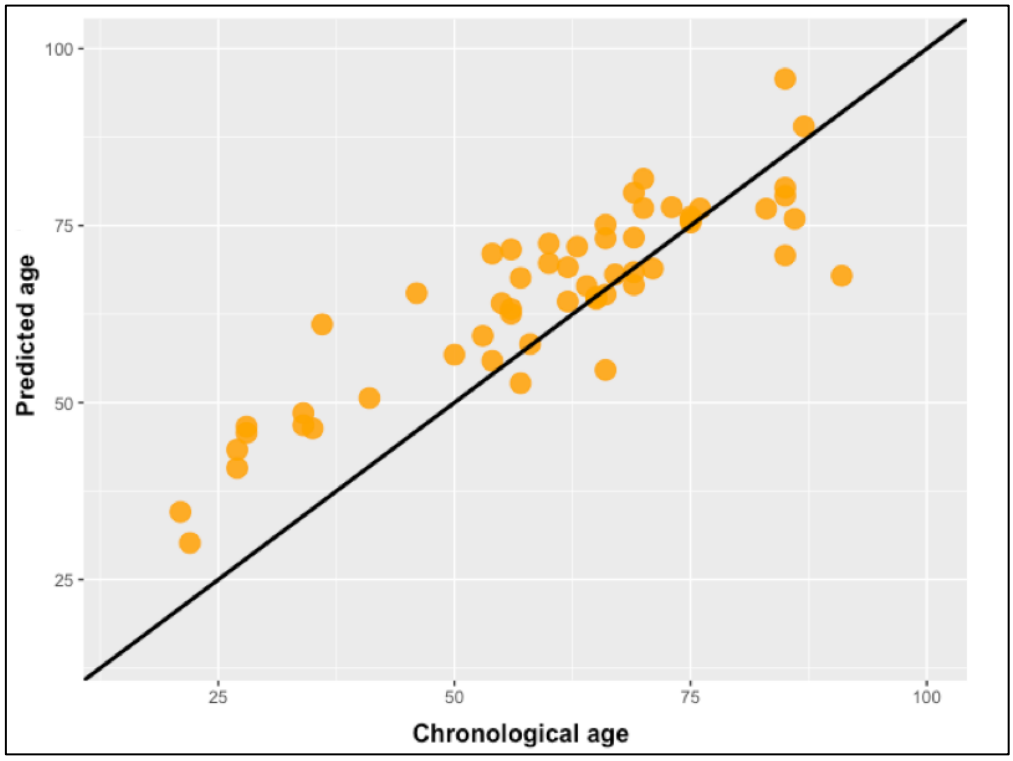

Figure 3.18 Chronological age against predicted age obtained from Hong at. al.'s model based on the testing data set. The Pearson's correlation coefficient (r) between predicted age and chronological age is 0.88.

3.7 Discussion

In previous forensic age estimation studies using AR CpG sites, there has been no consensus on the type of statistical method that should be used to identify AR markers, especially from high dimensional data containing hundreds of thousands of CpG sites. This has resulted in different findings in studies using the same types of tissues and the same types of genome-wide platforms. These differences in findings are likely to be at least in part due to the different methods applied, rather than biological differences [144,147]. A considerable number of papers in the field have carried out their statistical analyses assuming a linear association between age and methylation level, despite the fact that there have been no controlled studies to support that assumption [92,144,179]. The relationship between DNAm level and chronological age does not necessarily have to be linear, because the rate at which the process of methylation or demethylation proceeds at AR loci may not change constantly with age. Two separate studies, conducted by Horvath (2013) and Xu et al. (2015), have touched upon this matter. The former described the rate of change in DNAm level at the 353 CpG markers they identified across tissues as taking the form of a logarithmic relationship from childhood until adulthood and then changing to a linear relationship later in life [106]. The latter study also highlighted that a linear regression analysis is too simple to explain the complicated relationship between DNAm and chronological age [8]. Therefore, using the appropriate statistical method to measure the association between DNAm and chronological age requires a full understanding of their true relationship.

The findings of this study showed that, despite applying the same stringent conditions, Spearman's rank correlation test identified significant AR CpG sites than both Pearson's correlation test and simple linear regression (Figure 3.19). This suggests that DNAm level increases monotonically with age, based on the

fact that the Spearman's rank correlation algorithm measures the monotonic relationship between two continuous or ordinal variables, rather than a linear relationship. In contrast, the Pearson's correlation and simple linear regression tests measure linear associations between variables. Using any algorithm that tries to detect a linear relationship where one does not exist will therefore result in the discarding of a significant number of candidate markers that could have been detected by nonlinear correlation or nonlinear regression tests. However, the majority of studies in the literature have applied Pearson's correlation test or simple linear regression on their data instead of using the Spearman's rank correlation test. For example, Koch and Wagner (2011) carried out a study that aimed to identify AR CpG markers across tissues using five data sets from dermis, epidermis, cervical smear cells, T-cells and monocytes, which had been assayed on the HM27K platform. The experimental design of their study was based on using Pearson's correlation test with Beta values, which resulted in the identification of 19 CpG sites after applying stringent parameters (abs $r \geq 0.6$, and *P*-value $<10^{-13}$). Based on the findings presented here, more markers may have been obtained if the Spearman's rank correlation test with either Beta values or M values, or either Pearson's correlation test or simple linear regression with M values were used.

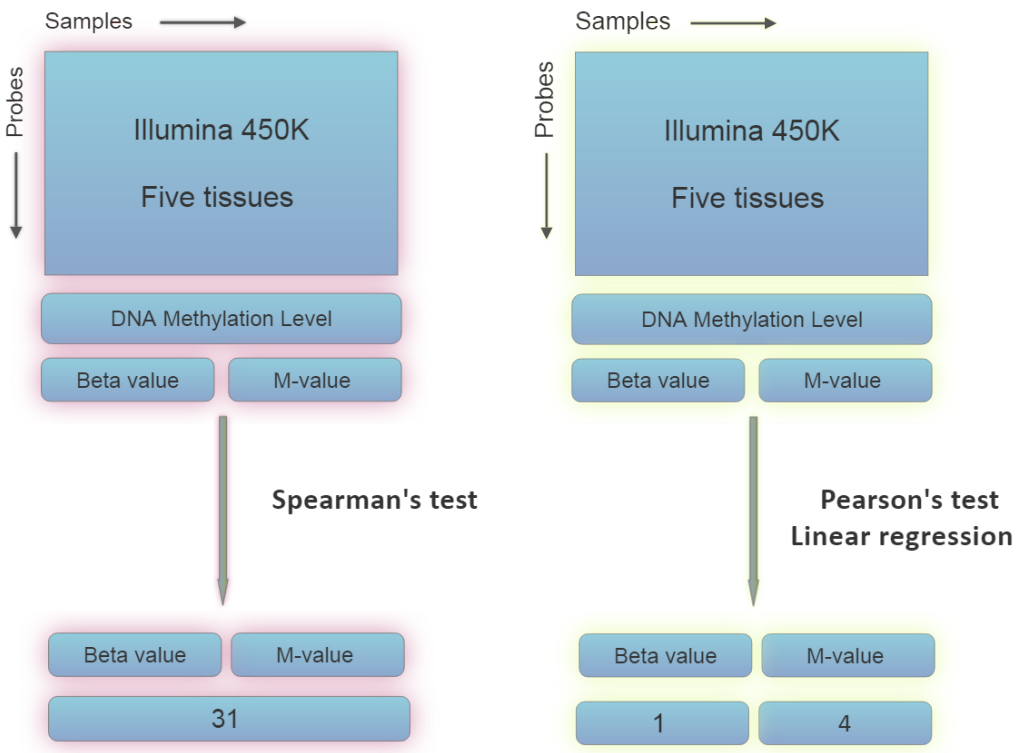

Figure 3.19 Schematic diagram summarising the outcomes of three statistical tests used to identify AR CpG sites across five tissues assayed on the Illumina HM450K BeadChip platform.

Another finding of the results presented here was the significant difference in outcomes between the two DNAm measurements, Beta values and M values. As can be seen in the results, M values outperformed Beta values, in terms of the number of significant AR CpG sites identified, when they were used in both the Pearson's correlation and simple linear regression tests to measure the association between DNAm level and chronological age. This indicates that, despite the fact that Beta values are widely used in AR studies, this has some limitations when parametric statistical tests are being implemented. A similar finding was reported by Du et al. (2010), who recommended using M values for conducting differential methylation analyses, as they perform better both in terms of the detection rate and of detecting true positives, for both highly methylated and unmethylated CpG sites [105]. This superior performance can be explained

by the fact that the Logit transformation of Beta values into M values reduces their heteroscedasticity [105]. That is, the resulting M values have more constant variation in DNAm levels in both highly methylated and unmethylated regions. Thus, the relationship between DNAm status and chronological age at the extreme values of methylation will become linear when Beta values are converted to M values.

A review of the literature showed that the majority of AR methylation studies have not used M values in their analyses. In fact, they have used Beta values not only for detecting AR CpG markers, but also in linear-based age-prediction models [89,145,157]. Consequently, this is likely to result in poor performance of these linear age-prediction models, as they only measure linear relationships rather than monotonic relationships. However, two exceptions to this were the studies conducted by Bekaert et al. (2015) and Xu et al. (2015) who, rather than using M values with linear regression analysis, used nonlinear (quadratic) regression methods to fit the best monotonic relationship between Beta values and chronological ages of donors. Furthermore, Bekaert et al. [143] compared the prediction accuracy of linear and nonlinear age-prediction models based on Beta values, and found that nonlinear (quadratic) regression had a better age-prediction accuracy. Our results indicate that, since linear methods such Pearson's correlation and simple linear regression tests were able to capture the association between DNAm level based on M values rather than Beta values, implementing a linear regression modelling system with M values would produce a better result than Beta values in terms of prediction accuracy. In contrast, Beta values should only be implemented with nonlinear methods such as Spearman's rank correlation tests and quadratic regression modelling systems.

The objective of the second part of this study was to implement identified optimum methods, which are to use Spearman's rank correlation (with either Beta or M values), and then using M values in stepwise regression analysis, and for

building the linear multivariate linear regression model. For this reason, an independent data set of DNAm profiles from 54 saliva samples were obtained from an online public repository, which have already been used to build a saliva-specific model by Hong et al. (2017), which is the best age-prediction model reported in the literature to date. Therefore, their data set represented a good opportunity to assess the utility of different types of statistical approaches, in order to see whether further enhancement could be achieved in terms of the accuracy of the age-prediction model in saliva.

Initially, 988 statistically significant saliva-specific AR CpG sites were identified using the criteria of Spearman's rho >0.6 at FDR <0.05, which were then further filtered to 49 candidate markers using a more stringent FDR value ($\leq 1e^{-7}$). Although the same training data set was used as that used by Hong et al. (2017), only seven of the 49 CpG sites overlapped with the 62 AR CpG sites they detected, and three of the 49 CpG sites overlapped with the seven CpG sites included in their model. Furthermore, their stepwise regression analysis yielded a model with only four AR CpG markers from the 62 markers they initially identified, whereas this study yielded a model containing nine AR CpG markers (Table 3.8) from the original 49 CpG sites (**Error! Reference source not found.**). This may be due to the fact that, in this study, M values were used rather than Beta values, which may have reduced the skewness of the data and the variation in DNAm values at AR CpG sites. This in turn may have rendered DNAm level more linearised with chronological age.

The nine AR CpG markers, selected based on the data from 54 saliva samples assayed on the HM450K BeadChip, explained 97.5% of the variation in DNAm level. All of these nine markers have previously been reported in different AR studies, but have never been used together in a single model to predict age from saliva samples. Only two sites (cg25124276 and cg00573770) have previously been found to be associated with age in saliva, the remaining seven

have previously been linked to aging in blood samples [136,146,153,180-184]. Two markers (cg10501210 and cg10804656) were not linked to any known gene, and the remaining seven markers mapped to seven different genes, namely: *ZEB2, OTUD7A, B3NT9, GRNI2C, ELOVL2, FAM123C, and GPR158*. The CpG markers associated with these genes were found to be hypermethylated with age, except for one CpG site (cg00573770) linked to *ZEB2*, which was found to be hypomethylated with age. Two of the genes (*OTUD7A* and *ELOVL2*) have frequently been reported in AR studies, and they encode for deubiquitinising enzyme, and fatty acid elongase 2, respectively [146,147,179,184].

In comparison to the original study conducted by Hong et al. (2017), it can be seen that both the selection of AR CpG sites and the construction of an age-prediction model have been enhanced by using Spearman's rank correlation test and M values in both stepwise regression analysis, and multivariate linear regression (Figure 3.20). In this study, the selected nine CpG sites explain 97.5% of the total variation in the training data set, as opposed to 96.9% explained by the seven CpG sites from the Hong et al. (2017) study. Furthermore, the performance of both models was tested on an African population known as the Khomani San, which is one of the most genetically diverse human populations in the world [180]. Despite the fact that the testing data set came from a genetically diverse population (Khomani San), the model reported here was able to predict their ages with an accuracy of 5.1 years (MAD), with no significant difference in age-prediction accuracy between the sexes. This performance is an improvement on the Hong et al. (2017) model, which gave an MAD value of 8.3 years on the same data set. These results also suggest that during the construction of an age prediction model, either M values should be used with a linear regression modelling system or Beta values used with a nonlinear (quadratic) regression modelling system.

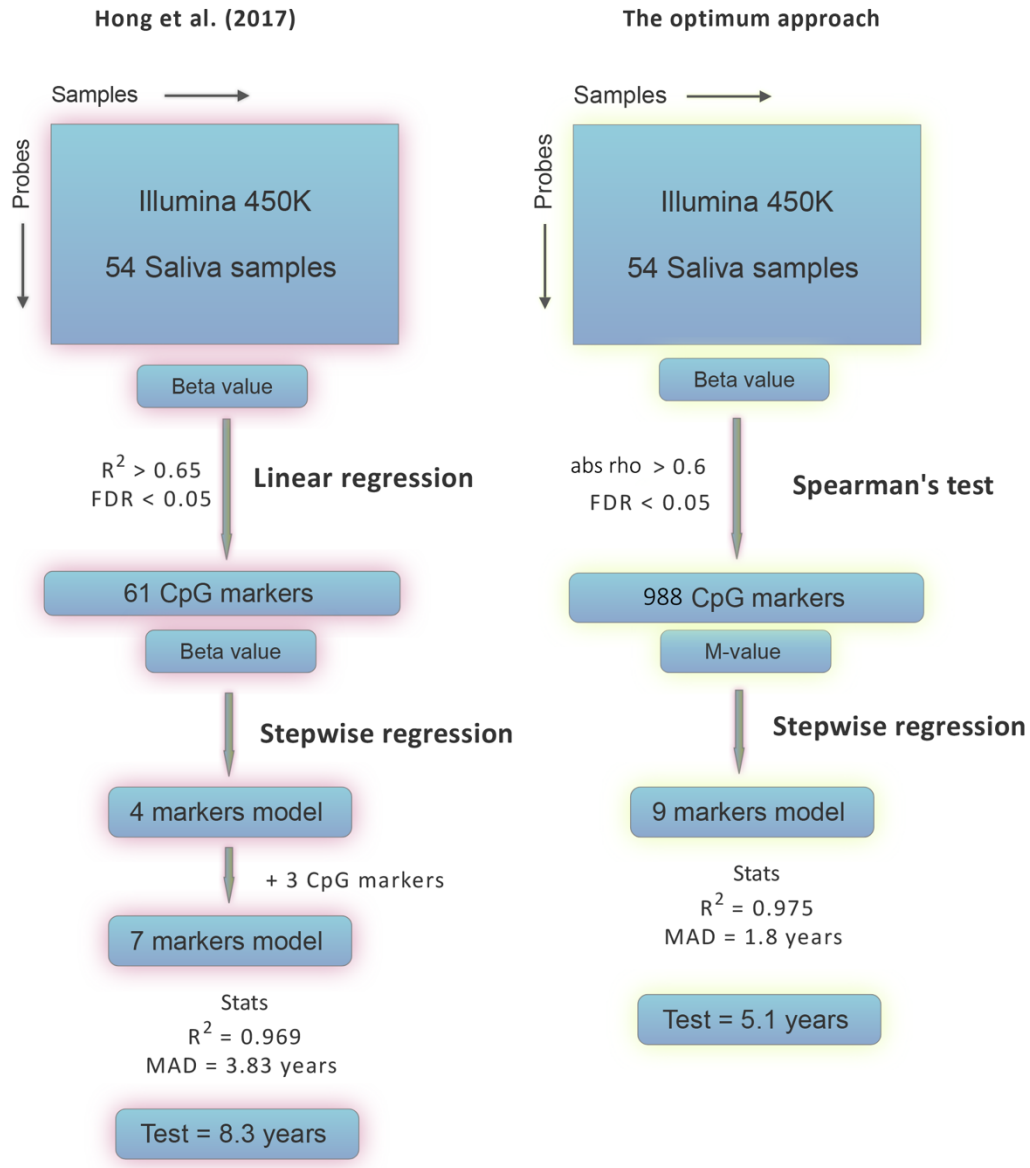

Figure 3.20 Schematic diagram showing the comparison between the HM450K model constructed using the optimum method identified in this study, and Hong et al's model.

3.8 Summary and conclusions

The main goal of this preliminary study was to enable researchers carrying out AR DNAm studies to make a more informed decision when it comes to selecting the right statistical method for identifying AR DNAm markers. It was found that the Spearman's rank correlation test detected a larger number of significant AR DNAm markers than both the Pearson's correlation test and simple linear regression analysis. These findings will assist other studies in selecting the most appropriate test for discovering candidate CpG sites and are likely to be of significance to both the forensic science and medical sciences fields. Furthermore, an analysis was conducted to study the effect of using different measures of DNAm levels on the outcomes of the correlation and regression analysis tests. The results suggest that using either Beta or M values with Spearman's correlation test is the best approach, however, Pearson's correlation test and simple linear regression detected fewer AR sites based on Beta values compared to M values. This also suggests that the relationship between DNAm levels and chronological age is monotonic and that researchers should consider this when selecting a statistical test for identifying AR DNAm sites and building age prediction models.

Taken together, the findings of this Chapter are summarised in Figure 3.21, which illustrates the optimum method that should be used to identify AR CpG sites with different DNAm measures. The first pathway is to use Spearman's rank correlation test with Beta values and then a nonlinear modelling system for building an age-prediction model as Beta values are monotonically associated with chronological age. The second procedure is to use Spearman's rank correlation test with M values and then a linear modelling system, as converting Beta values to M values will make the association between DNAm level and chronological age linear.

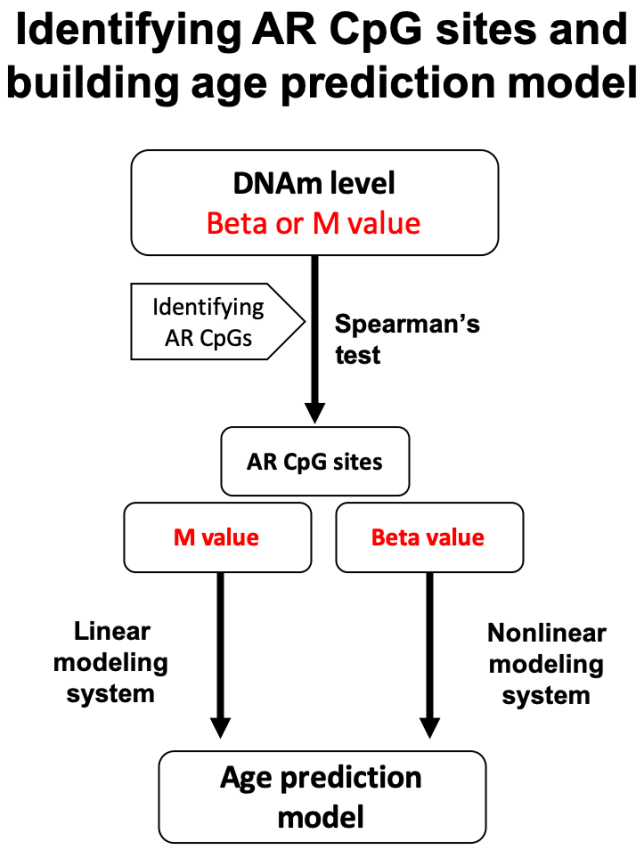

Figure 3.21 The optimum pathway for identifying AR CpG sites, and building age-prediction models using either Beta values or M values.

From a review of the literature, the identified nine saliva AR CpG sites in this study are strong candidates for achieving a better age prediction accuracy than those already reported in the literature. Validating these AR CpG sites on different DNAm assays would allow samples to be tested within a forensic DNA laboratory workflow. To achieve this, a proper assay design should be selected that could be used for the types of samples that are encountered in forensic cases. Forensic specimens are usually found in harsh environments that predispose the DNA to degradation, which results in low quantity and quality DNA, and thus, the selected assay should be suitable for use on this type of samples. Therefore, in

the next chapter the nine AR CpG sites along with the seven AR CpG sites from Hong et al.'s model are examined using bisulfite sequencing using the Illumina MiSeq® platform.

Chapter 4: DNA methylation-based age prediction from saliva samples using next-generation sequencing on the Illumina MiSeq® platform

4.1 Introduction

In forensic science, saliva constitutes a major source of DNA from various types of evidence collected at crime scenes, such as cigarette butts, chewing gum, toothbrushes, and drinking/eating items [185]. In addition, saliva samples can be taken in a non-invasive and convenient way, for medical screening and other diagnostic applications. For this reason, researchers have shown an increased interest in identifying age-related (AR) CpG markers for saliva samples [156,157,186]. The most promising AR CpG sites in the literature that have been identified for blood and other tissues have also been tested on saliva samples, in order to assess their ability to predict age in samples from other somatic tissues. Three blood-specific AR CpG sites associated with three different genes (*PDE4C*, *ASPA*, and *ITGA2B)* that were identified by Weidner et al. [145] were tested by Eipel et al. (2016) on 55 pyrosequencing profiles from buccal swab samples, and then validated on another independent data set of 55 buccal swab samples. This model (referred to as the "3-CpG-blood-model") had a mean absolute deviation (MAD) between predicted and chronological age of 4.3 years based on the training data set, and 7.03 years based on the testing data set. This model was further enhanced by adding an additional two saliva-specific CpG markers (associated with the genes *CD6* and *SERPINB5*), which, along with the three blood-specific CpG markers, had a MAD value of 5.12 years based on the testing

data set. Another two models were built by Silva et al. (2015), from two sets of AR CpG markers located in the *GRIA2* and *NPTX2* genes [155]. The first model (*GRIA2* model), containing three CpG markers, had a MAD value of 6.9 years based on the training data set, and the second model (*NPTX2* model) contained six CpG markers and had a MAD value of 9.2 years based on the training data set. These models were not validated on an independent set of samples (a testing data set). The final attempt at using AR CpG markers from other tissues was by Vidaki et al. (2017), who took 16 universal AR CpG markers from the pan-tissue model (353 CpG markers) created by Horvath et al. [106], and used them to predict age from saliva samples. Based on their training data set, Vidaki et al.'s model had a MAD value of 3.18 years, and 4 years based on an *in silico* testing data set [111].

Since the aforementioned markers and models were identified based on tissues other than saliva, researchers have also tried to improve the accuracy of DNA methylation (DNAm) age prediction models by identifying markers that are specific to saliva. Bocklandt et al. (2011) were the first researchers to identify three saliva-specific AR CpG sites, which were mapped to promoters associated with three different genes, namely *EDARADD*, *NPTX2*, and *TOM1L1*. The DNAm levels at these CpG sites were obtained from samples from 22 pairs of twins, 31 unrelated males and 29 unrelated females (aged 18–70 years), and modelled using multivariate linear regression, which explained 73% of the variation in the DNAm level. This model was able to predict age in the training data set with a MAD between predicted and chronological age of 5.2 years, but the model was not validated on an independent data set [156]. The most recent saliva-specific model was constructed by Hong et al. (2017), which consisted of six AR CpG markers (cg00481951, cg19671120, cg14361627, cg08928145, cg12757011, and cg07547549, mapped to the *SST*, *CNGA3*, *KLF14*, *TSSK6*, *TBR1*, and *SLC12A5* genes, respectively), in addition to one non-AR CpG marker (cg18384097 in the *PTPN7* gene) that was shown to be differentially methylated

in saliva samples. These markers were selected using 54 saliva samples assayed on the Illumina HumanMethylation450 (HM450K) BeadChip and based on a simple linear regression (R^2) value >0.65, a false discovery rate (FDR) <0.05, and a difference in DNAm level (based on Beta-values) between young and old individuals >0.1. This model based on seven CpG sites was then validated on a data set of 113 independent saliva samples assayed using SNaPshot minisequencing, and had a MAD of 3.15 years, representing the highest age prediction accuracy for saliva samples reported in the literature to date [157].

Even though these AR CpG markers have been identified and age prediction models built for saliva samples, there is still scope for improvement of the age prediction accuracy of these models. For instance, as shown in Chapter 3, selecting Spearman's rank correlation for identifying AR CpG sites and then using DNAm level measured as M values in stepwise regression for determining the best subset of markers produced an age prediction model with more significant AR CpG predictors. In addition, targeted sequencing of regions that are known to contain AR CpG sites could result in the discovery of adjacent CpG sites that could have a significantly greater association with age, which can further enhance the accuracy of age-prediction models [155].

4.2 Aims

The aim of this study was to validate the nine AR CpG sites identified in Chapter 3 using bisulfite sequencing on a high-resolution sequencing platform, the Illumina MiSeq®. The performance of these markers was compared with the seven AR CpG markers from the saliva-specific age prediction model created by Hong et al. (2017). The main reason for using Hong et al.'s model was because it is the most accurate saliva-specific model reported in the literature to date, thus using it as a benchmark to evaluate our model would provide an indication about its performance.

4.3 Objectives

- Collecting saliva samples from different individuals.
- Extracting the genomic DNA from saliva samples.
- The DNAm level at the nine AR CpG sites along with the CpG sites surrounding their genomic regions were measured by bisulfite sequencing using Illumina MiSeq®.
- Constructing saliva-specific age prediction model based on sequencing results.
- Testing the constructed age prediction model on independent saliva samples.
- Constructing Hong et al.'s model using their seven saliva-specific AR CpG sites.
- Testing Hong et al.'s model on independent saliva samples and compare the estimation accuracy with our nine AR CpG sites.

4.4 Materials and methods

The nine AR CpG markers in the HM450K model were further validated by targeted bisulfite sequencing using the Illumina MiSeq® platform. The sample collection, DNA extraction, and statistical analyses were carried out at University of Strathclyde. However, assessment of the quality/quantity of the genomic DNA (gDNA), primer design, sequencing of the targeted regions of interest (ROI), and sequence alignments were carried out by Zymo Research Corporation.

4.4.1 Samples

Saliva samples were collected from 192 individuals following ethical approval of the study by the University of Strathclyde Department of Pure and Applied Chemistry Departmental Ethics Committee. Prior to sample donation and

after receiving an information sheet explaining the study, participants signed a consent statement (Appendix B). The saliva samples were obtained from 88 males and 104 females aged 12 to 96 years (Figure 4.1). The saliva samples were collected in small vials and stored at 4°C. DNA was extracted and the concentration and quality of the gDNA was measured as described in Section 2.6.3 and 2.6.3.

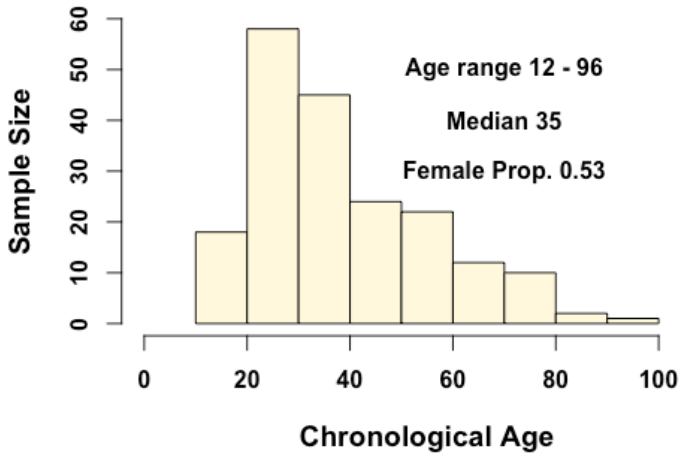

Figure 4.1 Age distribution for the donors of 196 saliva samples, including the age range, median age and proportion of females.

4.4.2 Targeted bisulfite sequencing

Primers targeting the nine AR CpG sites (Table 4.1), identified in Chapter 3 were designed as described in Section 2.6.4. Following primer validation, gDNA from the saliva samples was bisulfite converted and the specified ROI were amplified using the primers followed by massively parallel sequencing using the

Illumina MiSeq® system, as described in Section 2.6.5. Finally, the sequence reads were identified and processed as described in Section 2.6.6.

Table 4.1 The nine CpG markers in the saliva-specific HM450K model.

Probe ID	Chromosomal range	Gene
cg16867657	chr6:11044877	*ELOVL2*
cg10501210	chr1:207997020	*Unknown*
cg10804656	chr10:22623460	*Unknown*
cg04875128	chr15:31775895	*OTUD7A*
cg06279276	chr16:67184164	*B3GNT9*
cg00573770	chr2:145278485	*ZEB2*
cg07365960	chr17:72848535	*GRIN2C*
cg23606718	chr2:131513927	*FAM123C*
cg25124276	chr10:25464008	*GPR158*

4.4.3 Construction of an age-prediction model from bisulfite sequencing profiles

The sequenced regions of the nine candidate AR CpG sites in 192 saliva samples, generated using the Illumina MiSeq® platform, were used to retrain and validate the age prediction model. The samples were randomly divided using the *sample* function in R software as described in Section 2.5.3 into a training data set (60% of the samples) to construct the model and a testing data set (40% of the samples) to validate the prediction accuracy of the model. DNAm levels at the candidate AR CpG markers were regressed on the chronological ages of the

donors of the samples in the training data set using one of the options described in the conclusion of Chapter 3 and illustrated in Figure 3.21. The performance of this model was validated using samples in the independent testing data set. Construction and validation of the model were performed using R software, as described in Section 2.5.4.

4.4.4 Construction of Hong et al.'s model from bisulfite sequencing profiles

To further assess the performance of the saliva-specific HM450K model, the seven CpG sites (Table 4.2) from the model constructed by Hong et al. (2017) were also sequenced in order to compare them with the candidate markers selected in this study. Primer design and sequencing of the seven AR CpG markers were carried out as described in Section 2.6.4 and 2.6.5. The model constructed using the seven CpG markers was trained and tested on the same training and testing data sets that were used to construct and validate the saliva-specific HM450K model.

Table 4.2 The seven AR CpG markers identified by Hong et al. (2017) and included in their saliva-specific age-prediction model. Genomic locations are for the human genome assembly GRCh37, also known as hg19.

Illumina's Probe ID	Chromosomal range	Amplicon size (bp)	Number of CpG sites
cg18384097	chr1:202129366-202129766	95 bp	6
cg00481951	chr3:187387450-187387850	51 bp	4
cg19671120	chr2:98962774-98963174	85 bp	13
cg14361627	chr7:130418916-130419316	71 bp	6
cg08928145	chr19:19625164-19625564	82 bp	12
cg12757011	chr2:162280911-162281311	68 bp	4
cg07547549	chr20:44658025-44658425	82 bp	11

4.5 Results

4.5.1 Samples and primer QC (pre-sequencing)

The gDNA in each of the 192 samples was quantified and assessed for quality. The quality of the gDNA is indicated by DNA integrity number (DIN). Samples with ≥10ng/μL and DIN ≥3 were selected for the downstream analyses. Because bisulfite conversion treatment has a destructive effect on the gDNA, having samples with lower than these values will impact the overall amount of intact DNA template that is available for subsequent PCR amplification. The number of samples that passed these criteria was 168 of 192 samples. However, the samples with low quantity and/or degraded DNA were also included for sequencing, in order to test how this would affect the accuracy of age estimation, given that samples of this nature are frequently encountered in a forensic context. Thus, primers were designed so that they targeted smaller sized amplicons, in order to avoid dropout in the samples that contained degraded DNA. Primer pairs for each of the nine CpG sites successfully passed the validation step. The targeted region for each candidate CpG site contained at least four additional CpG sites (Table 4.3).

Table 4.3 Primer validation results for the nine saliva-specific candidate AR CpG sites.

Probe's ID	Chromosomal range	Amplicon size (bp)	Number of CpG sites
cg00573770	chr2:145278285-145278685	87 bp	4
cg04875128	chr15:31775595-31776195	195 bp	42
cg06279276	chr16:67183864-67184464	65 bp	11
cg07365960	chr17:72848235-72848835	136 bp	24
cg10501210	chr1:207996820-207997220	66 bp	9
cg10804656	chr10:22623260-22623660	105 bp	19
cg16867657	chr6:11044677-11045177	33 bp	9
cg23606718	chr2:131513627-131514227	153 bp	30
cg25124276	chr10:25463808-25464208	50 bp	9

4.5.2 Sequencing results

To study the effect of DNA quantity and quality on the sequencing results, the number of reads produced for each sample were plotted against the DNA concentration (ng/μL) and DIN value. Figure 4.2A and B show the number of reads with increasing DNA concentration and increasing DIN value in each sample, respectively. In addition, as Figure 4.2C illustrates, low intra-marker variation in read number was seen between samples within each maker, and high inter-marker variation among markers. Two of the identified AR CpG markers, namely cg04875128 and cg23606718, failed to produce results in 37% and 89% of the samples, respectively. Therefore, these markers were removed before carrying out downstream statistical analyses.

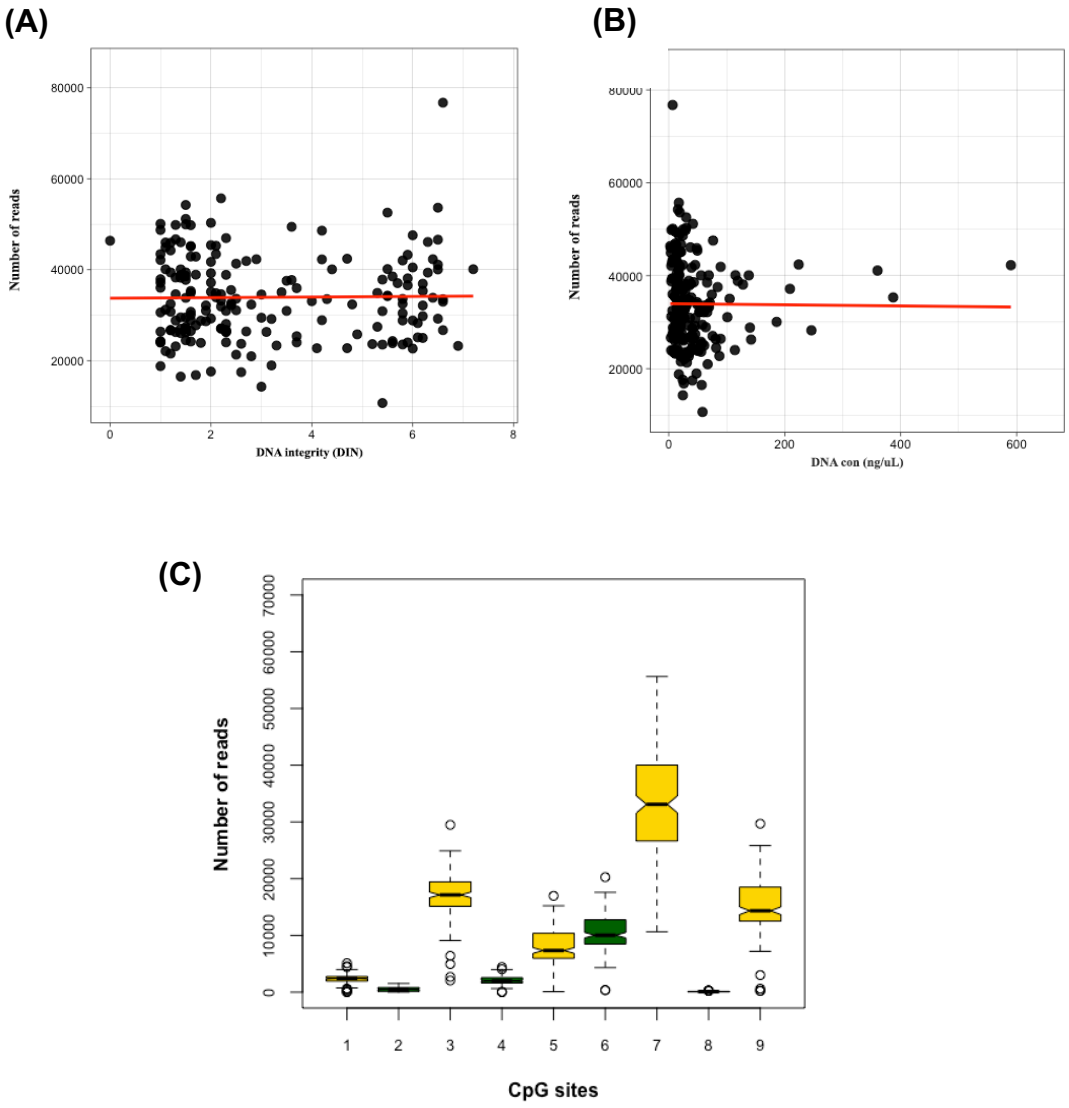

Figure 4.2 Number of Illumina MiSeq® sequencing reads versus **(A)** DNA integrity and **(B)** DNA concentration (ng/μL) in 192 saliva samples. The red lines in both graphs are lines of best fit, which show that there is no change in the number of reads with either DIN value or DNA concentration. **(C)** Box plot showing the number of reads covering the amplified regions for each of the nine targeted CpG sites.

4.5.3 Statistical analyses

To ensure reliable identification of AR CpG sites, only samples that had DNA concentrations of ≥10ng/μL (168 samples) were included in these

downstream analyses. In addition, to avoid biased selection of the makers, the samples were randomly split into a training data set of 100 (60% of the samples) and a testing data set of 68 (40% of the samples), with equal representation of age groups between the two data sets.

The relationship between DNAm at the remaining seven CpG sites and chronological age was compared between the HM450K and MiSeq® platforms (Table 4.4), to assess whether there was any discrepancy in terms of the positive or negative relationship (hyper- or hypomethylation); no discrepancies were identified. Lower correlation coefficients (rho) were seen for the sequencing data compared to the data from the HM450K BeadChip. These observed differences are partly due to the different sample sizes and age ranges between the two training data sets used in the initial study and this validation study. However, three markers (cg06279276, cg07365960, and cg25124276) did show a significant reduction in correlation coefficient between the BeadChip and sequencing data, from rho values of ~0.8 to 0.3, 0.5, and 0.5, respectively. Thus, they were removed from further downstream analyses.

Table 4.4 Spearman's rank correlation test between DNAm level and chronological age at the seven candidate AR CpG sites, based on 54 saliva samples assayed on the HM450K BeadChip, and based on the training data set of 100 saliva samples sequenced on the MiSeq® platform.

Illumina's Probe ID	CpG site position	HM540K Spearman's coefficient (rho)	MiSeq® Spearman's coefficient (rho)
cg00573770	chr2:145278484	-0.80	-0.72
cg06279276	chr16:67184164	0.76	0.34
cg07365960	chr17:72848534	0.80	0.50
cg10501210	chr1:207997020	-0.86	-0.70
cg10804656	chr10:22623460	0.80	0.76
cg16867657	chr6:11044877	0.90	0.76
cg25124276	chr10:25464008	0.77	0.51

As well as the seven-remaining target CpG sites, an additional 157 CpG sites found within these target regions were also included in the sequencing

results. All these additional CpG sites were tested for age association, in order to determine whether any markers had a stronger association with age than the original nine target sites.

Of the additional 157 CpG sites, 28 CpG sites were found to be strongly associated with chronological age. These newly-identified AR CpG sites were located in four amplicons: cg00573770 (*ZEB2*), cg10501210 (unknown gene), cg10804656 (unknown gene), and cg16867657 (*ELOVL2*). To distinguish between the newly-identified CpG sites, they were numbered from CpG1 to CpG28 (Table 4.5). At this stage, the total number of AR CpG sites resulting from the sequencing of seven genomic regions was 32.

Table 4.5 Spearman's rank correlation test between DNAm level and chronological age at the 28 adjacent CpG sites based on 100 saliva samples (training data set) sequenced on the MiSeq® platform. The CpG sites were designated by number (CpG site name) for the purposes of identification only.

Sequence amplicon (gene)	CpG site name	Chromosomal position	Spearman's coefficient (rho)	*P*-value
cg00573770 (*ZEB2)*	CpG1	chr2:145278476	-0.68	5.1E-15
	CpG2	chr2:145278508	-0.72	< 2.2E-16
	CpG3	chr2:145278563	-0.59	8.4E-11
cg10501210 (*unknown*)	CpG4	chr1:207997016	-0.72	< 2.2E-16
	CpG5	chr1:207997025	-0.74	< 2.2E-16
	CpG6	chr1:207997046	-0.75	< 2.2E-16
	CpG7	chr1:207997049	-0.69	2.6E-15
	CpG8	chr1:207997059	-0.78	< 2.2E-16
cg10804656 (*unknown*)	CpG9	chr10:22623379	0.69	1.7E-15
	CpG10	chr10:22623380	0.68	6.4E-15
	CpG11	chr10:22623391	0.74	< 2.2E-16
	CpG12	chr10:22623393	0.71	< 2.2E-16
	CpG13	chr10:22623401	0.71	< 2.2E-16
	CpG14	chr10:22623416	0.76	< 2.2E-16
	CpG15	chr10:22623429	0.73	< 2.2E-16
	CpG16	chr10:22623434	0.71	< 2.2E-16
	CpG17	chr10:22623436	0.74	< 2.2E-16
	CpG18	chr10:22623439	0.69	3.0E-15
	CpG19	chr10:22623445	0.76	< 2.2E-16
	CpG20	chr10:22623453	0.75	< 2.2E-16
	CpG21	chr10:22623479	0.75	< 2.2E-16
	CpG22	chr10:22623481	0.73	< 2.2E-16
cg16867657 (*ELOVL2*)	CpG23	chr6:11044860	0.71	< 2.2E-16
	CpG24	chr6:11044863	0.71	< 2.2E-16
	CpG25	chr6:11044866	0.76	< 2.2E-16
	CpG26	chr6:11044872	0.74	< 2.2E-16
	CpG27	chr6:11044874	0.78	< 2.2E-16
	CpG28	chr6:11044879	0.74	< 2.2E-16

4.5.4 Construction and validation of a saliva-specific age-prediction model

Due to the fact that several of the additional adjacent CpG sites showed a strong association with age, they were also exploited for the purposes of building a saliva-specific age-prediction model. Stepwise regression analysis was

implemented in order to re-build the model, selecting the best subset of the 32 AR CpG markers (four obtained from the HM450K data, and 28 adjacent CpG sites obtained from MiSeq® data). As one of the recommended options in Chapter 3, Beta values were used with a non-linear modelling system (quadratic regression) in order to capture the monotonic relationship between DNAm level at the AR CpG sites and the chronological age of the sample donors. This was done by including additional squared terms of the Beta values in the model, alongside the standard Beta value terms included in the stepwise regression analysis. The algorithm selected a model composed of ten different AR CpG markers (cg00573770, CpG5, CpG7, CpG16, CpG17, CpG18, CpG19, CpG21, CpG24, and CpG27), and one additional CpG marker (cg00573770), which appears twice in the model using both its standard Beta value and its squared Beta value (Figure 4.3). This quadratic model (Table 4.6) explained 92% of the total variation in DNAm levels in the 100 saliva samples in the training data set ($R^2 = 0.92$), and showed a high correlation between predicted and chronological ages (Pearson's coefficient (r) = 0.96), with a MAD value of 3.4 years (Figure 4.4A).

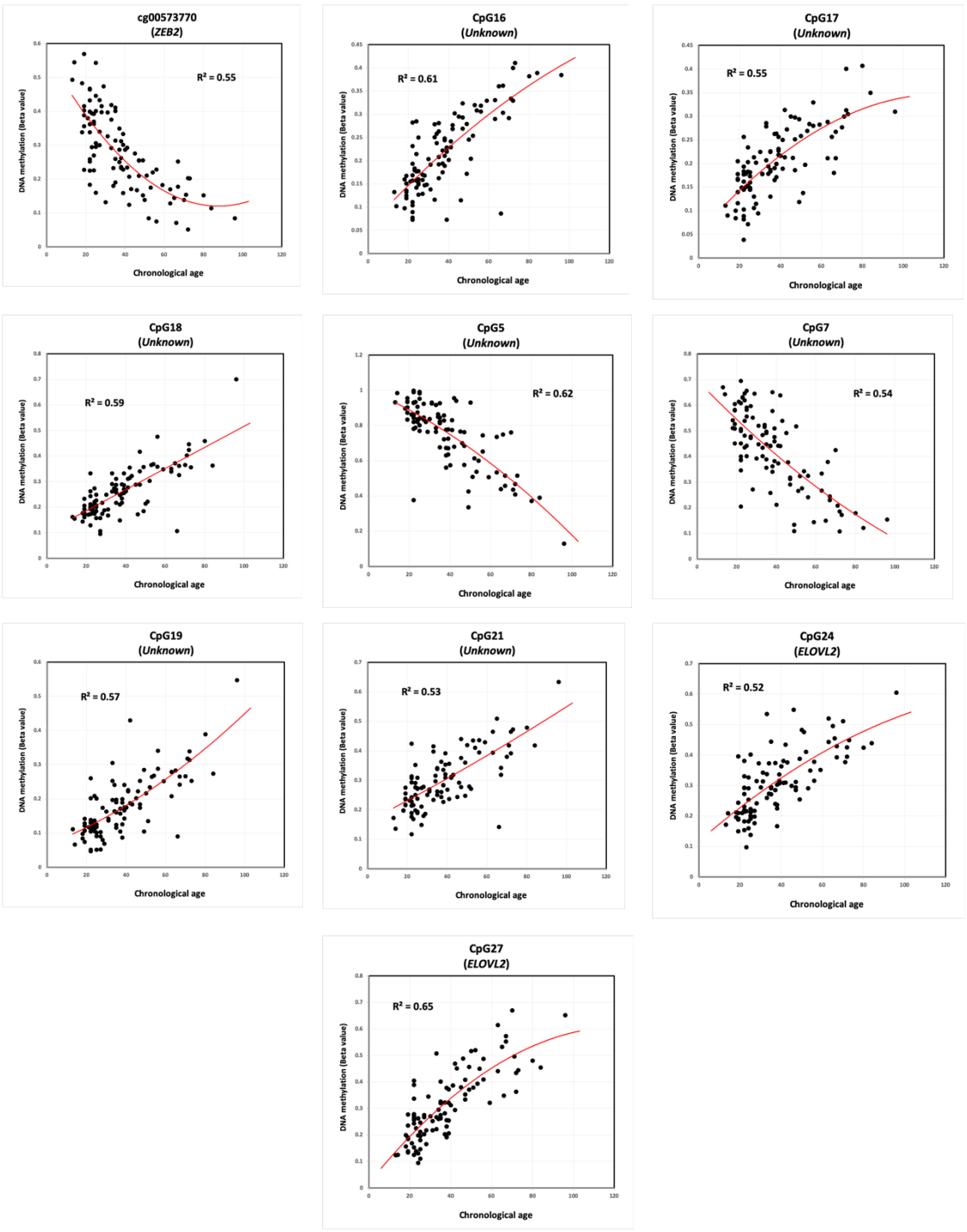

Figure 4.3 Scatter plots showing the change in DNAm level with age at ten saliva-specific AR CpG markers selected by stepwise regression. DNAm level was determined by targeted bisulfite sequencing using the Illumina MiSeq® platform.

Table 4.6 Quadratic regression model composed of ten AR CpG markers trained on methylation data obtained from Illumina MiSeq® sequencing of 100 saliva samples. The ^2 in the term column represents the squared Beta value of the marker.

Term	*Coefficients*	*P-value*	*R-squared*	*P-value*
Intercept	48.41	4.9E-14	0.92	1.76E-43
CG00573770	-72.11	2.8E-3		
CpG16	-44.88	5.0E-3		
CpG17	75.57	3.2E-5		
CpG18	-49.83	2.4E-4		
CG00573770^2	81.26	2.8E-2		
CpG5^2	-12.38	1.6E-3		
CpG7^2	-17.26	1.6E-2		
CpG19^2	47.86	8.2E-4		
CpG21^2	183.09	8.7E-10		
CpG24^2	59.52	1.3E-4		
CpG27^2	25.74	3.3E-2		

Subsequently, the performance of this model was validated on an independent data set of DNAm levels derived from targeted bisulfite sequencing of 68 additional saliva samples. The overall MAD between predicted and chronological age based on bootstrap analysis was 5.26 years (Pearson's correlation (r) = 0.88) with 95% confidence intervals of 5.24-5.27 years (Figure 4.4B). Interestingly, the prediction accuracy for individuals aged <30 was 3.36 years, whereas for individuals aged >40 years old it was 5.55 years. The higher deviation from chronological age in individuals of more advanced age has been observed in other DNAm-based age prediction models in different tissues [106,155,157].

To avoid sex bias in age estimation, male and female samples in the testing data set were separated and their MAD values assessed separately. Although a

t-test showed a non-significant (*P*-value = 0.6) difference in the prediction accuracy between the two sexes, the difference between DNAm age and chronological age was slightly higher in males (MAD = 5.6 years) compared to females (MAD = 5 years). This result is in line with Hannum et al. (2013), who reported that the methylome of men appears to change with age faster than that of women [136]. Finally, in order to test how low quantity/quality DNA samples affect age estimation, the model was used to predict age for 27 independent saliva samples with DNA <10ng/μl and DIN value <3, resulting in a MAD value of 11.5 years. Therefore, this suggests that carrying out age estimation on DNA samples with concentration <10ng/μL and DIN value <3 may give inaccurate results.

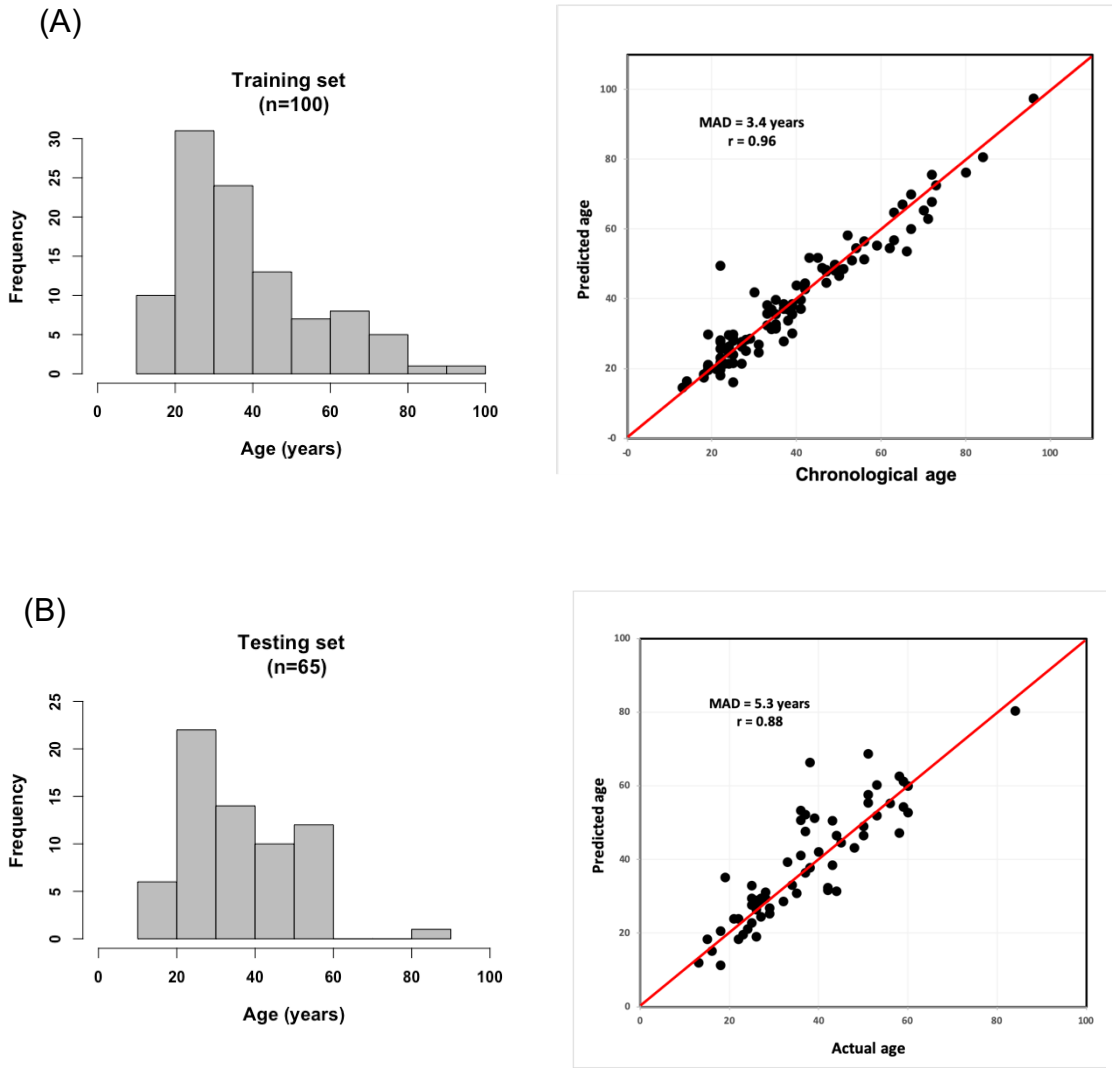

Figure 4.4 The performance of the quadratic regression model consisting of ten AR CpG markers using **(A)** the training data set of 100 samples and **(B)** the testing data set of 68 samples. The left panel shows a histogram illustrating the age range in the relevant data set, and the right panel shows a scatter plot illustrating the prediction accuracy of the model.

4.5.5 Construction and validation of Hong et al.'s model

Primer pairs for the seven CpG sites (six age-associated markers and one saliva-specific marker) included in the Hong et al. (2017) model successfully passed the validation and were sequenced on the Illumina MiSeq® platform in saliva samples from 192 individuals. The chronological ages of the donors of the 100 samples in the training data set were linearly regressed on DNAm level at each of these seven CpG markers. Table 4.7 compares the results of linear regression analysis based on SNaPshot minisequencing results obtained from the Hong et al. (2017) study and the MiSeq® platform results from this present study. The DNAm pattern with chronological age in the six AR CpG sites can be seen in Figure 4.5. The model was trained on the same training data set (100 samples) and tested on the same testing data set (68 samples) used in the previous section. This model (Table 4.8) explained 68% of the total variation in DNAm level in the 100 samples in the training data set (R^2=0.68), with a relatively high correlation between predicted and chronological ages (Pearson's coefficient (r) = 0.82), and a MAD value of 7.7 years (Figure 4.6A). In the testing data set the model gave a MAD value of 7.5 years, with a Pearson's correlation coefficient of r = 0.74 between the predicted and chronological ages (Figure 4.6B).

Table 4.7 The strength of linear association between DNAm level at six AR CpG markers and chronological age based on two different platforms, SNaPshot minisequencing and Illumina MiSeq®.

Illumina Probe ID	CpG site position	SNaPshot assay R^2 (n=54)	MiSeq® R^2 (n=100)
cg00481951	chr3:187387650	0.48	0.30
cg19671120	chr2:98962974	0.29	0.45
cg14361627	chr7:130419116	0.63	0.27
cg08928145	chr19:19625364	0.43	0.42
cg12757011	chr2:162281111	0.17	0.02
cg07547549	chr20:44658225	0.55	0.12

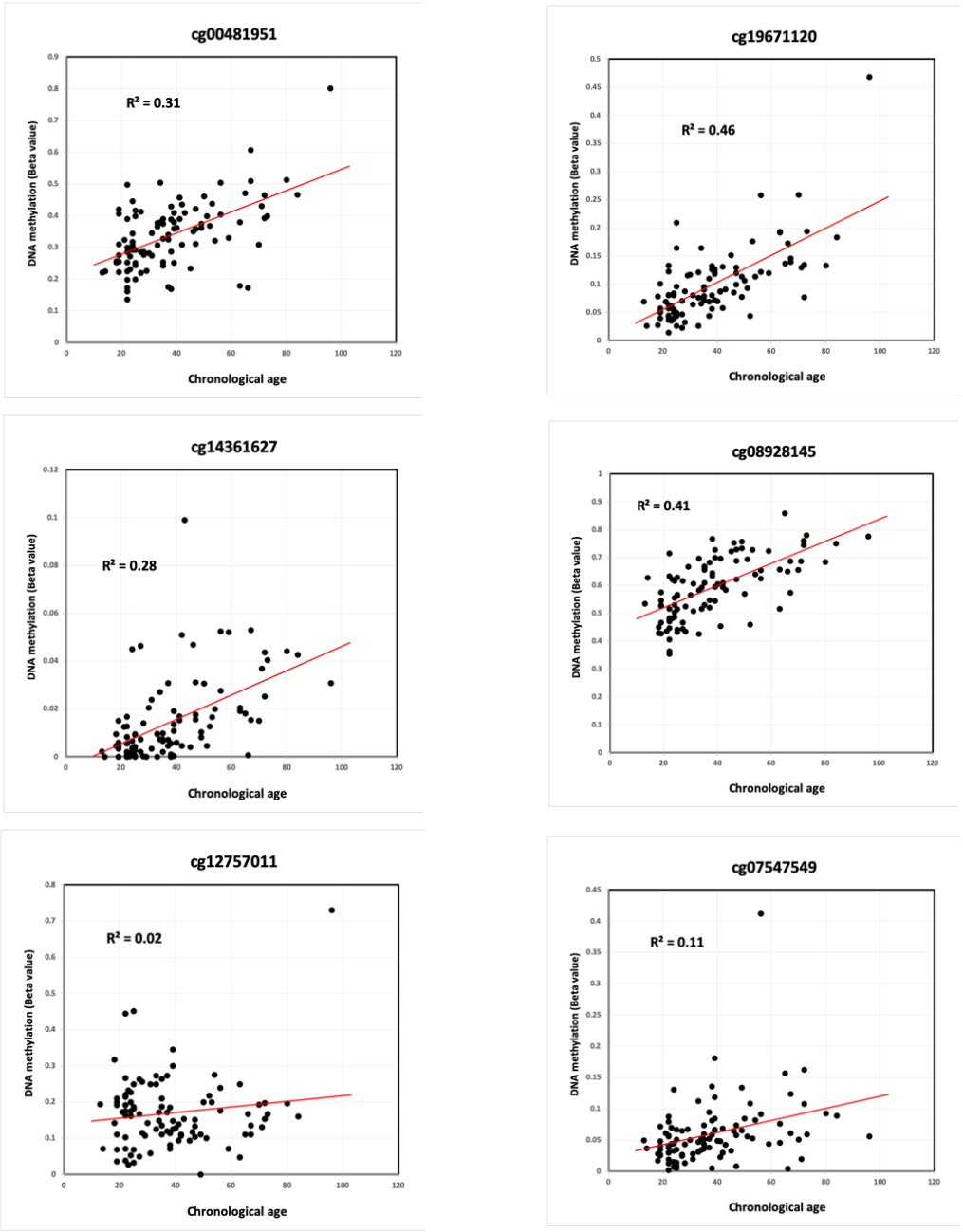

Figure 4.5 Scatter plots showing the change in DNAm level with age at the six saliva-specific AR CpG markers identified in the Hong et al. (2017) study. DNAm level was determined by targeted bisulfite sequencing using the Illumina MiSeq® platform.

Table 4.8 Regression model from the Hong et al. (2017) study composed of seven CpG markers trained on methylation data obtained from Illumina MiSeq® sequencing of 100 saliva samples.

Term	*Coefficients*	*P-value*	*R-squared*	*P-value*
Intercept	-11.21	0.16	0.68	3.0E-20
cg18384097	-5.90	0.38		
cg00481951	17.31	0.20		
cg19671120	121.65	0.00		
cg14361627	261.56	0.00		
cg08928145	51.55	0.00		
cg12757011	-10.38	0.44		
cg07547549	8.79	0.70		

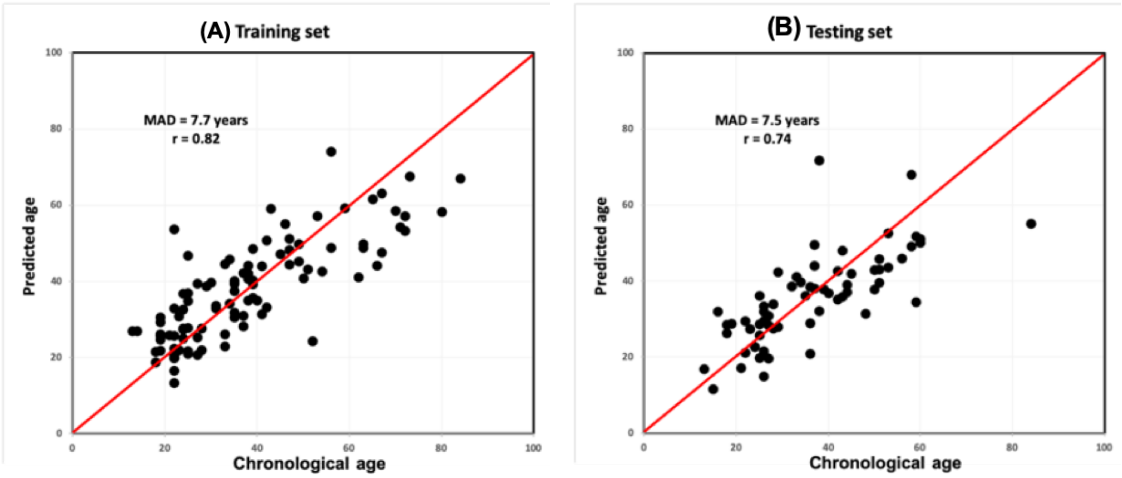

Figure 4.6 The performance of the Hong et al. (2017) model consisting of seven CpG markers. The scatter plots show the prediction accuracy of the model based on **(A)** the training data set of 100 samples and **(B)** the testing data set of 68 samples.

4.6 Discussion

In the last decade, researchers have identified AR DNAm markers for estimating the chronological age of individuals with high accuracy from samples of various tissue types, and these markers have outperformed all other known AR biomarkers, such as telomere shortening and mitochondrial deletions [187,188]. The aim of this study was to further enhance the age prediction accuracy of DNAm markers in saliva samples, by identifying greater numbers of AR CpG sites using the most appropriate statistical methods and modelling systems, which can capture the real relationship between DNAm level at AR CpG sites and the chronological age of sample donors.

Due to the fact that not all probes found on the Illumina HM27K and HM450K BeadChips are designed to target CpG sites with single-nucleotide resolution, thus, validating the HM27K and HM450K probe requires a methylation capture step in order to determine the exact genomic position of the CpG site under study. For this reason, the genomic regions surrounding the nine AR CpG sites identified in Chapter 3 were sequenced using bisulfite sequencing on a high-resolution sequencing platform, the Illumina MiSeq®. As expected, there were adjacent CpG sites that showed an association with age than the targeted CpG site under study. Thus, the model was re-built based on these adjacent AR CpG sites that showed a stronger association with age.

The final model that was built based on the Illumina MiSeq® data consisted of ten different AR CpG sites, including one CpG site (cg00573770) that was used twice (terms for both its Beta value and the square of its Beta value were included in the model). These markers are found within four clusters of CpG sites that are located on four different chromosomes: 1, 2, 6, and 10. Two of these clusters are linked to the genes *ZEB2* and *ELOVL2* (chromosomes 2 and 6, respectively), and the other two are not linked to any known genes (chromosome 1 and 10).The

presence of CpG sites from the *ELOVL2* gene in the final model is not surprising, as it is one of the most frequently included genes in age-prediction models for various types of tissues [148,189]. *ZEB2* (Zinc Finger E-Box Binding Homeobox 2) is a protein coding gene, which functions as a DNA-binding transcriptional repressor and is linked to Mowat-Wilson Syndrome and esophageal cancer [190]. Although they are not linked to any known genes, genomic locations 1q32.2 and 10p12.2 seem to be important in human aging, as they contain highly significant AR CpG sites. The former, containing the site cg10501210, was studied by Zbieć-Piekarska et al. [179] who used pyrosequencing in blood samples and found that all three of the identified CpG sites in that region were significantly associated with age. Finally, most of the markers (five CpG markers) in our model came from genomic location 10p12.2, which contains cg10804656, a site that has never been sequenced before but was mentioned by Florath et al. [146], who found it was associated with age in blood samples, based on HM450K data. The results of our sequencing results suggest that this region is going to be a promising location to focus on, as it contains a number of human age predictors.

The quadratic model described here explained 92% of the total variation in DNAm levels in a data set of 100 saliva samples, with age prediction accuracy of 3.4 years (MAD). This model was then validated on an independent data set of 68 saliva samples, resulting in a MAD of 5.3 years. Developing an age estimation assay on the MiSeq® platform is important, as in the near future it is likely to be the main DNA profiling technology in use in the majority of forensic laboratories. Although the sequencing coverage was above the recommended number of reads (1000) in all loci, even those samples with low quantity and/or quality, age estimation appears to be most reliable in saliva samples with DNA input >10ng/μL and DIN values >3. This finding does not contradict other studies who found that the minimum requirement for age estimation is 2.5ng of DNA, as this refers to the amount of bisulfite converted DNA, and not the DNA input before the bisulfite treatment [148,157]. Thus, the low prediction accuracy in low quantity/quality

samples may due to the bisulfite conversion step, which requires at least 500ng-1μg of gDNA in a maximum volume of 40μL to fully convert the unmethylated cytosine residues into uracil [155]. Therefore, an important step for the advancement of age estimation in forensic cases would be the development of a bisulfite treatment that is sensitive enough to deal with the small quantities of DNA obtained from forensic samples.

In order to evaluate the performance of our model, the model constructed by Hong et al. (2017) was used as a benchmark, as it had the highest R^2 (0.969), and the lowest MAD value (3.15 years based on testing data assayed using SNaPshot minisequencing) among all published saliva-specific age-prediction studies [157]. In this study, the seven CpG markers identified by Hong et al. (2017) were sequenced on the Illumina MiSeq®, and their model was reconstructed based on data from 100 saliva samples and validated on an independent data set of 68 saliva samples. Their model explained 68% of the total variation in DNAm levels, with four CpG markers (cg18384097, cg00481951, cg12757011, and cg07547549) being insignificant predictors of age (*P*-value >0.05) in the model. The prediction accuracy based on the 68 saliva samples was 7.5 years (MAD), which was 4.34 years higher than the value they reported based on the SNaPshot sequencing they conducted. One possible explanation for this result is that their CpG markers could have been over-fitted to the samples and the population being used for training and testing their model. Thus, when it was validated here using a sample that was not from this sample population, and which had a different sample size and age range, the deviation became more apparent. Finally, based on our data set of 168 saliva samples sequenced on the MiSeq® system, our model outperformed the model constructed by Hong et al. (2017), in terms of the amount of variation in DNAm levels explained by the model (92%), and the age prediction accuracy as measured by MAD (5.3 years).

4.7 Conclusion

The aim of this study was to identify saliva-specific AR CpG markers by implementing the most appropriate statistical methods and modelling systems, in order to further enhance the accuracy of age-prediction models in saliva samples. Initially, nine candidate markers were identified *in silico* using 54 DNAm profiles from saliva samples assayed on the Illumina HM450K BeadChip (Chapter 3), and then were validated by targeted bisulfite sequencing of another 192 saliva samples on the Illumina MiSeq® platform. All CpG sites on the sequenced amplicons were tested for age association, including sites adjacent to the target CpG sites, which resulted in the identification of additional AR markers with stronger associations with age. The best subset of these markers was selected by stepwise regression and then modelled using a quadratic (non-linear) modelling system. The model consisted of ten different AR CpG markers (cg00573770 in *ZEB2*, CpG16-CpG19, and CpG21 in genomic location 10p12.2, CpG5 and CpG7 in genomic location 1q32.2, and CpG24 and CpG27 in *ELOVL2*) with age prediction accuracy, based on an independent testing data set of 68 samples, of 5.3 years (MAD). This model could therefore be useful for providing intelligence to forensic investigations about the age of unidentified donors of saliva samples left at crime scenes. The model was compared with the model constructed by Hong et al. (2017), which was reconstructed using the same samples sequenced on the Illumina MiSeq® (a training data set of 100 samples and a testing data set of 68 samples), and the results showed a lower prediction accuracy (MAD 7.5 years) compared to that reported based on SNaPshot minisequencing (MAD 3.15 years). Summary of the findings in this chapter is illustrated in Figure 4.7. Since next generation sequencing platforms, particularly the MiSeq® platform, are likely to dominate forensic laboratories in the future, the quadratic model reported here could relatively easily be integrated into forensic laboratories in order to estimate age from saliva samples containing at least 10ng

of genomic DNA. On the other hand, the model can be also used on other analysis platforms such EpiTYPER, and/or SNaPshot, however, this requires retraining the model and then retesting it on the DNAm readings from the new assay system.

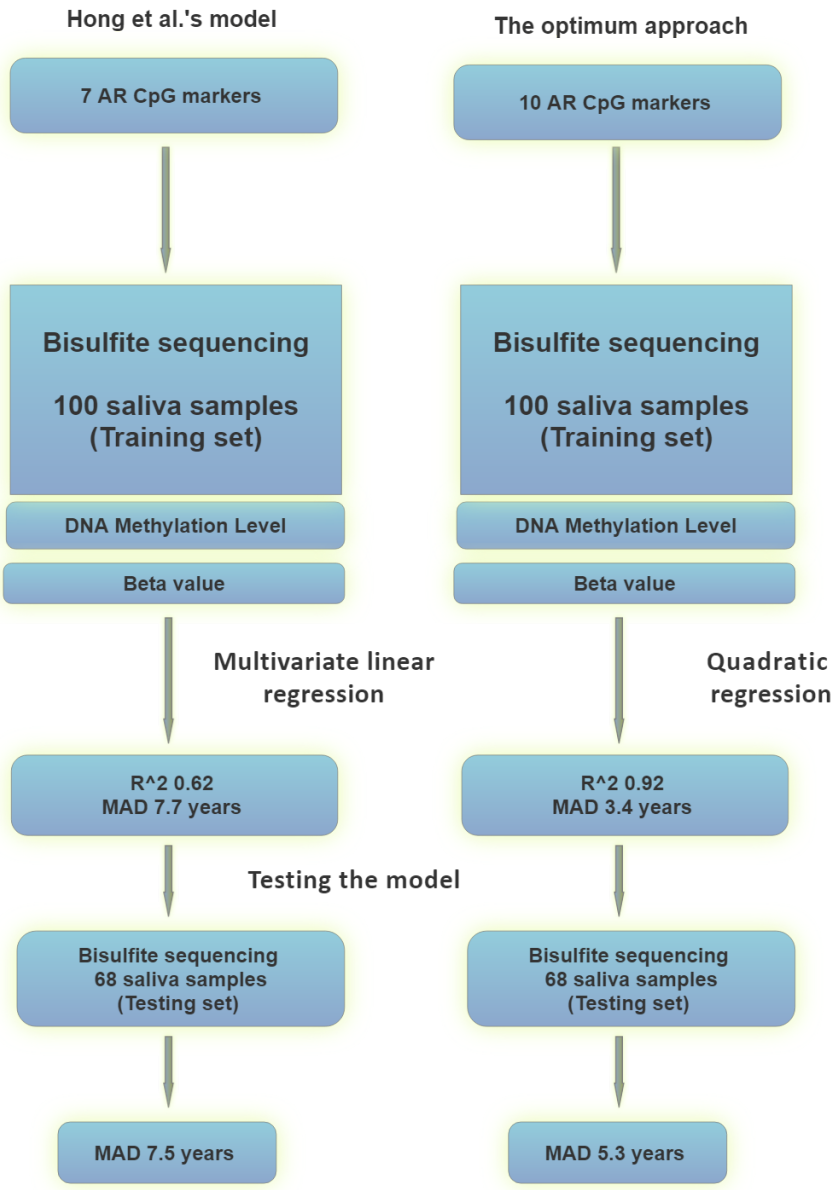

Figure 4.7 Schematic diagram showing the comparison between the HM450K model constructed using the bisulfite sequencing results using Hong et al's model.

Chapter 5: Identifying blood-specific age-related DNA methylation markers on the Illumina MethylationEPIC® BeadChip

5.1 Introduction

As discussed in Section 1.2.6, technologies to analyse DNA methylation (DNAm) in a gene-specific and genome-wide manner have developed significantly in recent years. For instance, gene-specific assays such as EpiTect, SNaPshot, EpiTYPER and targeted bisulfite sequencing have become prevalent in DNAm-related studies for their sensitive and reliable quantification of the DNAm level [150,157,191,192]. However, genome-wide assays that provide the opportunity to quantify methylation level at a single base level, such as the Illumina Infinium HumanMethylation BeadChip technology, have become the main choice for many research groups carrying out epigenome-wide association studies (EWAS). As described in Section 1.2.6.6, the introduction of two Illumina HumanMethylation BeadChips, namely the HumanMethylation27 (HM27K), and HumanMethylation450 (HM450K) BeadChips, was crucial for identifying a huge number of AR CpG sites and genes, many of which have been reported in the literature. In addition, the public genomic databases have become a rich source of epigenome-wide DNAm data, from a large body of epigenetic studies based on different human tissues [193].

Recently, a new array, the Illumina MethylationEPIC® (EPIC) BeadChip was introduced, containing over 860,000 probes, nearly double the number on the HM450K. Not all the probes on the HM450K BeadChip were included in the new EPIC BeadChip; ~90% of the HM450K probes were included, but others were

removed as a result of reports of poor performance [194]. As described in Section 1.2.6.6, the newly added probes provide a higher coverage of various genomic regions. The EPIC BeadChip is a promising tool to further our understanding of DNAm mechanisms in human development, disease, and also it could aid forensic science in offering more reliable age estimation.

5.2 Aims

This overall aim of this research was to take a comprehensive evaluation of blood-specific AR CpG sites found on the new EPIC BeadChip, and also to identify the genes associated with those AR CpG sites. This will provide new insights for epigenetic forensic researchers, by searching for new AR CpG sites on the EPIC BeadChip with better age prediction accuracy, which might enhance the performance of DNAm based age-prediction models, aiding forensic investigations in criminal cases where biological samples of unknown origin have been recovered.

5.3 Objectives

- DNAm profiles assay on the EPIC BeadChip were downloaded from an online genomic repository.
- Testing each probe on the EPIC BeadChip for age association using Spearman's rank correlation test.
- Developing blood-specific age prediction model from the highly AR CpG sites with a minimum number of CpG markers, which could have potential applications in forensic science.
- Using elastic net regression to build an age prediction model with the maximum number of markers that can be used for clinical purposes.

5.4 Materials and methods

All the R codes used in this Chapter can be seen in Appendix C2.

5.4.1 EPIC data sets

A total of 756 DNAm profiles, assayed on the EPIC BeadChip in individuals aged 0-88 years old, were assembled by combining three separate data sets retrieved from the National Centre for Biotechnology Information (NCBI) Gene Expression Omnibus (GEO) database using the procedures described in Section 2.3. The accession number of each data set and a brief description of the samples can be found in Table 5.1. To ensure identification of AR CpG sites was not biased towards a specific range of chronological ages, whole cord blood samples were included in this study, which represent time zero in human age terms (Figure 5.1).

Table 5.1 Description of the three data sets used in this study.

Accession number	DNA origin	n (Prop. female)	Median Age(range)	Citation
GSE103189	Whole cord blood	8 (0.38)	0 (0)	Dou et al. [195]
GSE123914	Whole blood	69 (1)	59 (51-65)	Zaimi et al. [196]
GSE116339	Whole blood	679 (0.59)	53 (23-88)	Curtis et al. [197]

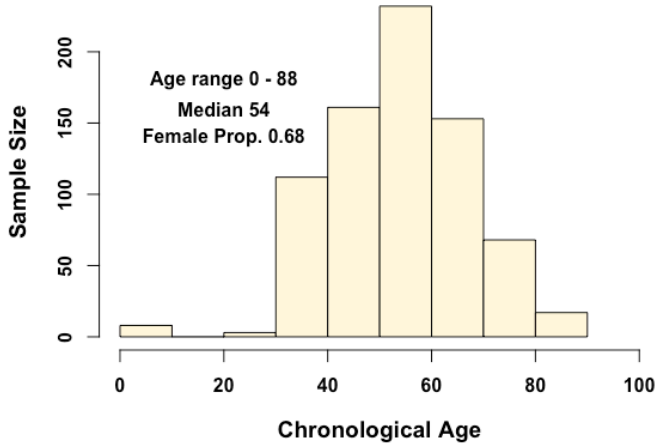

Figure 5.1 Distribution and descriptive statistics relating to the chronological ages of individuals who provided the samples used in this study.

The first data set (GSE103189) was obtained from a study conducted by Dou et al. [195], which aimed to evaluate the cell composition and DNAm differences between cord blood buffy coat and whole cord blood samples. The study revealed no significant differences between these samples, and thus they can be analytically combined and compared. The next data set (GSE123914) was obtained from a longitudinal study by Zaimi et al. [196], which aimed to examine the variation in DNAm level over a 1-year period in whole blood samples collected from 35 healthy women [196]. It was shown in this study that the median correlation coefficient between age and DNAm level at all CpG sites was 0.19, which suggests a wide variation in DNAm stability over a 1-year period. The last data set (GSE116339) contained 679 whole blood samples, retrieved from a study conducted by Curtis et al. [197], which aimed to investigate whether exposure to polybrominated biphenyl (PBB) was associated with DNAm changes in peripheral blood samples. In this study, a total of 1,890 CpG sites were identified at which DNAm was associated with total PBB levels [197].

5.4.2 EPIC data processing

The raw files of each data set were downloaded using *GEOquery* package, as described in Section 2.3. The downloaded files consisted of raw signal intensities from the red and green channels, which are the output from EPIC platform. The files were imported into R and converted into methylated and unmethylated signals by applying the *read.metharray.exp* function in the *minfi* package [173]. Although the number of CpG probes on the EPIC BeadChip is 865,918, the raw file comes with an additional 186,782 probes (giving a total of 1,052,641 probes). These additional probes were designed for quality control measures, such as background correction, negative controls, bisulfite conversion controls, and hybridisation controls [198].

As was the case on the HM450K BeadChip, probes on the EPIC BeadChip also have two chemistry designs, Infinium I and II, which possess different DNAm value distributions, introducing unwanted variation into the methylation values [98]. Thus, the two probe designs need to be normalised to render them comparable to each other, which was done using subset quantile normalisation, implemented in the *preprocessQuantile* function that is specially created for EPIC probes and implemented in the *minfi* package [85]. The same function was also used to filter out probes that did not meet the quality control measures described in Section 2.4.1. In addition, the function filtered out samples with significantly lower values in one of the two signal intensities (red/green channels) compared to the other, which is a quality control measure used to identify sample outliers. Figure 5.2 illustrates the two quality measures that were used to assess the DNAm levels in each data set, as described in Section 2.4.1. Finally, probes associated with known SNPs were removed from downstream analysis using the *dropLociWithSnps* function in the *minfi* package, as described in Section 2.4.3.

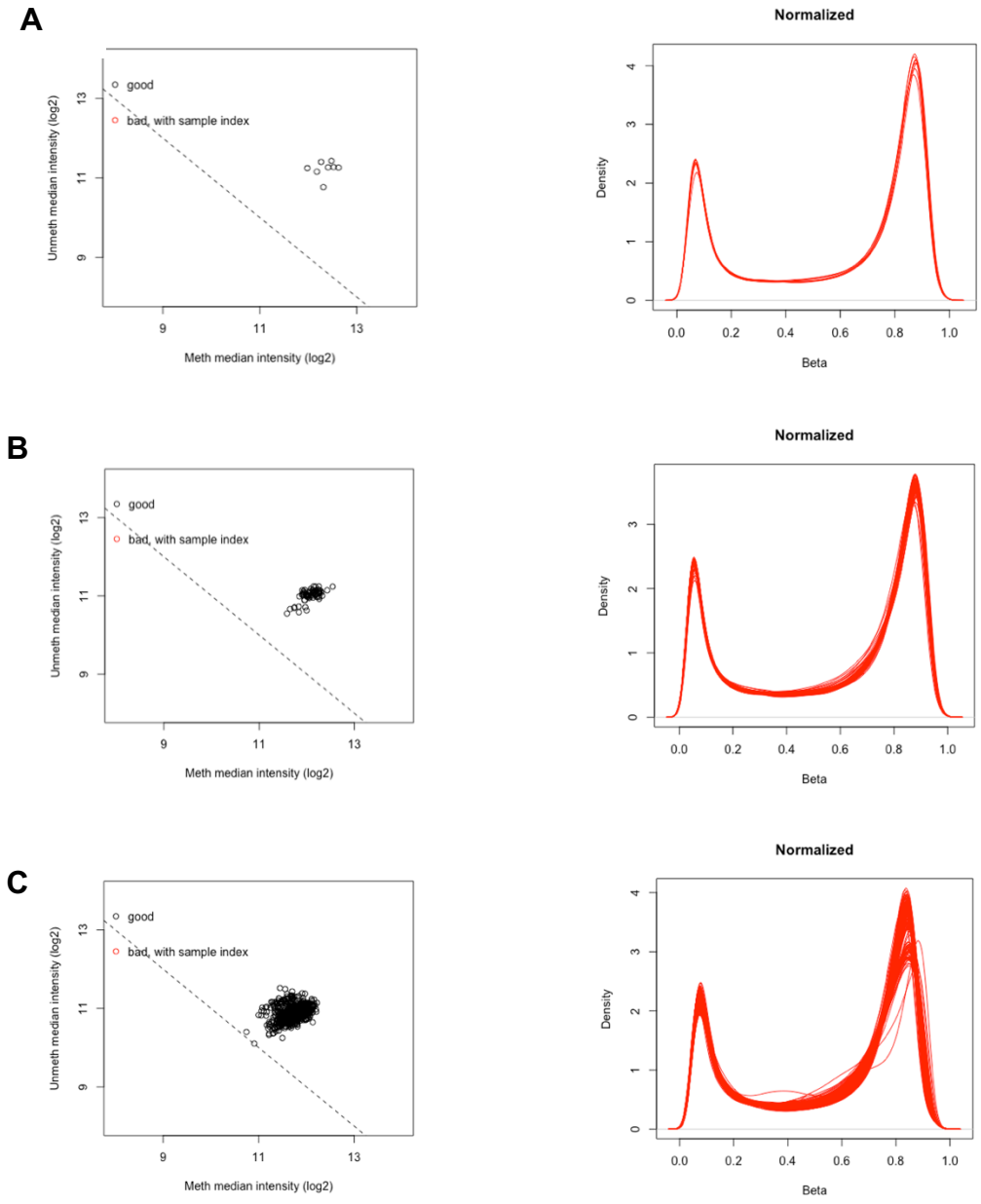

Figure 5.2 Outcomes of two quality checks for each data set used in this study; (A) GSE103189, (B) GSE123914, and (C) GSE116339. The quality measures used were (left panel) based on the log median intensity in the methylated and unmethylated channels for each sample, where good samples will be on the top half (above the dashed line). The second measure (on the right panel) is based on the distribution of Beta values in each sample, where normal samples show a bimodal distribution. From the density plot in (C), there were two samples with abnormal Beta value distributions, and these were removed from the analysis.

5.4.3 Testing for potential confounders

Given that variation in DNAm has been found to be associated with various factors such as cell type, sex, alcohol intake, smoking, obesity, and certain drugs, it is important to account for these factors as they may cause a confounding effect in EWAS [167]. The Singular Value Decomposition (SVD) method was used, as described in Section 2.4.4, to discover any hidden relationship between the samples, based on batch and sex as they were supplied by the authors. After implementing SVD on the combined data set, segregation was found between the samples, which was based on sex (Figure 5.3A). For this reason, CpG probes targeting sex chromosomes were filtered out from the downstream statistical analysis (Figure 5.3B). Batch effects were removed using the *Combat* function in R and then visualised using SVD to ensure there was no hidden structure in the data set. Figure 5.4A and B illustrate the batch effect before and after batch correction, respectively.

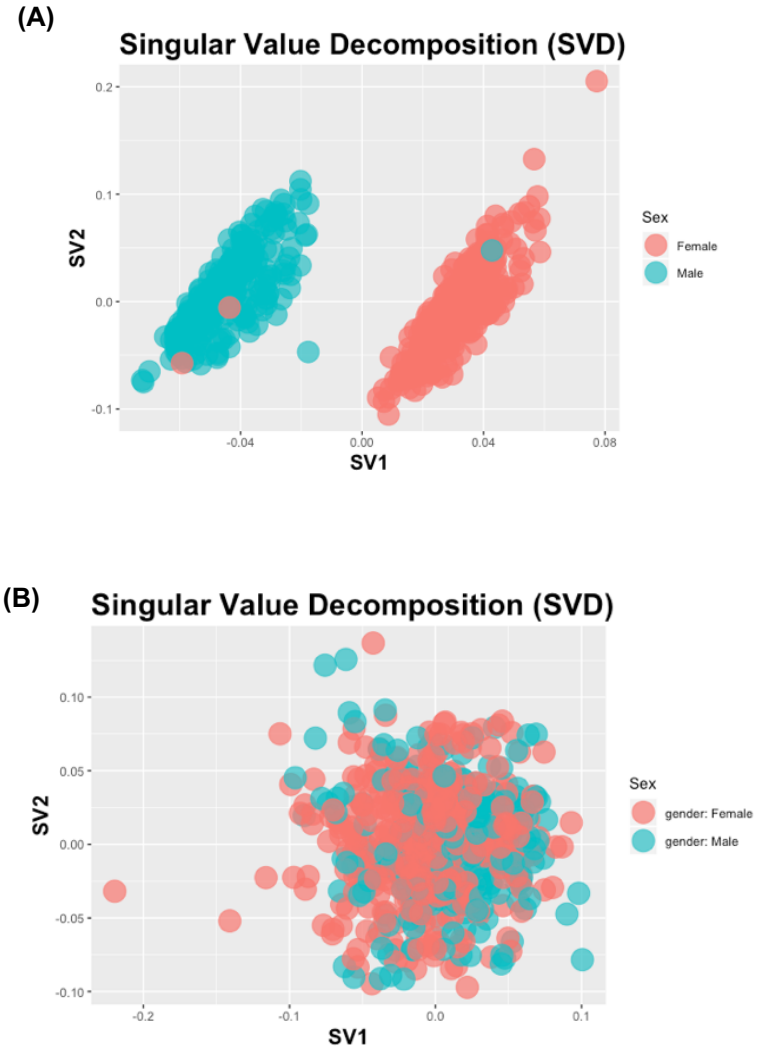

Figure 5.3 SVD analysis showing the data before (**A**) and after (**B**) removing probes targeting CpG sites on the sex chromosomes. Sex information was obtained from the original authors, and it can be seen in (A) that there were three samples wrongly labelled by the original authors.

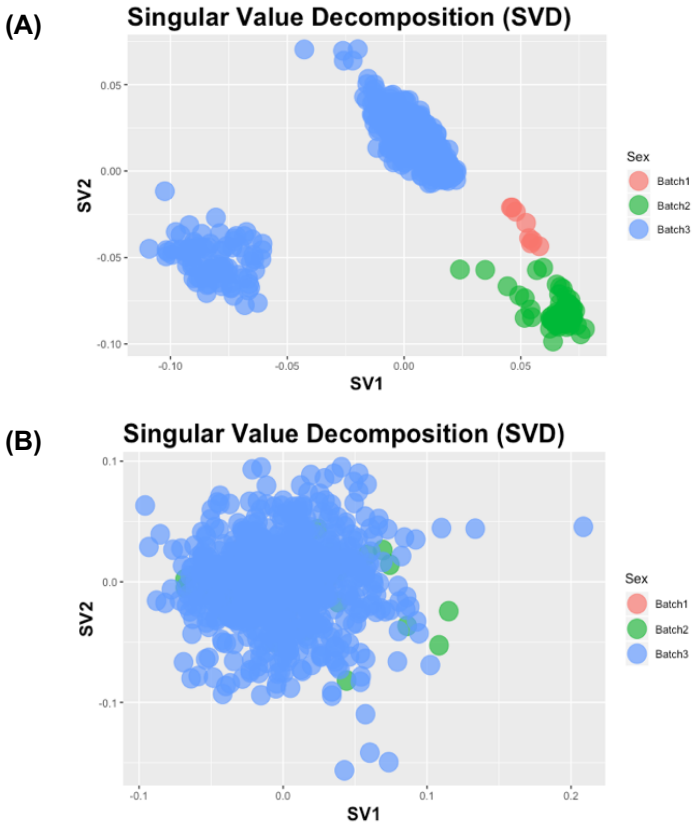

Figure 5.4 SVD analysis showing the data before (**A**) and after (**B**) removing batch effect as described in section 2.4.4.

Another potential confounder in this study was the PBB level, which was measured in the blood samples in the third data set (GSE116339). Since it has been shown that the level of PBB in blood has a significant effect on DNAm at 1,890 CpG sites, this could also have a confounding effect if it is found to be associated with chronological age. Thus, regression analysis was conducted between PBB level and chronological ages for the third data set, to reveal any linear association between them. Although the *P*-value of the test was significant (*P*-value = 1.4×10^{-9}), the R^2 was extremely low (0.05), which indicates that age only explains 5% of the variation in the level of PBB in blood. Finally, batch effects

were removed in the data set using a nonparametric empirical Bayes framework method implemented in the *Combat* function within the *SVR* package [168,199].

5.4.4 Estimating and adjusting for cell type composition

The blood cell composition was estimated using the *estimateCellCounts* function in the *minfi* package, and the identified AR CpG sites were tested for any association with cell type proportion as described in Section 2.4.5.

5.4.5 Evaluating the AR CpG sites on EPIC BeadChip

The aim of this section was to identify and evaluate AR CpG sites from the newly added probes on the EPIC BeadChip. The DNAm Beta values in the data set were converted to M values. Spearman's correlation coefficients between the DNAm level at each CpG probe and the chronological ages of the samples were calculated using R software as described in Section 2.5.1.2. The selection criteria for AR CpG probe candidates were based on two criteria: absolute (abs) Spearman's rho $\geq$0.6, and false discovery rate (FDR) at $\leq$0.05. The adjusted *P*-value was as calculated as described in Section 2.5.1.4. The genomic details of the resulting AR CpG probes were obtained using the "Infinium MethylationEPIC v1.0 B4 Manifest File," released by Illumina, which is based on the hg19/GRCh37 human genome assembly.

5.4.6 Building age prediction models

The EPIC BeadChip data were then used to build an age prediction model, to determine whether the EPIC BeadChip CpG probes have better age estimation capabilities compared to those found on the older Illumina HumanMethylation platforms such as HM27K and HM450K. The intended downstream application of the age prediction model determines the type of method that should be used to build it. For example, if the model will be used for health applications, the number

of markers in the model would not need to be limited, since the DNA in the sample would usually be in relatively large quantities. However, for forensic applications, the number of markers in the model should be kept to a minimum, as the quantity of DNA in the majority of forensic samples is low and surveying large numbers of makers requires reasonably large amounts of DNA, due to the destructive procedures involved in DNAm analysis. Therefore, two methods were used to build prediction models, elastic net regression, which creates an unlimited size model and multivariate linear regression for the smallest possible model.

5.4.7 Elastic net regression

As described in Section 2.5.3, the data set was randomly split into a training set and a testing set, whilst maintaining equal relative representation of the various age groups within the sets. The number of samples in the training set was 527, which is 70% of the original set, and 227 samples in the testing set (30%). Elastic net regression was performed as described in Section 2.5.3.3, to selected subset of CpG markers and then validated on the 227 independent testing samples.

5.4.8 Multivariate linear regression

To build an age prediction model with a minimum number of CpG markers was done in three steps as described in Section 2.5. For variable reduction, the age was linearly regressed on the DNAm level at each CpG site in the training data set, as described in Section 2.5.1.3, and then markers with $R^2 \geq 0.6$ at FDR ≤ 0.05 were selected. The selected markers were input into a stepwise regression to select the best subset for use in the age prediction model. The selected CpG markers were then combined in a multivariate linear regression to build the model, and then validated on the testing data set. The model was re-evaluated by bootstrap analysis, to ensure its prediction robustness as described in Section 2.5.4.

5.5 Results

5.5.1 EPIC data sets

The purpose of the work reported in this Chapter was to identify AR CpG markers on the EPIC BeadChip. The analysis initially included 756 samples from three different data sets, however two blood samples (GSM3228582, and GSM3228722) from data set GSE116339 had abnormal Beta value distributions, as shown in the density plot in Figure 5.2C, and thus were removed from the downstream analysis. The number of samples remaining for analysis was 754 samples, and the number of CpG sites after probe filtration was 816,127 probes. Testing for confounding variables was performed by examining how PBB level variation can be explained by age. The results (Table 5.2) showed that age only explains 5% (*P*-value $<1.4 \times 10^{-9}$) of the variation in PBB levels in blood. Therefore, the PBB level was considered not to be a confounding variable and was ignored.

Table 5.2 Linear regression between PBB level in each sample in the third data set and chronological age of individual donors.

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

5.5.2 Estimating and adjusting for cell type composition

The composition of different cell types in each sample was estimated and then tested for association with chronological age. As can be seen in Figure 5.5, CD4+ T cells, and natural killer (NK) cells had the strongest correlation with age (Spearman's rho = -0.35 and 0.32 respectively) compared to the other cell types (monocytes, CD4+, granulocytes, and B-cells), which gave rho values of ≤ 0.19. Therefore, if not adjusted, the change in DNAm level at AR CpG sites could be

explained by the change in cell composition with age, rather than by aging in individuals. For this reason, and to avoid identifying false positive AR markers, each identified AR CpG site in this study was adjusted for cell composition using multivariate linear regression.

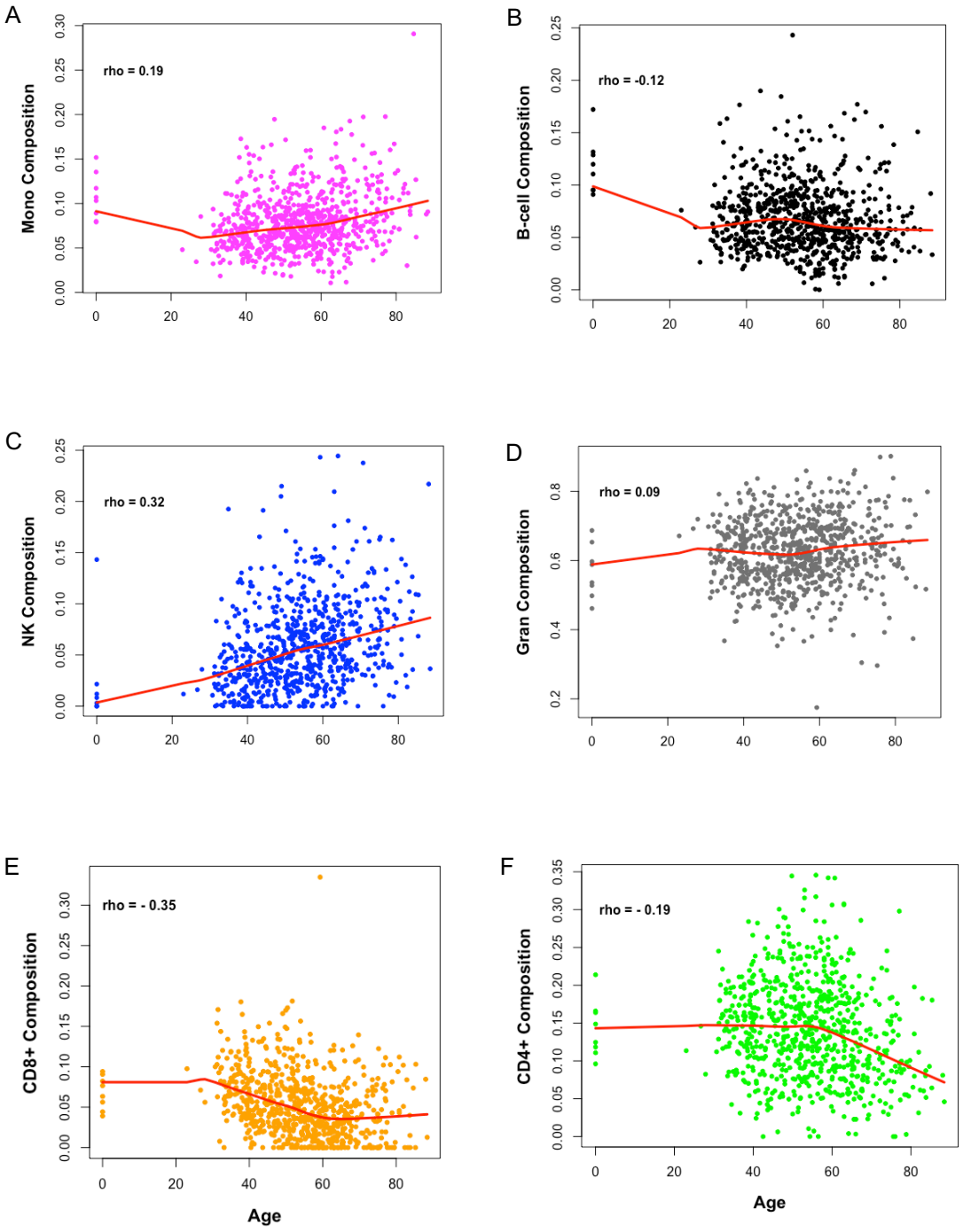

Figure 5.5 Change in blood cell composition with age. The estimated proportions of the six blood cells based on DNAm pattern [172]; (A) monocytes (Mono), (B) B cells, (C) natural killer (NK) cells, (D) granulocytes (Gran), (E) CD8+ T cells, and (F) CD4+ T cells in the samples are plotted against age. Spearman's correlation coefficients are reported for each composition proportion estimate and age. The red lines are weighted regression (Loess) lines fitted to the data.

5.5.3 AR CpG markers on the EPIC BeadChip

AR CpG sites were selected on the basis of the Spearman's rank correlation test between chronological age and DNAm level, using M values. The cut-off value for selecting AR markers was an absolute Spearman's coefficient (rho) >0.6 at FDR <0.05, as recommended by various studies [8,94,145]. A total of 52 AR CpG sites passed these conditions (Figure 5.6A), 19 of which were positively correlated (hypermethylated) and 33 negatively correlated (hypomethylated) with age (Table 5.3). The results of the cell type composition test showed that the change of DNAm level at the 52 AR CpG sites is due to aging rather than cell type composition.

The AR CpG sites with the top two highest correlation coefficients, were located in the *ELOVL2* gene, which is the most prominent gene associated with age in various tissues, as found in a number of studies (Figure 5.6B) [148,152,189]. Many of the markers identified were also identified in other studies that used a similar study design but using the Illumina HM450K BeadChip. For example, of the nine AR markers discovered by Garagnani et al. (2012), five were also identified in this study. However, of the remaining four sites, one was dropped by SNP filtration and three had abs rho <0.5. In another study, Florath et al. (2015) identified 162 AR CpG sites, of which ten were absent from the EPIC BeadChip, two were dropped after SNP filtration, and only 53 were found with abs rho >0.5. In comparing the correlation coefficients of the same AR probes on the HM450K and EPIC platforms, it was observed that the magnitude of the coefficient values was smaller on the EPIC platform, and for some probes their age association is no longer observed. For instance, nine markers identified by Xu et al. (2015) as highly AR CpG sites (with abs rho of at least 0.8) on the HM450K platform, were found to have abs rho <0.38 on the EPIC BeadChip, which is under the threshold for significant association with age. Finally, the results of the cell type composition

test, showed that the change of DNAm level at the identified 52 AR CpG sites is due to aging rather than cell type composition.

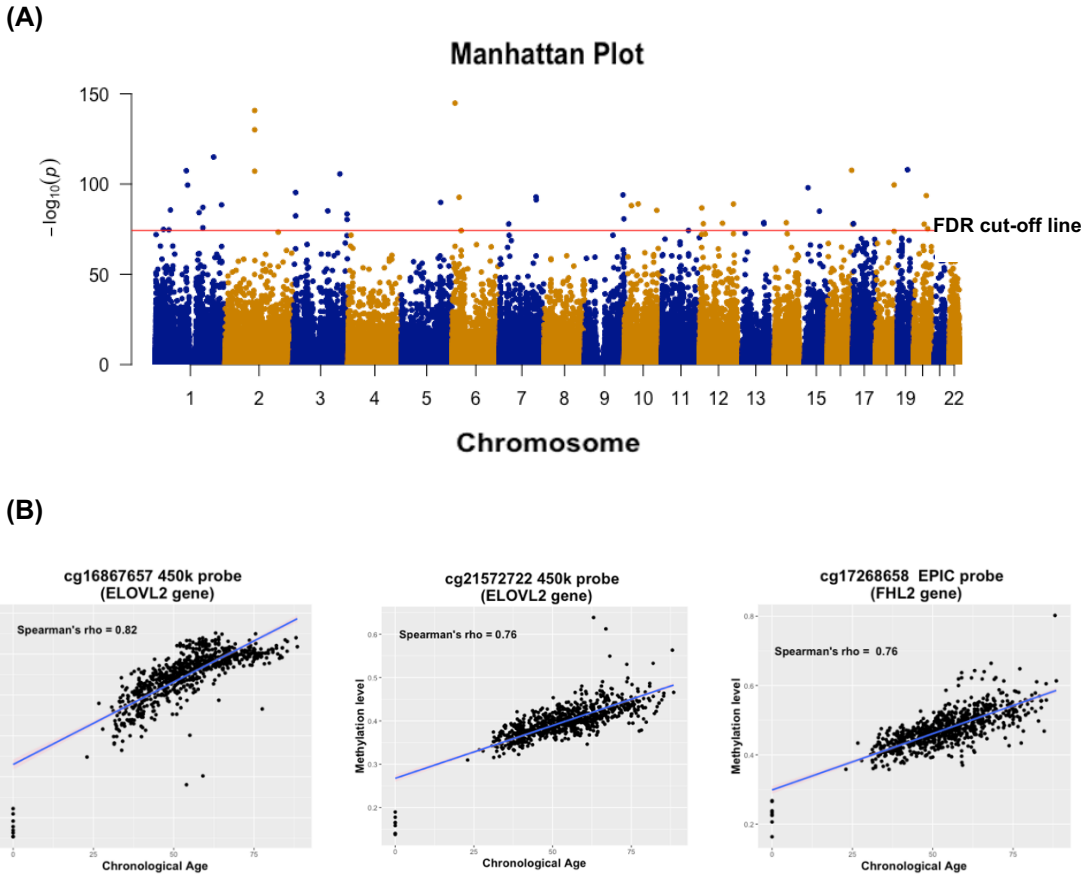

Figure 5.6 (A) Manhattan plot of P-values from Spearman's correlation test between DNAm level at each CpG site and chronological ages in the data set. (B) Scatter plots for the top three AR CpG sites found on the EPIC BeadChip.

Table 5.3 AR CpG sites found on the Illumina EPIC BeadChip, identified by Spearman's rank correlation test with cut-off value of abs rho >0.6 at FDR <0.05. Probes are sorted from the highest positively to the highest negatively correlated with age.

Probe ID	UCSC[1] Ref. Gene name	Probe type	Chr.[2]	Pos.[3]	Spearman's rho
cg16867657	*ELOVL2*	HM450K	chr6	11044877	0.82
cg17268658	*FHL2*	EPIC	chr2	106015745	0.76
cg21572722	*ELOVL2*	HM450K	chr6	11044894	0.76
cg06639320	*FHL2*	HM450K	chr2	106015739	0.74
cg22454769	*FHL2*	HM450K	chr2	106015767	0.69
cg17110586	*unknown*	HM450K	chr19	36454623	0.69
cg04875128	*OTUD7A*	HM450K	chr15	31775895	0.67
cg24866418	*LHFPL4*	EPIC	chr3	9594082	0.66
cg13649056	*unknown*	HM450K	chr9	136474626	0.66
cg07547549	*SLC12A5*	HM450K	chr20	44658225	0.66
cg23500537	*unknown*	HM450K	chr5	140419819	0.65
cg06493994	*SCGN*	HM27K	chr6	25652602	0.65
cg08097417	*KLF14*	HM450K	chr7	130419133	0.65
cg13206721	*GPR158*	EPIC	chr10	25463350	0.64
cg06784991	*ZYG11A*	HM450K	chr1	53308768	0.63
cg12841266	*LHFPL4*	EPIC	chr3	9594093	0.63
cg27099280	*CELF6*	EPIC	chr15	72612204	0.63
cg03032497	*unknown*	HM450K	chr14	61108227	0.61
cg25410668	*RPA2*	HM450K	chr1	28241577	0.6
cg03650729	*TAL1*	EPIC	chr1	47692625	-0.6
cg18651026	*COL11A2*	HM450K	chr6	33140660	-0.6
cg25371036	*AMOTL1*	HM450K	chr11	94500749	-0.6
cg15109150	*FAM65C*	EPIC	chr20	49308830	-0.6
cg09240238	*LOC730668*	EPIC	chr22	46402573	-0.6
cg16054275	*F5*	HM450K	chr1	169556022	-0.61
cg16789844	*PDE1C*	EPIC	chr7	32339213	-0.61
cg01855540	*DUSP16*	EPIC	chr12	12716653	-0.61
cg11649376	*ACSS3*	HM450K	chr12	81473234	-0.61
cg26685941	*ABCC4*	HM450K	chr13	95952902	-0.61
cg05412028	*ABCC4*	HM450K	chr13	95952937	-0.61

cg04503319	*ANKRD11*	HM450K	chr16	89368367	-0.61
cg17457912	*C17orf91*	HM450K	chr17	1617102	-0.61
cg17015290	*KIAA1755*	EPIC	chr20	36850842	-0.61
cg03776853	*unknown*	EPIC	chr22	36461577	-0.61
cg23719650	*unknown*	EPIC	chr3	193988507	-0.62
cg00876267	*unknown*	HM450K	chr9	139588516	-0.62
cg05308819	*unknown*	HM450K	chr1	155959156	-0.63
cg25167618	*SLC12A8*	EPIC	chr3	124840296	-0.63
cg11218872	*unknown*	EPIC	chr3	193988737	-0.63
cg08587685	*ABLIM1*	EPIC	chr10	116392206	-0.63
cg08745595	*F5*	EPIC	chr1	169556012	-0.64
cg09809672	*EDARADD*	HM27K	chr1	236557682	-0.64
cg22796704	*ARHGAP22*	HM450K	chr10	49673534	-0.64
cg05179292	*C1R*	EPIC	chr12	7244621	-0.64
cg17403084	*PXN*	EPIC	chr12	120704034	-0.64
cg03473532	*MKLN1*	HM450K	chr7	131008743	-0.65
cg16008966	*unknown*	HM450K	chr1	114761794	-0.67
cg13552692	*CCDC102B*	EPIC	chr18	66389447	-0.67
cg18933331	*unknown*	HM450K	chr1	110186418	-0.69
cg07323488	*EGFEM1P*	EPIC	chr3	168185313	-0.69
cg07082267	*unknown*	HM450K	chr16	85429035	-0.69
cg10501210	*unknown*	HM450K	chr1	207997020	-0.71

[1] Based on UCSC Genome Browser database
[2] Chromosome
[3] Position based on the human assembly GRCh37, also known as hg19.

5.5.4 Novel AR CpG sites on EPIC BeadChip

From the 52 AR CpG sites identified in this study, 21 were from the newly added probes on the EPIC BeadChip, and so these can be considered novel AR CpG sites since they have not been reported in the literature before (Table 5.4). In addition, they map to 18 genes, nine of which (*LHFPL4, SLC12A8, EGFEM1P, GPR158, TAL1, KIAA1755, LOC730668, DUSP16,* and *FAM65C*) have also never been reported in the literature as being associated with age. The majority

of these sites (16) were hypomethylated, and five were hypermethylated with age (Figure 5.7). The highest positively correlated novel AR CpG site was cg17268658 with rho = 0.76 (*P*-value $1.9x10^{-141}$), associated with the *FHL2* gene, which has been reported in many age association studies [146,179,200]. The highest negatively correlated CpG site was cg07323488 with rho = -0.69 (*P*-value $2.6x10^{-106}$), which is linked to a pseudogene known as *EGFEM1P*. Scatter plots of age versus DNAm level for the top four most highly correlated AR CpG sites can be seen in Figure 5.8.

Table 5.4 The 21 novel AR CpG sites from the newly added probes on the Illumina MethylationEPIC BeadChip, identified by Spearman's correlation test with a cut-off value of abs rho >0.6 at FDR <0.05. Probes are sorted from the highest positively to the highest negatively correlated with age.

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

[1] Based on UCSC Genome Browser database

[2] Chromosome

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

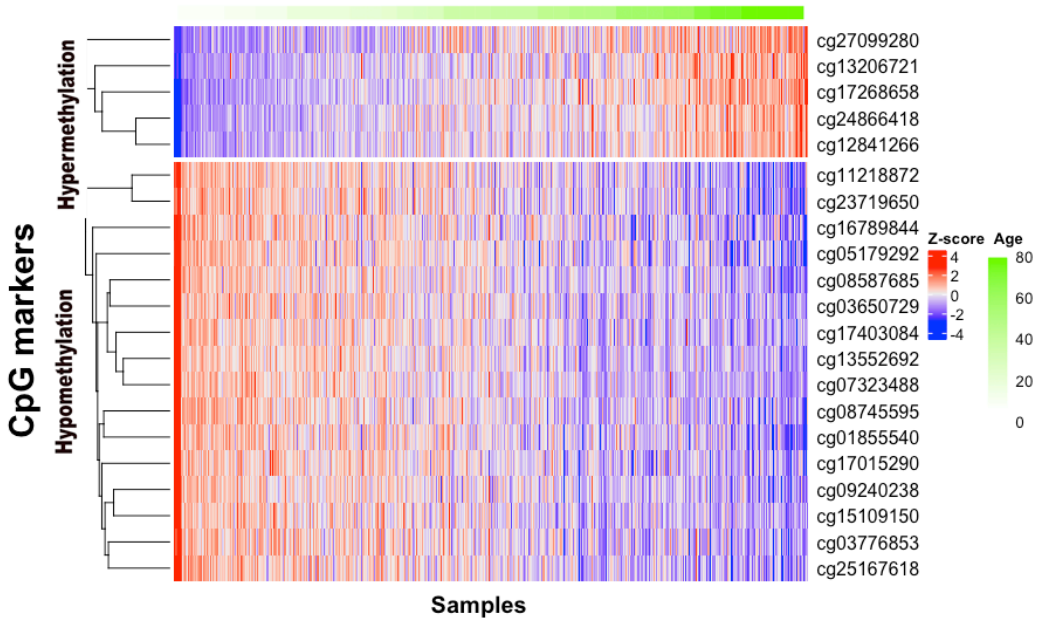

Figure 5.7 Heat map illustrating methylation level at 21 novel AR CpG markers for each sample in the training data set, ordered by chronological age across samples. The methylation level in each sample is indicated by the Z-score, where red indicates a site is hypermethylated and blue is hypomethylated. Hierarchical clustering of the CpG markers is presented on the left-hand side of the heat map.

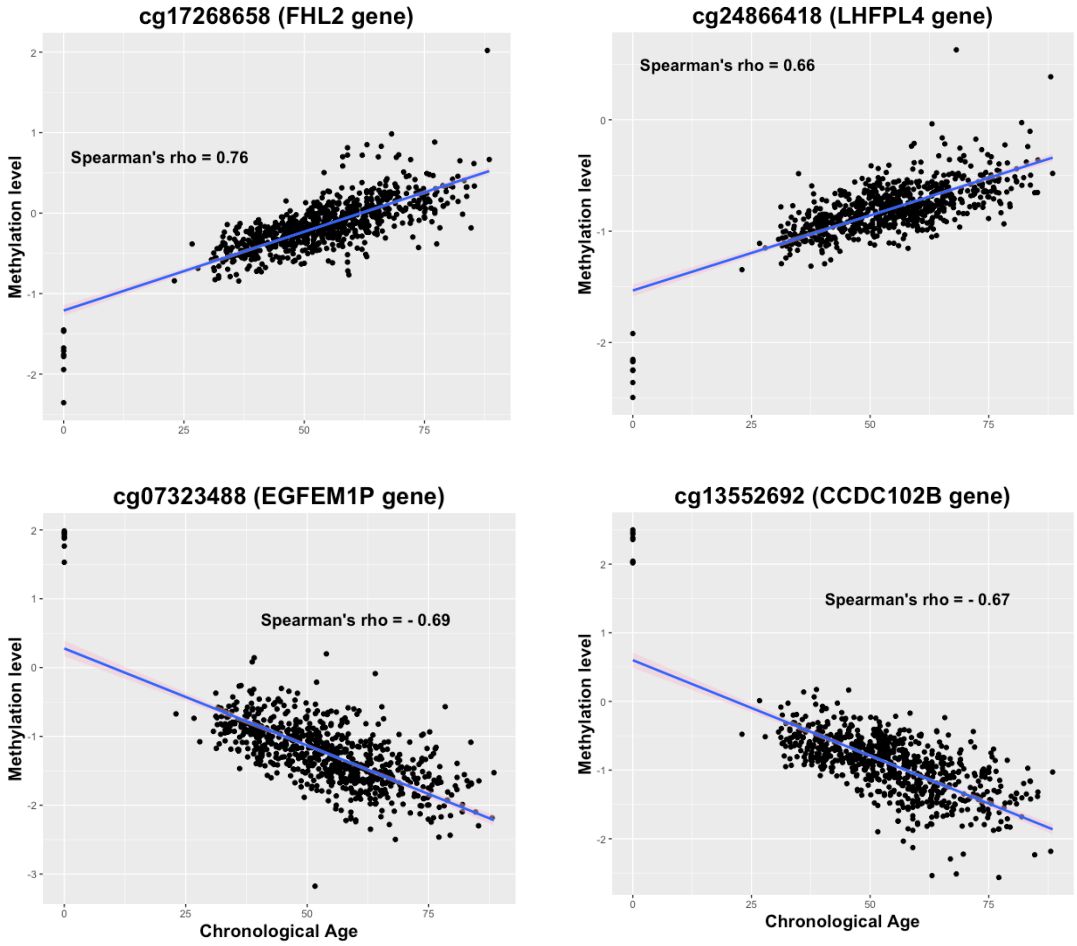

Figure 5.8 Scatter plots of M values versus chronological age for the top two positively and two negatively correlated AR CpG sites from the newly added probes on the EPIC BeadChip.

5.5.5 Blood specific age prediction models

5.5.5.1 Elastic net regression model

In several previous studies where age has been modelled as a function of CpG methylation status for sites in the genome, elastic net regression has been used to perform both feature selection and model building [106,136]. Elastic net regression is ideal for constructing predictive models with large number of markers that have high prediction accuracy [136]. Elastic net regression was performed on the data set of 816,127 CpG sites and automatically selected 425

AR CpG sites (Table A1) across 527 blood samples. From these markers, 160 AR CpG sites were from the newly added probes on EPIC BeadChip. The prediction model containing the selected markers was evaluated on the training data set using one round of ten-fold cross-validation. The prediction accuracy of the model containing the 425 CpG markers based on the training data set was equal to 0.68 years (MAD). Furthermore, its performance was evaluated using an independent validation data set containing 227 blood samples, which resulted in an MAD of 2.6 years, and a Pearson's correlation coefficient (r) between the predicted and chronological age of 0.97 (95% CI: 0.96–0.98).

5.5.5.2 Multivariate linear regression model

Although the model constructed using elastic net regression had very high prediction accuracy, the number of markers in the model (425 CpG sites) is very high. An assay containing this number of sites could not be implemented in a forensic science context, where assays with limited numbers of markers are required. Thus, to build an age prediction model with the minimum number of AR CpG markers, three steps were carried out (variable reduction, selection, and building the model). The first step was regressing age on DNAm level at each CpG site in the training data set, and then markers with R^2 >0.6 at FDR $\leq$0.05 were selected. Ten CpG markers met this condition, and only two of them were from the newly added probes on the EPIC BeadChip. The second step was to select the best subset of these sites to build an age prediction model. The stepwise regression selected six markers as the best subset for age prediction, which contained only one newly added EPIC BeadChip probe (Table 5.5). This model explained 81% of the total DNAm levels in the blood samples with prediction accuracy of 4.5 years MAD based on the training data set, and 4.6 years based on the testing data set. The accuracy rate based on bootstrap analysis was 4.5 years, with 95% confidence intervals (CI) of 4.56 to 4.57 years. The correlation (r) between predicted age and chronological age was 0.9 (95%

CI: 0.88 – 0.93) (Figure 5.9). Finally, to avoid gender bias in age prediction, male and female samples in the testing data were separated and their MAD values were assessed separately, to determine whether the difference between them was significant. A t-test showed that there was a non-significant (*P*-value = 0.3) difference in the prediction accuracy for males (MAD = 4.4 years) compared to females (MAD = 4.9 years).

Table 5.5 Multivariate linear regression analysis between DNAm levels at six CpG sites and age in the training data set. The CpG marker in bold is the only site exclusively found on the EPIC BeadChip.

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

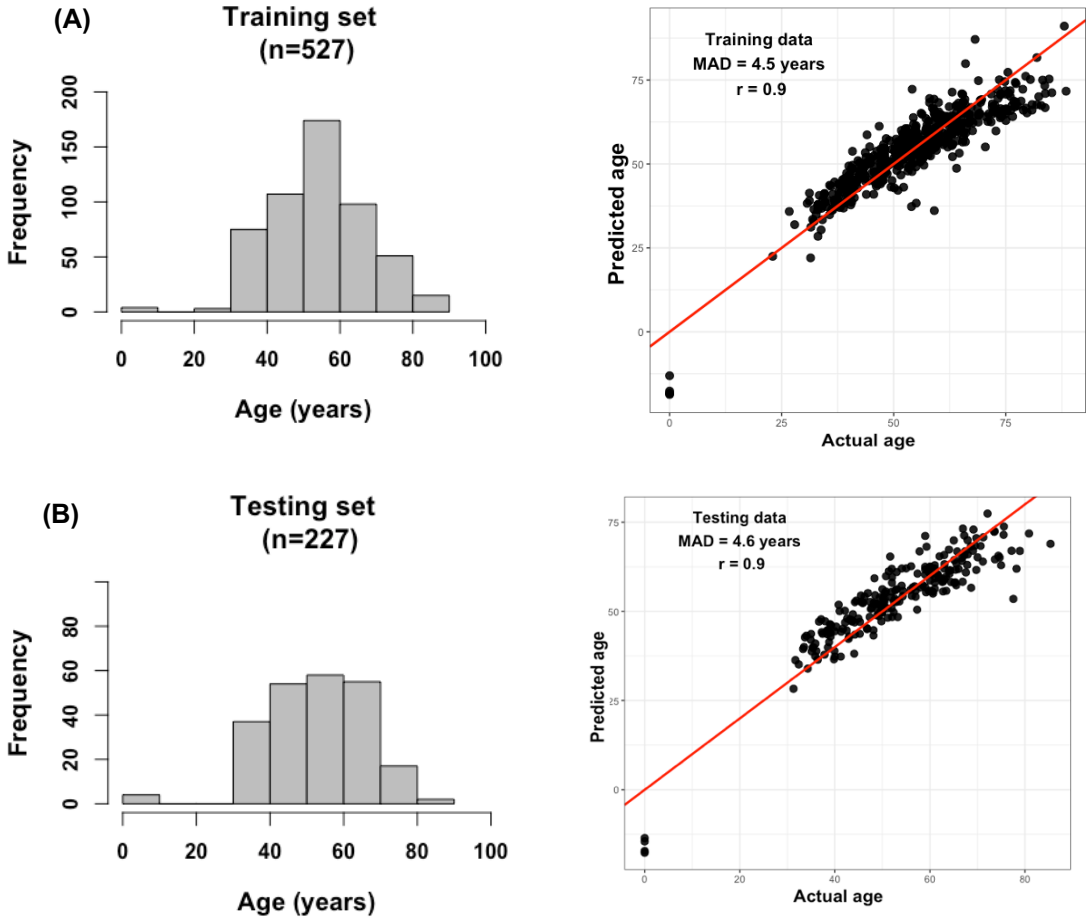

Figure 5.9 Performance of the multivariate linear regression model consisting of six AR CpG markers. The histograms show age range in the data and the scatter plots show the accuracy of the model in (**A**) the training set of 527 samples, and (**B**) the testing set of 227 samples.

5.6 Discussion

In this present study, 754 whole blood DNAm profiles assayed on the Illumina EPIC BeadChip were examined, and 52 AR CpG sites were found, of which 31 were from the HM27K and HM450K platforms, and 21 were novel probes added to the EPIC BeadChip. Apart from these 21 novel sites, all identified AR sites were previously found by different studies that used blood DNAm profiles assayed on the HM450K BeadChip. However, their correlation coefficient values on the EPIC BeadChip were lower compared to their values on the HM450K array. Although these differences between studies are expected, and may be due to differences in sample size, age range, and the ethnicity of individuals, an unexpected outcome was that some AR CpG sites with high correlation coefficients on the HM450K platform were not associated with age on the EPIC BeadChip. Probes that completely lost their association with age in this study were originally identified in studies with sample sizes below the recommended, which is 100 samples [106]. For example, the number of samples in Xu et al. (2015) was 16 samples, and all their identified AR CpG sites were found to be weakly abs rho <0.38) associated with age in our study. This suggests that AR probes identified in studies with a small sample size would be more likely to be sample-specific than tissue-specific.

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

genomic locations within these same genes, significant age association has been identified. The final gene identified in this study was from a gene newly targeted on the EPIC BeadChip, *EGFEM1P*.

From the nine newly identified AR genes, two become hypermethylated with age (*LHFPL4* and *GPR158*), and seven become hypomethylated with age (*SLC12A8, TAL1, KIAA1755, LOC730668, DUSP16, EGFEM1P and FAM65C*). *LHFPL4* is located on chromosome three and encodes a subset of the superfamily of tetraspan transmembrane proteins, which is a critical regulator of postsynaptic GABA clustering in hippocampal pyramidal neurons [201]. Its differential methylation has previously been found to be a biomarker for the early detection of cervical cancer [202]. *GPR158* is located on chromosome ten and encodes a G protein-coupled receptor, which is implicated in many physiological and disease processes [203]. The protein encoded by *DUSP16* on chromosome 12 is a dual specificity phosphatase, implicated in various cellular processes including cell differentiation [204]. *EGFEM1P* is a pseudogene located on chromosome three and was shown by one study to be differentially methylated in obese asthmatics, and by another to be significantly hypermethylated in patients with chronic lymphatic leukemia [205,206]. *KIAA1755* encodes for an uncharacterised protein, and contains a SNP variant (rs6127471) that has been associated with individuals who have increased heart rate [207,208]. *FAM65C* encodes a protein that is a member of extracellular complex that generally regulates cellular processes in response to stimuli, but its main molecular function is still obscure [209]. The hypomethylated CpG site linked to *LOC730668*, which is a Dynein Heavy Chain-Like pseudogene located on chromosome 22, has been reported in two different studies to be differentially hypomethylated in individuals with temporal lobe epilepsy, and in individuals with psoriatic epidermis [210,211].

Studying genes without knowing how they correlate with chronological age could introduce false positives. Thus, if not adjusted, age could be a potential

confounder in case-control studies. For example, a study conducted by Fluhr et al. [212] found *SLC12A8* (which was significantly hypermethylated with age in this study) to be differentially methylated in children with juvenile myelomonocytic leukemia (JMML). However, this study was based on children with JMML versus healthy adults, and the AR markers would be expected to be differentially methylated between children and adults regardless of JMML-status. Another example is the *TAL1* gene located on chromosome one, which encodes a protein that has been associated with Precursor T-Cell Acute Lymphoblastic Leukaemia and T-Cell Childhood Acute Lymphocytic Leukaemia. In a study conducted by Musialik et al. [213], methylation levels in the promoter of the *TAL1* gene were found to be slightly elevated in patients aged ≥ ten years with Precursor B-cell acute lymphoblastic leukaemia, suggesting it was a potential predictor for the disease. Again, since methylation level was not adjusted for age, this association could be confounded by age.

Recently, the search for AR CpG sites and attempts to build DNAm-based age prediction models with high accuracy have been of major interest within the fields of forensic science, and epidemiology. For this reason, this study examined whether the EPIC BeadChip contains AR CpG markers with a better prediction accuracy than those found on the previous Illumina platforms (HM27K and HM450K). Two methods were used to build age prediction models, elastic net regression and multivariate linear regression. The optimum model selected by elastic net regression contained a set of 425 CpG sites, 160 (38%) of which were probes that were newly added to the EPIC BeadChip. This model had a high prediction accuracy, based on the validation data set, of 2.6 years (MAD). Comparing this result with a study conducted by Hannum et al. (2013) that had a similar experimental design but used Illumina HM450K data, their prediction model, also selected by elastic net regression, consisted of 71 CpG markers with a prediction accuracy of 4.89 years (MAD). Building a prediction model for use in forensic investigations requires a small number of markers due to the minute

quantities of DNA that are frequently recovered from forensic samples [15]. A second model was therefore constructed using multivariate linear regression. The six AR CpG sites selected by this stepwise regression, which contained only one CpG marker that was newly added to the EPIC BeadChip, had a MAD value of 4.6 years based on the validation set. A review of the literature shows that the range of MAD values achieved by forensic researchers for models based on blood samples was 3.2 to 7.9 years, using two to 17 CpG markers [144,214,215]. Therefore, the prediction accuracy of data generated using the EPIC BeadChip falls within the range of MAD values reported in previous studies.

5.7 Summary and conclusions

The purpose of the study presented here was to use blood-based Illumina MethylationEPIC BeadChip data to identify AR CpG markers from probes that were new on this platform. Fifty-two AR CpG sites were identified, 21 of which were novel AR CpG sites that mapped to 18 genes, nine of which have never been reported in the literature as being associated with age (Figure 5.10). This finding will provide new insights for researchers in both clinical and forensic epigenetics. For instance, in clinical epigenetics this will allow researchers to account for the aging effect of these genes, which will significantly limit the false positives in their genome- and epigenome-wide association studies. In addition, although the newly introduced probes on the EPIC BeadChip did not improve the accuracy of age-prediction models when compared to other models reported in the literature, the new genomic locations harbouring AR CpG sites can be further studied by forensic geneticists using targeted bisulfite sequencing, which may result in the discovery of additional AR sites with high age prediction accuracy, that can be exploited for forensic purposes.

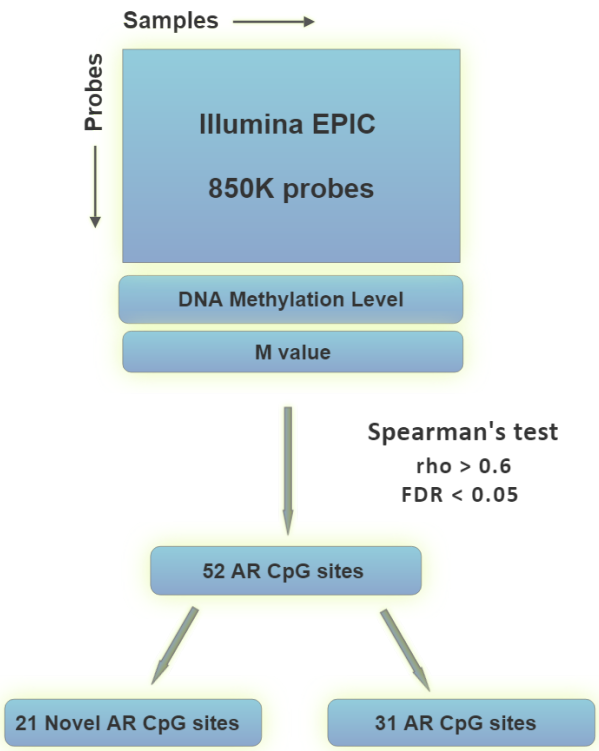

Figure 5.10 Schematic diagram summarising the main findings of this chapter.

Chapter 6: A multi-tissue age prediction model based on DNA methylation analysis

6.1 Introduction

Age-related (AR) CpG markers can be either tissue-specific or common across multiple tissues. Tissue-specific AR markers have been successfully identified for some tissues of forensic interest, namely blood, saliva, semen, and teeth [89,147,157,216]. However, using the methylation level for one set of tissue-specific AR markers to predict chronological age from other tissues has been shown to result in poor prediction accuracy [155,156]. Although some researchers have built multi-tissue age prediction models that can be used across multiple tissues, these models have several limitations if implemented in forensic science [94,106,141]. Unfortunately, the forensic literature at the time of conducting this research has only focused on the identification of CpGs correlated with age in single body fluid/tissue types, and there is no research that has focused on the identification of universal AR CpG biomarkers that can be used to predict age from tissues that are frequently recovered from the crime scenes. Constructing an age prediction model from these universal AR CpG markers (a multi-tissue age prediction model) would be beneficial in predicting the age from samples of unknown tissue types, which would bypass the necessity to first identify the tissue type, a step that is not only time-consuming but can also expose the valuable DNA evidence to chemical destruction.

Developing a multi-tissue age prediction model could be very important in terms of estimating the age of a biological sample that is recovered from a crime scene, because its identity (e.g. blood, saliva, semen, etc.) may often be unknown. Koch and Wagner [94] proposed 19 universal AR CpG markers across

tissues and built a multi-tissue age prediction model based on a data set containing five tissues (dermis, epidermis, cervical smear cells, T-cells and monocytes) profiled on Illumina HumanMethylation27 (HM27K) BeadChips. However, this research did not incorporate tissue types of forensic interest. In addition, its age prediction accuracy was low when tested on multiple tissues, with mean absolute deviation (MAD) between predicted and chronological age of 11 years. This level of accuracy would result in an age range that would make it very difficult to significantly reduce a list of suspects by age, making the assay less useful when applied in a forensic context. In an attempt to overcome some of the limitations of the Koch and Wagner (2011) study, Horvath (2013) used 51 different tissues to build a multi-tissue age prediction model. Although the MAD between predicted and chronological age for this model was very low (3.6 years), the large number of CpG markers that it contains (353) limits its practical application in forensic casework on samples that are typically low quantity and/or degraded. [157].

6.2 Aims

The aim of the current study was to identify a small subset of universal AR CpG sites that could be used to build a multi-tissue age prediction model to predict chronological age for forensic purposes, especially across tissues that are frequently recovered from crime scenes, such as blood, semen, saliva, menstrual blood, and vaginal secretions. In addition, a multi-tissue age prediction model would be beneficial, since using an existing tissue-specific age prediction model on other tissues will produce inaccurate age estimation. The reasoning behind restricting the number of CpG markers to the minimum was to facilitate the design of a PCR-based DNAm assay that only requires a small amount of starting DNA template, so that this method could be implemented in any forensic laboratory.

6.3 Objectives

- Downloading DNAm profiles of different tissues assayed on the Illumina HM27K or HM450K array platforms from different genomic repositories.
- After combining all the tissues together in one large data set, elastic net regression was implemented to build a multi-tissue age prediction models with minimum and maximum number of CpG markers.
- Demonstrating how the number of CpG markers in the age prediction model affect the prediction accuracy.

6.4 Materials and methods

All the R codes used in Chapter can be seen in Appendix C3.

6.4.1 Training data set

In total of 28 individual data sets (Table 6.1) assayed on the Illumina HM27K or HM450K BeadChip platforms were downloaded from three different genomic repositories; National Centre for Biotechnology Information (NCBI) Gene Expression Omnibus (GEO), the Cancer Genome Atlas (TCGA), and ArrayExpress, and compiled into one large training data set in R using the protocol described in Section 2.3. The training data set consisted of samples from 3,020 healthy individuals, obtained from 22 different tissues and cell types. These samples were derived from 1,484 males, and 1,352 females (the remaining 184 were from individuals of unknown gender), with ages ranging from 0 to 101 years old. The sample specifically excluded individuals with any known disease, because there are certain diseases such as cancer that affect overall DNAm

levels at CpG sites, which consequently will affect the prediction accuracy of the age prediction model [106,134,136]. Data sets retrieved from TCGA involved only normal tissues.

Table 6.1 Description of the 28 data sets used in the training data set.

No.	DNA origin	Accession no.	No. of samples (prop. female)	Platform	Median age (range)	Genomic repository
1	Blood	GSE41037	715 (0.38)	HM27K	33 (16,88)	GEO
2	Blood	GSE40279	656 (0.52)	HM450K	65 (19,100)	GEO
3	Blood PBMC	GSE36064	78 (0)	HM450K	3.1 (1,16)	GEO
4	Blood PBMC	GSE32148	48 (0.52)	HM450K	15 (3.5,76)	GEO
5	Cord blood	GSE27317	216 (0.51)	HM27K	0 (0,0)	GEO
6	Brain	GSE38873	168	HM27k	45 (20,70)	GEO
7	Breast	GSE32393	23 (1)	HM27K	46 (19,75)	GEO
8	Buccal swab	GSE25892	109 (0.61)	HM27K	15 (15,15)	GEO
9	Colon	GSE32146	24 (0.54)	HM450K	14 (3.5,19)	GEO
10	Dermal fibroblast	GSE22595	14 (1)	HM27K	20 (6,73)	GEO
11	Bone marrow	GSE17448	16 (0.38)	HM27K	52 (21,85)	GEO
12	Placenta	GSE36642	28 (1)	HM27K	0 (0,0)	GEO
13	Prostate	GSE26126	70 (0)	HM27K	61 (44,73)	GEO
14	Saliva	GSE34035	181 (0.015)	HM27K	29 (21,55)	GEO
15	Uterine Cervix	GSE30758	152 (1)	HM27K	25 (19,55)	GEO
16	Muscle	GSE38291	22 (0.55)	HM27K	66 (53,78)	GEO
17	CD4+ CD14	GSE20242	50 (0.68)	HM27K	34 (16,69)	GEO
18	Sperm	GSE26974	19 (1)	HM27K	0 (0,0)	GEO
19	Kidney	TCGA, KIRP	45 (0.3)	HM450K	66 (31,83)	TCGA
20	Colon	TCGA, COAD	37 (0.63)	HM27K	74 (43,90)	TCGA
21	Lung	TCGA, LUSC	27 (0.15)	HM27K	69 (52,83)	TCGA
22	Lung	TCGA, LUAD	24 (0.58)	HM27K	66 (51,77)	TCGA
23	Lung	TCGA, LUSC	42 (0.32)	HM450K	73 (40,85)	TCGA
24	Prostate	TCGA, PRAD	50 (0)	HM450K	63 (44,72)	TCGA
25	Stomach	TCGA, STAD	41 (0.51)	HM27K	69 (43,87)	TCGA
26	Head and Neck	TCGA, HNSC	50 (0.24)	HM450K	62 (26,87)	TCGA
27	Breast	TCGA, BRCA	27 (1)	HM27K	51 (35,88)	TCGA
28	Breast	TCGA, BRCA	88 (1)	HM450K	57(28-90)	TCGA

6.4.2 Illumina HumanMethylation data processing

The data sets from the two platforms (Illumina HM27K and HM450K) in the training data set were merged together by focusing on only the overlapping CpG probes (21,368) that are present on both platforms. Although, this will result in the elimination of CpG probes that could be strongly associated with age, this is offset by the advantage of being able to combine a large number of data sets containing different type tissues. Each of the data sets in the training data set were individually tested and samples that showed abnormal DNAm levels were removed from downstream statistical analyses. The training set was not normalised as described Section 2.4.2, instead it was normalised using a normalisation algorithm created by Horvath (2013), which was developed to improve the accuracy of the resulting age prediction model constructed by elastic net regression [106]. For this reason, Horvath's normalisation algorithm was obtained, which can be found in the supplementary materials files under the name of "Additional file 24" [106]. The normalisation algorithm consists of custom scripts that run on R software.

6.4.3 Singular value decomposition (SVD)

SVD was used to assess how the DNAm profiles observed in the different tissue types would separate them from each other, and also to identify samples that do not cluster with their original tissue type, which indicates possible outliers [168]. The SVD analysis was carried out using the *svd* function in R software, as described in Section 2.4.4.

6.4.4 Identifying universal CpG sites and building the multi-tissue age prediction model

For variable reduction and model selection, elastic net regression was implemented using the *glmnet* package in R software, as described in Section

2.5.3.3. Horvath (2013) found that it was advantageous to transform the chronological age values in the training data set before conducting elastic net regression, as this improves the prediction accuracy of the constructed age prediction model, using the following functions:

$$F(age) = \log(age + 1) - \log(20 + 1), \quad if\ age \leq 20$$

$$F(age) = \frac{(age + 1)}{(20 + 1)}, \quad if\ age > 20$$

The transformed chronological ages were regressed on Beta values for all CpG probes (21,368) in the training data set. The reason for using Beta values instead of M values, is that the process of transforming Beta to M values introduces infinite (Inf) values into the data set. These Inf values can be input into correlation and regression tests in R software, but not elastic net regression, and thus Beta values were used. As described in Section 2.5.3.3, the number of markers in the best model, which is determined by the lambda value, is automatically selected by elastic net regression. However, the lambda value can, alternatively, be controlled, and the number of markers in the final model can be manually selected. Since the aim of this study was to construct multi-tissue age prediction models for forensic purposes, lambda values that correspond to models containing 20 CpG markers or fewer were selected.

6.4.5 Validating the multi-tissue age prediction model

The constructed models were subsequently validated on ten additional individual data sets (Table 6.2), which were assayed on the Illumina HM27K or HM450K BeadChip platforms and downloaded from GEO, TCGA, and ArrayExpress, as described in Section 2.3. The testing data set consisted of 661 samples derived from six body fluids, namely blood, saliva, menstrual blood, vaginal secretions, uterine endometrium (included as a similar cell type to vaginal

secretions), and semen, from individuals with ages ranging from 6 to 90 years old. The main reason for using these types of tissues in the testing data set is because these are the types of tissue commonly found at crime scenes, and the model was created to be applied on these types of samples.

Table 6.2 Description of the 10 data sets used in the testing data set.

No.	DNA origin	Accession no.	Platform	n (Prop.Female)	Genomic repository
1	Saliva	GSE28746	HM27K	69 (0)	GEO
2	Semen, blood, menstrual blood, vaginal secretions, and saliva	GSE59509	HM450K	42	GEO
3	Menstrual and vaginal sect	GSE77283	HM450K	9 (1)	GEO
4	Uterine Endometrium	TCGA, UCEC	HM450K	34 (1)	TCGA
5	Buccal swab	E-MTAB-6730	HM450K	179	ArrayExpress
6	Blood	GSE76169	HM450K	63	GEO
7	Blood	GSE104812	HM450K	48	GEO
8	Buccal swab	GSE94876	HM27K	120	GEO
9	Semen	GSE115920	HM450	6 (0)	GEO
10	Blood	GSE41169	HM450	90 (0.28)	GEO

The data sets from the different Illumina platforms (HM27K and HM450K) in the testing data set were merged together by focusing on only the overlapping CpG probes (21,368). Similar to the training data set, each data set in the testing data set was individually tested and samples that showed abnormal DNAm level were removed from the downstream statistical analyses. The testing set was normalised using the normalisation algorithm developed by Horvath (2013) [106]. The samples in the testing data set were used to test the multi-tissue age prediction model created by elastic net regression, by using it to predict their

chronological ages, and then the MAD value calculated as described in Section 2.5.4.

6.5 Results

6.5.1 Illumina HumanMethylation data processing

The number of samples in the data sets downloaded from the genomic repositories for the training data set was initially 3,020 samples. However, when the DNAm profiles in each data set were evaluated individually, a number of samples did not pass the quality control measures and showed abnormal distribution of the methylation Beta values. Therefore, they were removed from downstream statistical analyses. Table 6.3 shows the number of outliers that were removed from each data set. After removing the outliers and combining all the samples in each of the data sets together, the number of samples in the training data set was 2,881 samples consisting of samples from 1,414 males, and 1,283 females (the remaining 184 were from individuals of unknown gender), with ages ranging from 0 to 101 years old (Figure 6.1).

Table 6.3 The number of abnormal samples in each data set in the training data set, and the number of samples that passed the quality control measures.

No.	DNA origin	Accession no.	No. of abnormal samples	No. of normal samples
1	Blood	GSE41037	57	658
2	Blood	GSE40279	12	644
3	Blood PBMC	GSE36064	0	78
4	Blood PBMC	GSE32148	2	46
5	Cord blood	GSE27317	50	166
6	Brain	GSE38873	0	168
7	Breast	GSE32393	0	23
8	Buccal swab	GSE25892	0	109
9	Colon	GSE32146	0	24
10	Dermal fibroblast	GSE22595	0	14
11	Bone marrow	GSE17448	0	16
12	Placenta	GSE36642	0	28
13	Prostate	GSE26126	0	70
14	Saliva	GSE34035	0	181
15	Uterine Cervix	GSE30758	0	152
16	Muscle	GSE38291	0	22
17	Blood CD4+CD14	GSE20242	0	50
18	Sperm	GSE26974	0	19
19	Kidney	TCGA, KIRP	0	45
20	Colon	TCGA, COAD	0	37
21	Lung	TCGA, LUSC	0	27
22	Lung	TCGA, LUAD	2	22
23	Lung	TCGA, LUSC	0	42
24	Prostate	TCGA, PRAD	0	50
25	Stomach	TCGA, STAD	16	25
26	Head and Neck	TCGA, HNSC	0	50
27	Breast	TCGA, BRCA	0	27
28	Breast	TCGA, BRCA	0	88

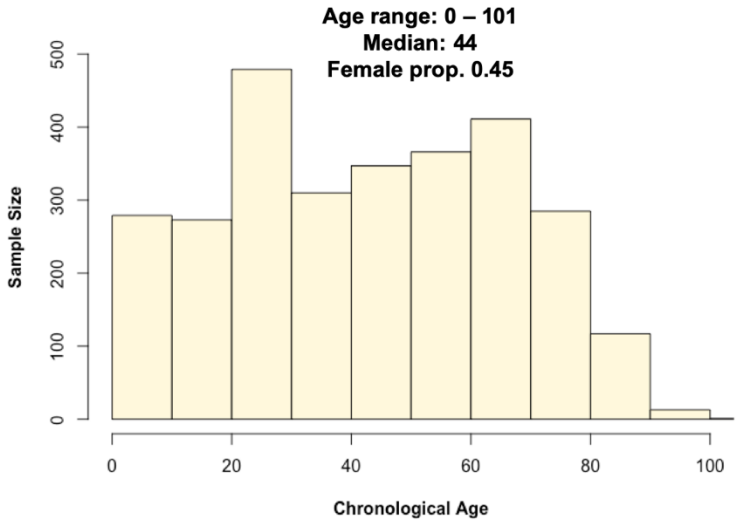

Figure 6.1 Age range in the training data set consisting of 2881 samples from 22 different tissues and cell types.

6.5.2 Singular value decomposition

The two principle reasons for conducting SVD analysis was to further assess the samples in the training data set by identifying samples not clustering with their labelled tissue, and to examine how distinctive DNAm patterns are between the different types of tissue. As Figure 6.3 shows, no outlier samples remained after the data processing described above, and all samples clustered with other samples labelled as being from the same tissue. One clear pattern seen in the graph is that DNAm pattern is tissue specific, and the samples in each tissue tend to cluster together and are not mixed with samples from other tissues. Another interesting aspect of the data is that the semen samples clustered away from all other body tissues, in the top right corner of the graph. This indicates that semen samples have very distinctive DNAm patterns compared to other body tissues.

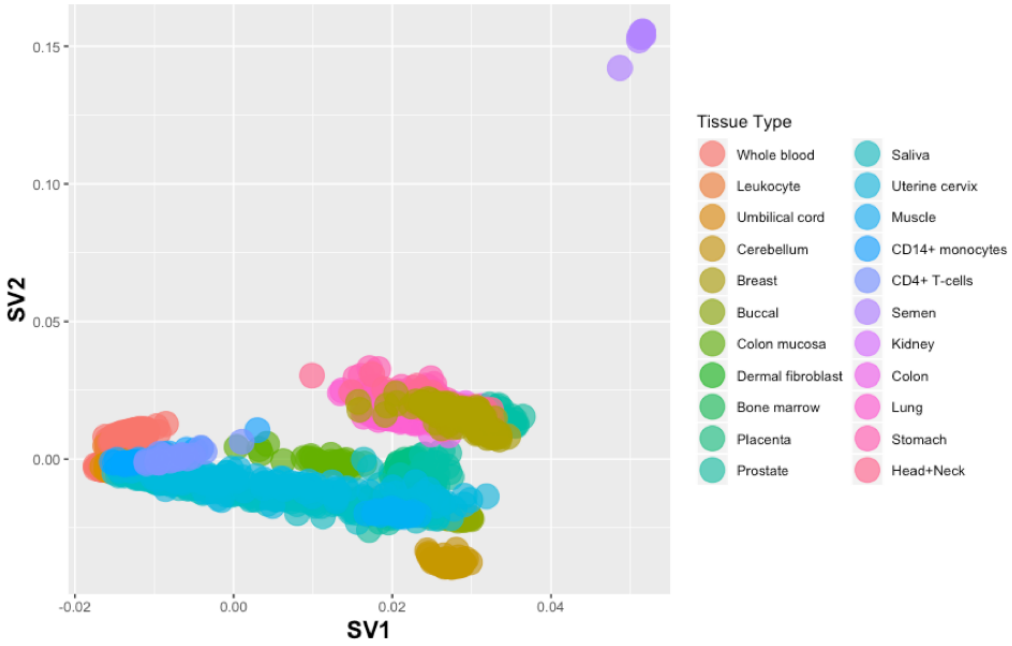

Figure 6.2 Singular Value Decomposition plot for 2,881 samples in the training data set, based on 21,368 CpG probes. The colours represent different tissue types.

6.5.3 Identifying universal CpG sites and building the multi-tissue age prediction model

Elastic net regression was performed on the training data set, containing 21,368 CpG sites, in 22 different tissue types. Elastic net regression constructs prediction models starting with single marker and then begins to add more markers to the model until it reaches the lowest mean squared error between the predicted age and chronological age, that is the best prediction accuracy. Initially the model was constructed without restricting the number of CpG sites that could be selected. The model that was automatically selected using elastic net regression that reached the lowest mean squared error had 267 AR CpG sites (Table A2) with an MAD value across all tissue types of 3.9 years (r = 0.97, *P*-value<$2.2x10^{-16}$) (Figure 6.3). Table 6.4 shows the prediction accuracy calculated separately for each tissue.

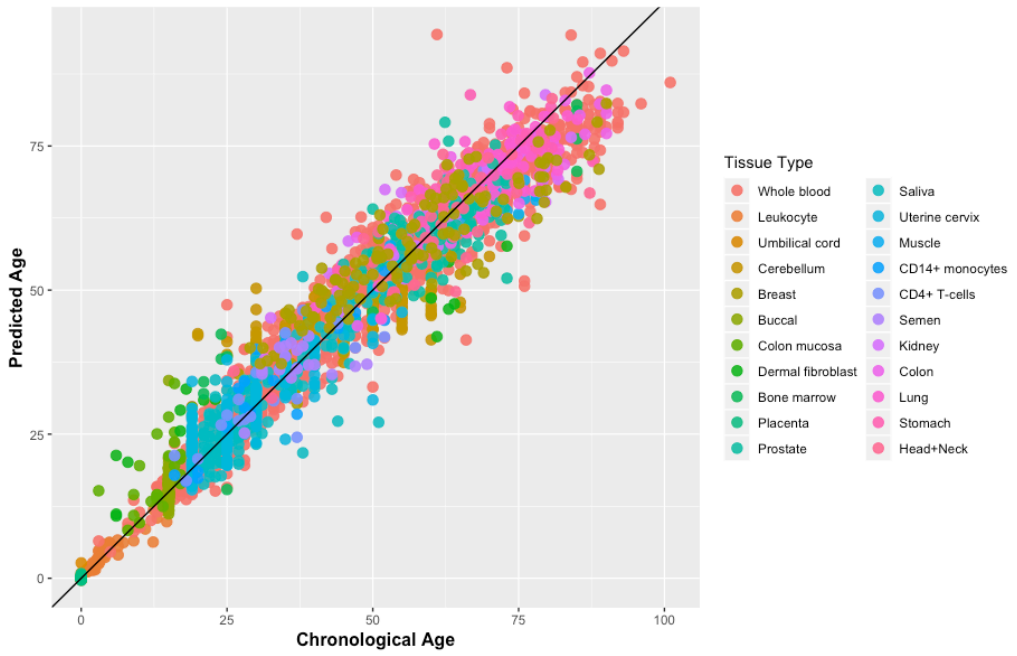

Figure 6.3 Predicted age versus chronological age in the training data set. Across all the samples (2,881 samples) in the training data set, the correlation between the predicted and chronological age is 0.97 (P-value < 2.2×10^{-16}) and the MAD value is 3.9 years.

Table 6.4 Age prediction accuracy (MAD) of the model selected by elastic net regression, containing 267 AR CpG sites, for each of the different tissues in the training data set.

No.	**DNA origin**	**MAD value (years)**
1	Whole blood	4.1
2	Leukocyte	0.87
3	Umbilical cord	0.14
4	Cerebellum	6.0
5	Breast	5.81
6	Buccal	1.87
7	Colon mucosa	5.44
8	Dermal fibroblast	11.8
9	Bone marrow	9.1
10	Placenta	0.21
11	Prostate	4.56
12	Saliva	3.32
13	Uterine cervix	3.42
14	Muscle	5.43
15	CD14+ monocytes	3.46
16	CD4+ T-cells	4.22
17	Semen	4.44
18	Kidney	5.23
19	Colon	5.32
20	Lung	4
21	Stomach	5.52
22	Head and neck	4.96

The next step in the analysis was to study the relationship between the number of CpG sites in the age prediction models and the corresponding MAD values. This was done in order to demonstrate how using a smaller subset of AR CpG sites in the model would affect age prediction accuracy. To achieve this, the MAD value for each model starting with a model based on 1 CpG site and adding sites up to the model based on 267 CpG sites was calculated and then plotted against each the number of sites. As Figure 6.4 shows, the age prediction accuracy across all tissues increases as the number of CpG sites in the model increases. However, as the number of markers in the model increases, the prediction accuracy reaches a plateau, and the MAD values remain steady and no further significant decrease is observed.

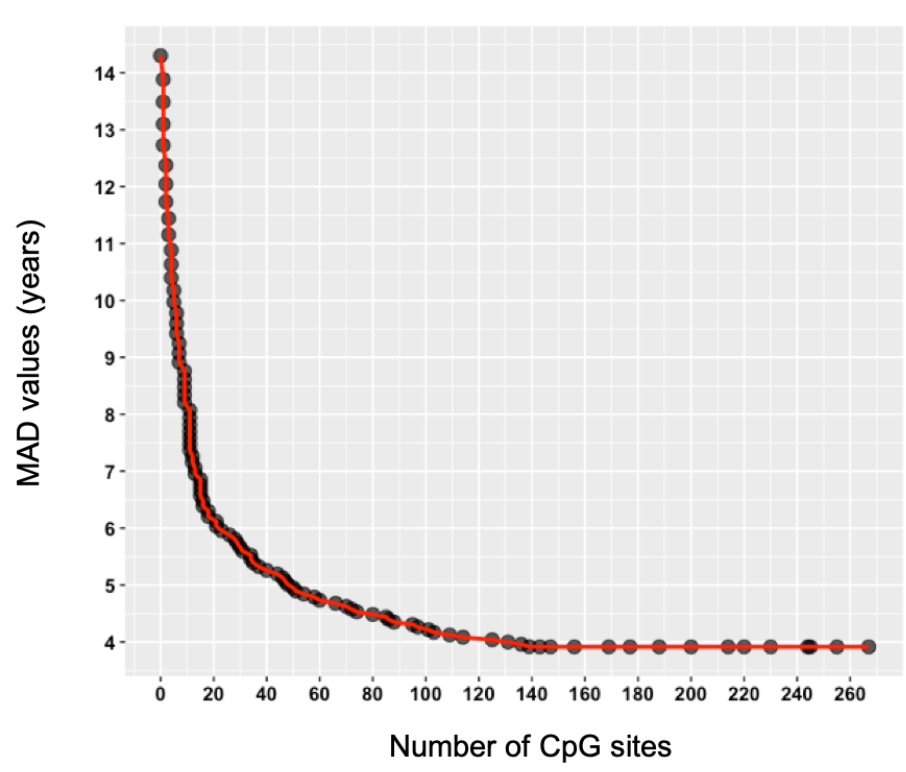

Figure 6.4 Age prediction accuracy of the models constructed using elastic net regression containing 1 to 267 CpG sites. The MAD value for each model was calculated based on the samples in the training data set for each of the 267 models.

6.5.4 The mini multi-tissue age prediction model

As shown in the previous section, the prediction accuracy of a model decreases as the number of the CpG sites included in it decreases. Given that in the literature, the number of CpG sites in any model constructed for forensic purposes did not exceed 16 CpG sites, the model with the same number of CpG sites was selected in this study in order to build a 'mini multi-tissue model', containing a small enough number of sites that would allow the model to be developed into a PCR-based assay in future. The identity of the 16 CpG sites can be seen in Table 6.5, and the change in DNAm level with age across all tissues

is illustrated in Figure 6.5. The overall prediction accuracy (MAD) of these 16 CpG sites based on the training data set was equal to 6.39 years (r 0.94) (Figure 6.6), and the MAD value for each individual tissue can be seen in Table 6.6. Furthermore, the performance of this model was also evaluated using testing data set, which can be seen in the following section.

Table 6.5 Identity of the 16 CpG markers selected by elastic net regression.

No.	Illumina's ID	Gene	Genomic location
1	cg01459453	*SELP*	chr1:169599212
2	cg01511567	*SSRP1*	chr11:57103631
3	cg06268694	*CELSR1*	chr22:46932642
4	cg06493994	*SCGN*	chr6:25652602
5	cg07388493	*NDUFS5*	chr1:39491459
6	cg07588779	*GPR137*	chr11:64051753
7	cg08996521	*CISH*	chr3:50649994
8	cg10893437	*ZNF828*	chr13:115079492
9	cg17324128	*RASSF4*	chr10:45455500
10	cg17861230	*PDE4C*	chr19:18343901
11	cg19722847	*IPO8*	chr12:30849114
12	cg21801378	*BRUNOL6*	chr15:72612125
13	cg22736354	*NHLRC1*	chr6:18122719
14	cg25809905	*ITGA2B*	chr17:42467728
15	cg26394940	*C22orf26*	chr22:46449461
16	cg26614073	*SCAP*	chr3:47517819

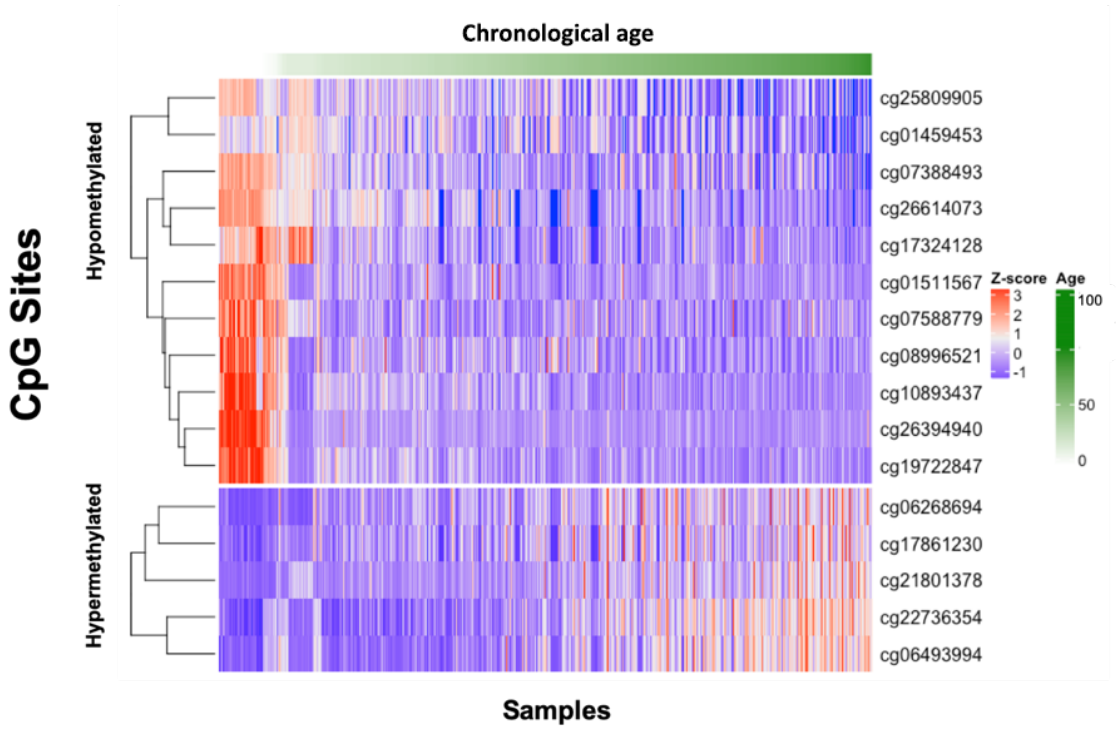

Figure 6.5 Heat map illustrating methylation level across all tissues at the 16 CpG sites selected by elastic net regression in the training data set, ordered by chronological age (indicated in green across the top of the figure). The methylation level in each sample is indicated by the Z-score colour code, where red indicates a site is hypermethylated and blue is hypomethylated. The branching patterns on the left indicate hierarchical clustering of the CpG sites.

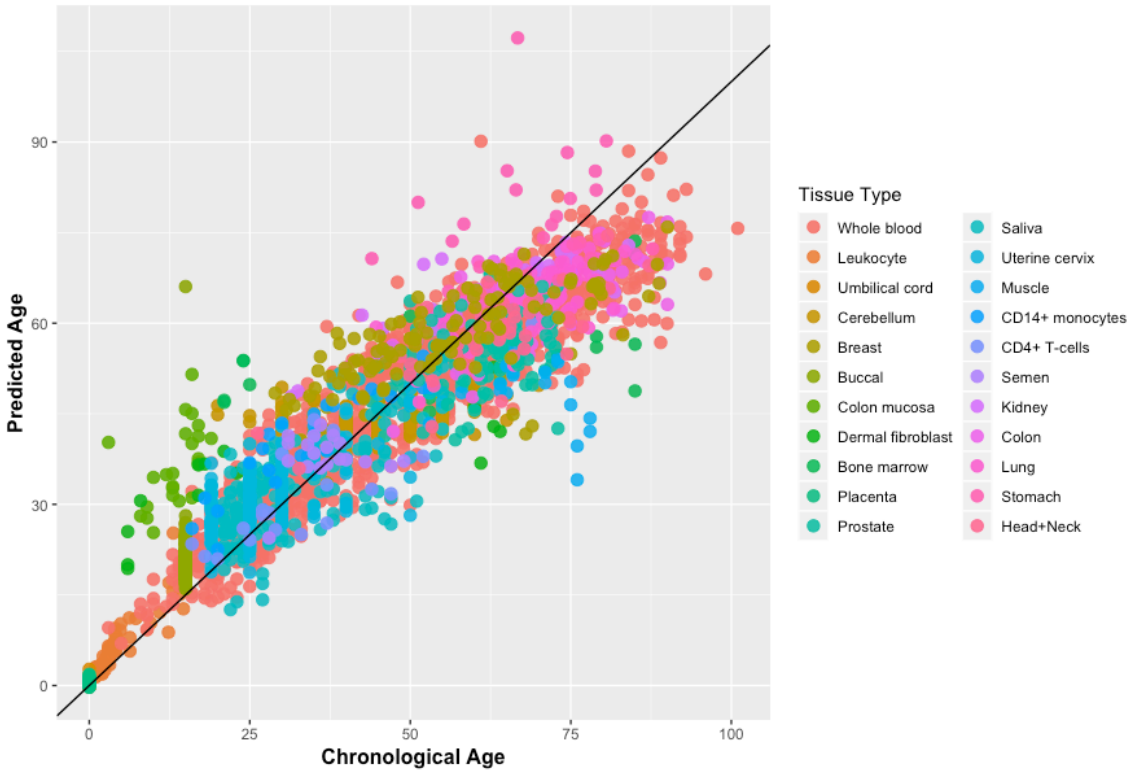

Figure 6.6 Predicted age versus chronological age in the training data based on a model containing 16 universal AR CpG sites. Across all training data samples, the correlation between the predicted and chronological age is 0.94 (P-value < 2.2×10^{-16}) and the MAD value is 6.39 years.

Table 6.6 Age prediction accuracy (MAD) of the model selected by elastic net regression, containing 16 AR CpG sites, for each of the different tissues in the training data set.

No.	DNA origin	MAD value (years)
1	Whole blood	6.2
2	Leukocyte	2.4
3	Umbilical cord	0.43
4	Cerebellum	7.3
5	Breast	8.2
6	Buccal	6.4
7	Colon mucosa	23
8	Dermal fibroblast	17.4
9	Bone marrow	19.2
10	Placenta	0.67
11	Prostate	7.8
12	Saliva	5.5
13	Uterine cervix	6.2
14	Muscle	19
15	CD14+ monocytes	7.2
16	CD4+ T-cells	6.1
17	Semen	5
18	Kidney	6.4
19	Colon	11
20	Lung	6.5
21	Stomach	12
22	Head and neck	6.4

6.5.5 Validating the multi-tissue age prediction model

The number of samples in the data sets downloaded from the genomic repositories for the testing data set was initially 661 samples. Before conducting validation testing, a single saliva sample (sample ID GSM1438496 from

accession number GSE59509) was removed from the testing data set, as it showed an abnormal Beta value distribution. The remaining samples that showed a normal DNAm pattern were combined into one testing data set containing 660 samples derived from six body fluids, namely blood, saliva, menstrual blood, vaginal secretions, uterine endometrium, and semen, from individuals with ages ranging from 6 to 90 years old (Figure 6.7). To further confirm how the number of the CpG sites in the model affected its prediction accuracy, the MAD value for each of the 267 models created by elastic net regression was calculated based on the samples in the testing data set. As expected, the MAD value for the testing data increased as the number of CpG sites in the model decreased (Figure 6.8).

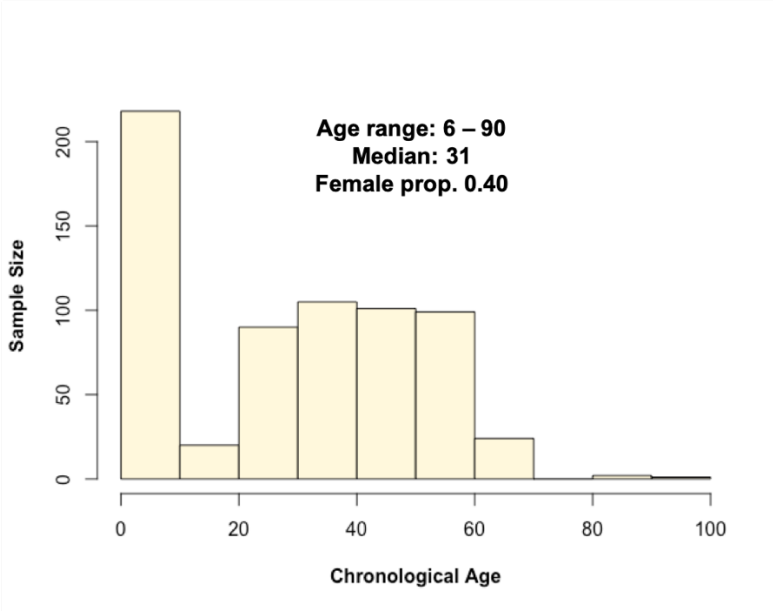

Figure 6.7 Age range in the testing data set consisting of 660 samples from six different tissues.

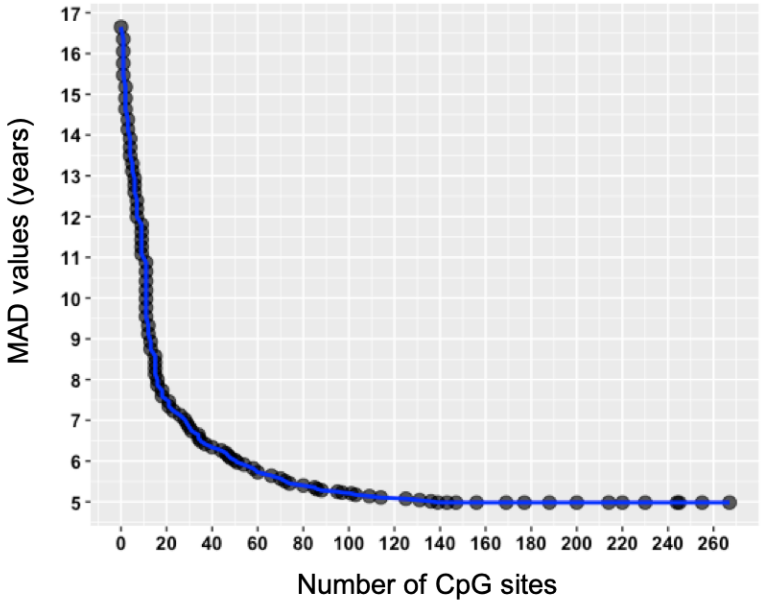

Figure 6.8 Age prediction accuracy of the models constructed using elastic net regression containing 1 to 267 CpG sites. The MAD value for each model was calculated based on the samples in the testing data set for each of the 267 models.

Given that in the literature, the number of CpG sites in any model constructed for forensic purposes did not exceed 16 CpG sites, the model with the same number of CpG sites was selected in this study in order to build a 'mini multi-tissue model', containing a small enough number of sites that would allow the model to be developed into a PCR-based assay in future. This model contained 16 universal AR CpG sites, and was validated separately on the testing data set. The prediction accuracy (MAD) of this model was 7.9 years (r = is 0.91, *P*-value < $2.2x10^{-16}$) across all of the six tested tissues (Figure 6.9A), however, the model showed very different performances on the different tissue types. For example, saliva samples exhibited the lowest MAD value, which was equal to 4.5 years, whereas semen and menstrual blood samples had the highest MAD values, at 11.7 and 12.8 years, respectively (Figure 6.9B-H). The low prediction accuracy for the menstrual blood and uterine endometrium samples may be due

to the menstrual cycle and concomitant increases in cell proliferation. This is in line with the Horvath (2013) study, who classified uterine endometrium as a poorly-age-predicted tissue, along with breast tissue, dermal fibroblasts, and heart tissue [106].

The high MAD value for the semen samples was also expected and is likely to be due to their distinctive DNAm pattern compared to the other body tissues, as shown in the SVD analysis in Section 6.5.2. This distinctive DNAm pattern in semen can be explained by the fact that germline cells go through epigenetic reprogramming during early development, in order to prevent the inheritance of aberrant DNAm patterns that might adversely influence gene expression in the offspring [18,19]. However, from Figure 6.9, it can be seen that there is less variation in the predicted DNAm ages in some tissues, especially whole blood, saliva, and buccal swabs, which indicates a consistency in the performance of the model in these tissues, across a wide range of chronological ages.

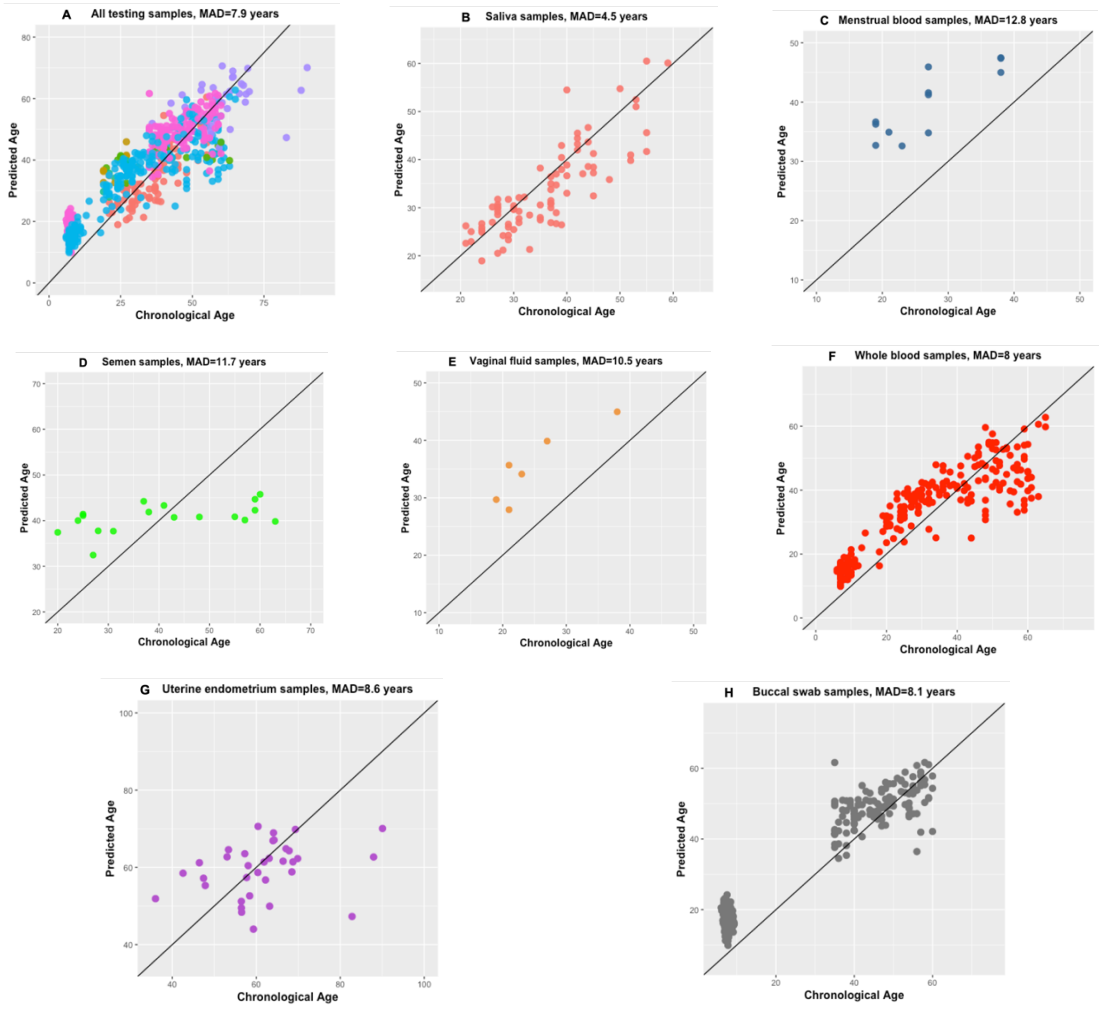

Figure 6.9 Predicted age versus chronological age in the training data based on a model containing 16 CpG sites. Across all testing data, the correlation (r) between the predicted and chronological age is 0.94 (*P*-value < $2.2x10^{-16}$) and the MAD value is 6.39 years. (**A**) Across all testing data samples, the correlation between the predicted and chronological age is 0.91 (*P*-value < $2.2x10^{-16}$) and the MAD value is 7.9 years. (**B**) Saliva samples (r = 0.83, *P*-value < $2.2x10^{-16}$, MAD = 4.5 years). (**C**) Menstrual blood samples (r = 0.84, *P*-value = $6x10^{-4}$, MAD = 12.8 years). (**D**) Semen samples (r = 0.51, *P*-value =0.03, MAD = 11.7 years). (**E**) Vaginal fluid samples (r = 0.89, *P*-value =0.02, MAD = 10.5 years). (**F**) Whole blood samples (r = 0.86, *P*-value < $2.2x10^{-16}$, MAD = 8 years). (**G**) Uterine endometrium samples (r = 0.3, *P*-value =0.1, MAD = 8.6 years). (**H**) Buccal swab samples (r = 0.96, *P*-value < $2.2x10^{-16}$, MAD = 8.1 years).

6.6 Discussion

A number of studies have demonstrated a statistically significant relationship between DNAm levels at specific CpG sites and chronological age in human tissues [106,131,145,156]. These CpG markers have the potential to be used in forensic investigations to estimate the age of an individual from various body fluids and tissues such as blood, saliva, semen and teeth, with high estimation accuracy [89,92,144,149,156]. However, in many forensic cases, the tissue source of the DNA evidence is unknown and requires an additional pre-processing step in order to identify the source before conducting a tissue-specific age prediction assay. As well as being time-consuming, this pre-processing step may consume a quantity of the DNA evidence, which is usually present in limited amounts. Thus, finding universal age-related CpG markers that can be used to estimate the age of an individual for forensic purposes across all types of forensically relevant tissues would be of major benefit.

In the literature, there are two multi-tissue age-prediction models, however these two models have several limitations, including the large number of CpG markers (353 CpG sites), and poor prediction accuracy (11.4 years MAD) across tissues [94,106]. Thus, the main aim of this research was to construct a model containing a limited number of universal AR CpG markers, while maintaining a good prediction accuracy across a wide range of tissues, which could be implemented for forensic purposes. To achieve this, elastic net regression has been implemented to perform both the identification of the universal AR CpG sites and construction of the age prediction model using a data set containing samples from 22 different tissue and cell types [106,136]. The reason for using a wide range of tissue and cell types with a wide range of chronological ages (from 0 to 101 years old), was to ensure the identified markers had the capability of predicting age across a wide range of ages, regardless of the tissue and/or cell type source.

The SVD analysis in Figure 6.2 showed that samples clustered together based on their tissue and cell type, which indicates that the DNAm patterns exhibit tissue-specific properties. This explains why DNAm markers, including AR CpG sites, perform very well in tissue-specific models, compared to models constructed across multiple tissues. For this reason, finding AR CpG sites that show similar correlations with chronological age across tissues is very challenging. Furthermore, the distinctive separation of semen samples from other body tissues confirms the phenomenon discussed in Section 1.2.5, in which the epigenome in the germ line is erased and reprogrammed during spermatogenesis, such that semen samples exhibit completely different DNAm patterns than samples from other tissues [18,19].

Elastic net regression was implemented in this study because it was used to build the most prominent multi-tissue age prediction model in the literature [106]. Although this algorithm selected a prediction model with an MAD value of 3.9 years across all tissue types in the training data set, the number of markers it contained (267 AR CpG sites) was very high. It would not be feasible to implement a PCR-based assay containing this number of sites for use on forensic samples. Therefore, the challenging task in this study was to select the model with the highest possible number of markers that could still be implemented in forensic laboratories. To solve this, the number of markers was selected based on the maximum number markers that have been used in forensic-based age prediction models reported in the literature, which was 16 CpG sites in a blood-specific model created by Vidaki et al. (2017). All of the 16 identified universal AR CpG sites are associated with genes (see Table 6.5) and 13 of them have been previously identified in the literature as being associated with age in different tissues. This included 11 CpG sites overlapping with the 353 CpG sites in Horvath's (2013) pan-tissue model (cg01459453, cg01511567, cg06493994, cg07388493, cg17324128, cg19722847, cg21801378, cg22736354, cg25809905, cg26394940, and cg26614073), one (cg17861230) overlapping with

Koch and Wagner's (2011) multi-tissue model [94,106], and one (cg08996521) significantly associated with neonatal gestational age [217]. The remaining three AR CpG sites (cg06268694, cg07588779, and cg10893437) have never previously been reported as being associated with age in any tissues.

The prediction accuracy (MAD) of the selected 16 universal AR CpG sites across all tissues was 7.9 years, which outperformed Koch and Wagner's (2011) multi-tissue model by 3.5 years (their MAD across tissues was 11.4 years). However, when compared to the majority of tissue-specific age prediction models, the multi-tissue model is less accurate, which was as expected. For example, using the model reported here to predict age from blood and semen samples resulted in MAD values of 8 and 11.7 years, respectively, whereas two blood- and semen-specific age prediction models reported in the literature had MAD values of 3.2 and 5.4 years, respectively [92,215]. However, due to the number of tissues and samples that were used in this study to construct the model, the MAD value achieved across tissues is the best prediction accuracy that any multi-tissue model has reached using this number of markers. The only way to improve the prediction accuracy of multi-tissue models for forensic applications would be to find an assay system that can profile a large number of sites from small/degraded DNA samples.

6.7 Summary and conclusions

The purpose of the study presented here was to use published data to screen the epigenome, to identify a small sub-set of universal AR CpG sites for age estimation across multiple forensically relevant tissues with a reasonable accuracy, which could potentially be incorporated into a multiplex PCR-based assay. This study has identified 16 universal age correlated CpG sites across 2,881 samples from 22 different tissues retrieved from individuals aged from 0 to 101 years old (Figure 6.10). These 16 AR markers were selected using elastic

net regression to build a multi-tissue age prediction model. This model displayed good prediction accuracy on a testing data set consisting of 660 samples from six body fluids, with a prediction accuracy (MAD) across tissues of 7.9 years. This suggests that the selected universal markers could be used for age estimation on the types of biological samples that are most frequently found at crime scenes, such as blood, semen, saliva, menstrual blood, and vaginal secretions. Although the results in Chapter 5 suggested that upgrading to the newer Illumina MethylationEPIC platform did not produce an age prediction model with better accuracy in blood, it might be possible to improve the multi-tissue model using MethylationEPIC data for a wide spectrum of tissues and a broader range of ages, and then develop a PCR-based assay that could easily be implemented in forensic laboratories.

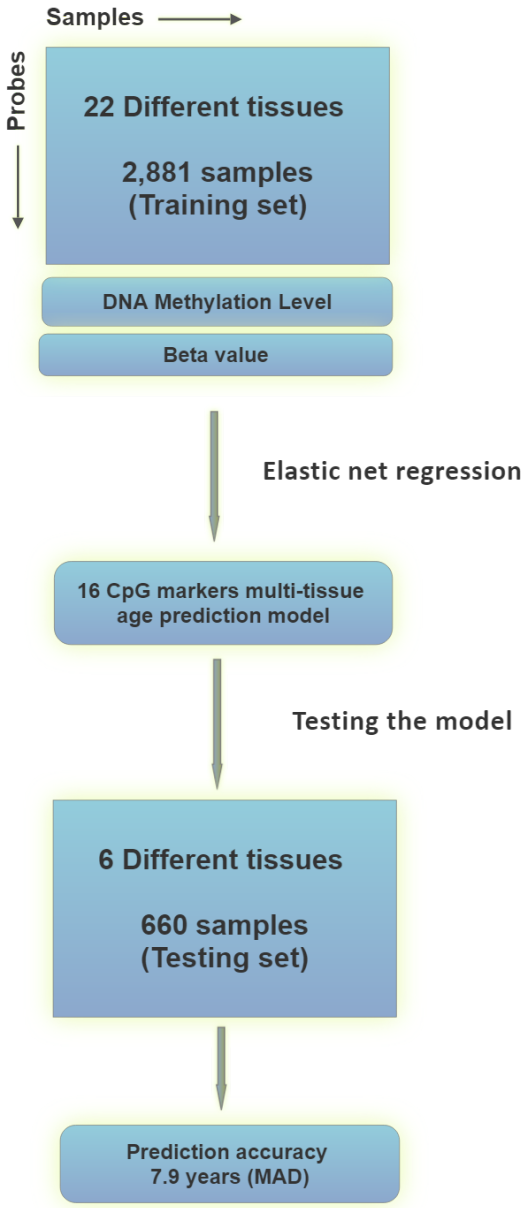

Figure 6.10 Schematic diagram summarising the main findings of this chapter.

Chapter 7: General discussion, conclusions, and recommendations for future work

DNA methylation (DNAm) CpG markers present a unique opportunity to answer a wide range of questions in forensic science that cannot be answered by conventional STR markers. This thesis focused on four research themes related to DNAm markers. Firstly, identifying the optimum statistical method for discovering age-related (AR) DNAm markers, and then using this method to build a saliva-specific age prediction model. Secondly, validating this model using next-generation sequencing using the Illumina MiSeq® platform. Thirdly, identifying blood-specific AR DNAm markers using the newly introduced Illumina MethylationEPIC® BeadChip, and finally, building a multi-tissue age prediction model with a small number of universal CpG sites that are capable of predicting the age of individuals regardless of the type of tissue being used.

In the last decade, forensic geneticists have shown an increased interest in the use of DNAm markers for age estimation in forensic casework. This has led to the introduction of various statistical methods in order to identify AR CpG sites for use in constructing age prediction models. Chapter 3 explored the use of three statistical methods, namely Spearman's rank and Pearson correlation tests, along with simple linear regression, and selected a standard set of procedures that were optimum for identifying AR CpG sites from high dimensional data generated using the HumanMethylation BeadChip platforms. In addition, the performance of these three methods were examined based on the two different DNAm measurements, Beta and M values.

The outcomes presented in Chapter 3 support the use of Spearman's rank correlation test over either Pearson's correlation or simple linear regression in identifying significant AR CpG sites. Based on the algorithm of the Spearman's

rank correlation, which measures monotonic relationships between variables, this finding suggests that the DNAm level at AR CpG sites increases monotonically with age. This is in line with a study conducted by Horvath (2013), who described the rate of change in DNAm level at AR CpG markers across tissues as taking the form of a logarithmic relationship from childhood until adulthood and then changing to a linear relationship later in life [106]. Furthermore, another study conducted by Xu et al. (2015) also highlighted that linear tests such as Pearson's correlation and linear regression are too simple to explain the complicated relationship between DNAm levels at AR markers and chronological age [8].

Another significant factor that was shown in Chapter 3 to affect the outcomes of the statistical methods for identifying AR CpG sites was the type of DNAm measurement (Beta or M values) being used. The reason behind this is that the Logit transformation of Beta values into M values alters the relationship between DNAm levels at AR markers and age from a monotonic to a linear relationship. Thus, using M values in Pearson's correlation and linear regression tests resulted in more AR markers than using Beta values. A similar finding was reported by Du et al. (2012) who recommended using M values with linear tests for conducting differential methylation analyses, as they perform better in terms of the detection rate and in terms of detecting true positives for both highly methylated and unmethylated CpG sites [105]. On the whole, the outcomes of Chapter 3 recommend using either Beta or M values with nonlinear methods such Spearman's rank correlation and quadratic regression. However, when linear methods such as Pearson's correlation and linear regression methods are implemented, M values should be used (Figure 3.21).

An interesting further extension of these findings in Chapter 3 was the implementation of the selected statistical method on a data set of 54 saliva samples retrieved from the National Centre for Biotechnology Information (NCBI) Gene Expression Omnibus (GEO) database, which were assayed on the

HM450K, in order to construct a saliva-specific age prediction model that outperformed the models in the literature. Nine CpG markers were successfully identified using Spearman's rank correlation test and selected for construction into a prediction model by stepwise regression. These nine markers were integrated into a multivariate linear regression model that explained 97.5% of the total variation in the data, with a prediction accuracy of 1.3 years (mean absolute deviation (MAD) between predicted and chronological age), which outperforms the best previously reported models in the literature. This model was validated *in silico* on an independent data set downloaded from NCBI repository, consisting of 56 saliva samples that were collected from the Khomani San population living in the South African Kalahari Desert. The prediction accuracy of the model on this testing data set was 5.1 years (MAD). The identification of an optimal method for identification of AR DNAm markers is beneficial for researchers in many disciplines aiming to identify AR DNAm markers across tissues.

In Chapter 4, the nine sites identified in Chapter 3 were then validated by targeted bisulfite sequencing of DNA from an additional 192 saliva samples, using the Illumina MiSeq® platform. Sequencing the nine candidate CpG sites resulted in the identification of neighbouring AR CpG sites with stronger association with age, in the genes *ELOVL2* and *ZEB2*, as well as genomic locations 10p12.2 and 1q32.2. The best subset of these adjacent AR CpG sites was selected by stepwise regression and then modelled using a quadratic modelling system. The quadratic model was composed of 10 different AR CpG markers, and was trained on a data set of 100 saliva samples (age range 13-96 years). The model explained 92% (R^2 = 0.92) of the total variation in DNAm levels, with mean absolute deviation (MAD) between predicted and chronological age of 3.4 years, and a Pearson's correlation coefficient (r) of 0.96. Subsequently, the performance of this model was validated on an independent data set of 65 additional saliva samples (age range 13-84 years), which produced a prediction accuracy based on bootstrap analysis of 5.26 years (95% confidence intervals 5.24-5.27 years), and a

Pearson's correlation coefficient (r) of 0.88. The performance of this model was further assessed by comparison with the best saliva-specific age prediction model reported in the literature, which was created by Hong et al. (2017) and composed of seven CpG markers. Based on the same training and testing sets that were used to build and validate the saliva-specific HM450K model, their model explained 68% (R^2 = 0.68) of the total variation in DNAm levels, with MAD values of 7.7 years, and 7.5 years, respectively. Since next generation sequencing platforms such as the MiSeq® are likely to dominate forensic laboratories in the near future, the quadratic model reported here can be integrated into the routine forensic laboratory workflow in order to estimate chronological age from saliva samples.

Two microarray platforms, the Illumina HumanMethylation27 (HM27K), and HumanMethylation450 (HM450K) BeadChips, played an important role in identifying a large number of AR CpG sites that have been used to build many age prediction models for forensic purposes. In addition, they have enriched the public databases of epigenome-wide DNAm profiles, which now contain samples from a large body of epigenetic studies based on different human tissues [193]. Thus, introducing a new array with over 860,000 probes, nearly double the number on the HM450K, will attract researchers working in the field of age estimation to study these probes for potential AR CpG sites that could improve the age prediction accuracy of models implemented in forensic science.

This new Illumina platform is the MethylationEPIC® (EPIC) BeadChip, which was examined in Chapter 5, in order to identify novel blood-specific AR CpG sites that could potentially be used for forensic purposes. The newly-added probes were examined using a large cohort of 754 blood DNAm profiles assayed on the EPIC BeadChip, from individuals aged 0-88 years old. Interestingly, 21 novel AR CpG sites were discovered that mapped to 18 genes, nine of which (*LHFPL4, SLC12A8, EGFEM1P, GPR158, TAL1, KIAA1755, LOC730668,*

DUSP16, and *FAM65C)* have never previously been reported in the literature to be associated with age. Discovering new genes harbouring AR CpG sites can aid forensic geneticists to further study that regions by targeted bisulfite sequencing, which may result in the identification of additional AR sites with high age prediction accuracy, which could be exploited for forensic science.

The uses of age prediction models are not restricted to forensic science, but also have clinical applications. For instance, predicted age (DNAm age) has been found to be related to frailty [218], cognitive/physical fitness in the elderly [219], Parkinson's disease, Alzheimer's disease-related neuropathology [220], and can predict overall mortality in humans [221]. However, the major difference between models created for forensic and clinical uses, is that the former requires a small number of markers, due to the nature of the forensic specimens, whereas the former can include unlimited numbers of markers as clinical samples tend to be more abundant. For this reason, the data were split into a 527-sample training set and a 227-sample testing set, and two separate models were created based on EPIC DNAm profiles. One model was constructed using multivariate regression (using the variable reduction, selection, and model building procedures described in Section 2.5), which contained six AR CpG sites, and the other model was constructed using elastic net regression, which contained 425 AR CpG sites.

The elastic net regression model contained 160 (38%) AR CpG sites that came from the newly-added probes on the EPIC BeadChip. The accuracy of this model based on the independent testing set was 2.6 years (MAD), which was highly accurate compared to other models reported in the literature. However, the outperformance of this model compared to other models in the literature may due to the number of CpG sites it contains, rather than the type of markers. For example, in a study conducted by Hannum et al. (2013), which had a similar experimental design but used Illumina HM450K data, their elastic net regression

model contained only 71 AR CpG markers and had a prediction accuracy of 4.89 years (MAD).

Building a prediction model for use in forensic investigations requires a small number of markers due to the minute quantities of DNA that are frequently recovered from forensic samples [15]. A second model was therefore constructed using multivariate linear regression. The six AR CpG sites selected by this stepwise regression, which contained only one CpG marker that was newly-added to the EPIC BeadChip, explained 81% of age-correlated variation in DNAm levels and had a MAD value of 4.6 years, with 95% confidence intervals of 4.56 to 4.57 years, based on the testing data set. A review of the literature shows that the range of MAD values achieved by forensic researchers for models based on blood samples was 3.2 to 7.9 years, using two to 17 CpG markers [144,214,215]. Therefore, the prediction accuracy of data generated using the EPIC BeadChip falls within the range of MAD values reported in previous studies.

Finally, one of the limitations of AR CpG sites is that they tend to be tissue specific and their accuracy in age estimation is highly linked to the type of tissue they are specific for. This means that using DNAm level for one set of tissue-specific AR markers to predict chronological age from other tissues has been shown to result in poor prediction accuracy [155,156]. Despite the fact that there are pre-processing steps that can be used to determine the tissue source before applying the right age prediction model, these steps are time-consuming, and can consume a large amount of the DNA evidence, which is usually present in limited amounts. In the literature, there are two different researchers who have built multi-tissue age prediction models that can be used across multiple tissues, however their models have several limitations, including the large number of markers (353 in Horvath's (2013) model), which could not be implemented in a forensic laboratory, and poor prediction accuracy (11.4 years in Koch and Wagner's model (2011)) [94,106]. The main focus of Chapter 6 was therefore on the identification of universal AR CpG sites that can be used to predict age from tissues that are frequently

recovered from crime scenes, such as blood, semen, saliva, menstrual blood, and vaginal secretions.

A multi-tissue age prediction model with 16 universal AR markers was constructed using elastic net regression, based on a training data set containing 2,881 samples from donors with ages ranging from 0-101, retrieved from 22 different tissues and cell types. Training the model with this large number of samples with a wide age range, and a wide range of tissues and cell types was done in order to ensure that the identified universal markers could predict age with as high as possible accuracy from any tissue source and cell type. This is very important as in many forensic cases DNA is recovered from unidentified tissue sources. The multi-tissue model was validated on a testing data set of 660 samples from six different forensically-relevant tissues (blood, saliva, semen, menstrual blood, vaginal secretions, and uterine endometrium), and the results showed a prediction accuracy (MAD) of 7.9 years. By ranking the tissues from high to low prediction accuracy, it was observed that age could be predicted from saliva with the highest accuracy (4.5 years), then blood (8 years), buccal swabs (8.1 years), uterine endometrium (8.6 years), vaginal secretions (10.5 years), semen (11.7 years), and finally menstrual blood (12.8 years).

The results in Chapter 6 suggest that the multi-tissue age prediction model could be implemented for forensic purposes in cases where the tissue source is unknown, however the prediction accuracy (7.9 years) is not as high as the tissue-specific age prediction models. Furthermore, one of the ways to enhance the multi-tissue model is by incorporating more markers, because as the number of markers increases the prediction accuracy of the model will also increase. In addition, although the results in Chapter 5 suggested that upgrading to the newer Illumina EPIC platform did not enhance age prediction accuracy in blood, it might be possible to improve the multi-tissue model using EPIC data for a wide spectrum of tissues and broader ages.

Overall, this thesis has taken a broad look at age estimation using DNAm AR CpG sites and their application in forensic science. The results presented suggest that, if the appropriate statistical method is implemented, epigenome-wide platforms such as the Illumina HumanMethylation BeadChips (HM27K, HM450K, and EPIC) can provide DNAm CpG sites that are reliable biomarkers for age estimation (Chapter 3). Furthermore, these AR CpG markers can be detected by next generation sequencing using the MiSeq® platform (Chapter 4), which is a technology that is more likely to dominate forensic laboratories in the near future. Thus, combining both DNAm analysis for age estimation and DNA sequence variation for human identification in a single streamlined process using an NGS platform will be the next target for many forensic researchers. However, from the results in Chapter 5, it can be seen that the upcoming age-related studies will reach a certain level of age prediction accuracy beyond which they cannot enhance anymore. The reason for this is that, although there will be novel AR markers coming from the newly-added probes on the new epigenome-wide platforms, the strength of the age association is likely to remain within the range of those found on the older platforms. Thus, the only way to further enhance the performance of age prediction models will be by looking for other factors such single nucleotide polymorphism (SNP) markers, that can be incorporated along with AR DNAm markers and enhance their prediction ability. Finding such factors may help increase the prediction accuracy of multi-tissue age prediction models, which have less accurate estimation compared to tissue-specific models. Finally, the outcomes of this work are not only applicable to forensic science, but also in clinical research where AR CpG sites are also associated with various diseases, and thus can be used to study the epigenetic basis of disease.

References

1. Kayser M, Schneider PM. DNA-based prediction of human externally visible characteristics in forensics: Motivations, scientific challenges, and ethical considerations. Forensic Science International: Genetics. Elsevier; 2009 Jun;3(3):154–61.

2. Kidd KK, Pakstis AJ, Speed WC, Grigorenko EL, Kajuna SLB, Karoma NJ, et al. Developing a SNP panel for forensic identification of individuals. Forensic Science International. Elsevier; 2006 Dec 1;164(1):20–32.

3. Beaumont KA, Shekar SN, Cook AL, Duffy DL, Sturm RA. Red hair is the null phenotype of MC1R. Human Mutation. Wiley Subscription Services, Inc., A Wiley Company; 2008 Aug 1;29(8):E88–E94.

4. Sulem P, Gudbjartsson DF, Stacey SN, Helgason A, Rafnar T, Magnusson KP, et al. Genetic determinants of hair, eye and skin pigmentation in Europeans. Nature Genetics. Nature Publishing Group; 2007 Dec 1;39(12):1443–5.

5. Shiota K. DNA methylation profiles of CpG islands for cellular differentiation and development in mammals. Cytogenet Genome Res. Karger Publishers; 2004 Jul 14;105(2-4):325–34.

6. Frumkin D, Wasserstrom A, Budowle B, Davidson A. DNA methylation-based forensic tissue identification. Forensic Science International: Genetics. Elsevier Ireland Ltd; 2011 Nov 1;5(5):517–24.

7. Yi SH, Jia YS, Mei K, Yang RZ, Huang DX. Age-related DNA methylation changes for forensic age-prediction. Int J Legal Med. 2014 Nov 16;129(2):237–44.

8. Xu C, Qu H, Wang G, Xie B, Shi Y, Yang Y, et al. A novel strategy for forensic age prediction by DNA methylation and support vector regression model. Scientific Reports. Nature Publishing Group; 2015 Dec 4;5(1):17788.

9. Kader F, Ghai M. DNA methylation and application in forensic sciences. Forensic Science International. Elsevier Ireland Ltd; 2015 Apr;249:255–65.

10. Bird A. Perceptions of epigenetics. Nature. 2007 May 24;447(7143):396–8.

11. Robinson BWS, Erle DJ, Jones DA, Shapiro S, Metzger WJ, Albelda SM, et al. Recent advances in molecular biological techniques and their relevance to pulmonary research. Thorax. BMJ Publishing Group Ltd and British Thoracic Society; 2000 Apr 1;55(4):329–39.

12. Russo VE, Martienssen RA, Riggs AD. Epigenetic mechanisms of gene regulation. Cold Spring Harbor Laboratory Press; 1996. pp. 1–4.

13. Lieb JD, Beck S, Bulyk ML, Farnham P, Hattori N, Henikoff S, et al. Applying whole-genome studies of epigenetic regulation to study human disease. Cytogenet Genome Res. Karger Publishers; 2006 May 24;114(1):1–15.

14. Schaukowitch K, Kim TK. REVIEWEMERGING EPIGENETIC MECHANISMS OF LONG NON-CODING RNAS. Neuroscience. IBRO; 2014 Apr 4;264(C):25–38.

15. Vidaki A, Daniel B, Court DS. Forensic DNA methylation profiling—Potential opportunities and challenges. Forensic Science International: Genetics. 2013 Sep;7(5):499–507.

16. Espada J, Esteller M. DNA methylation and the functional organization of the nuclear compartment. Seminars in Cell and Developmental Biology. Elsevier Ltd; 2010 Apr 1;21(2):238–46.

17. Rando OJ, Verstrepen KJ. Timescales of genetic and epigenetic inheritance. Cell. Elsevier; 2007 Feb 23;128(4):655–68.

18. Tang WWC, Dietmann S, Irie N, Leitch HG, Floros VI, Bradshaw CR, et al. A unique gene regulatory network resets the human germline epigenome for development. Cell. The Authors; 2015 Jun 4;161(6):1453–67.

19. Gkountela S, Zhang KX, Shafiq TA, Liao W-W, Hargan-Calvopiña J, Chen P-Y, et al. DNA Demethylation Dynamics in the Human Prenatal Germline. Cell. Elsevier Inc; 2015 Jun 4;161(6):1425–36.

20. Jablonka E, Lamb MJ. The inheritance of acquired epigenetic variations. Int J Epidemiol. Oxford University Press; 2015 Aug;44(4):1094–103.

21. Yagi S, Hirabayashi K, Sato S, Li W, Hattori N. DNA methylation profile of tissue-dependent and differentially methylated regions (T-DMRs) in mouse promoter regions demonstrating tissue-specific gene expression. Genome Res. 2008;18:1969–78.

22. Huh I, Zeng J, Park T, Yi SV. DNA methylation and transcriptional noise. Epigenetics Chromatin. 2013;6(1):9.

23. Shemer R, Birger Y, Dean WL, Reik W, Riggs AD, Razin A. Dynamic methylation adjustment and counting as part of imprinting mechanisms. PNAS. National Acad Sciences; 1996 Jun 25;93(13):6371–6.

24. Conerly M, Grady WM. Insights into the role of DNA methylation in disease through the use of mouse models. Disease Models & Mechanisms. The Company of Biologists Ltd; 2010 May 1;3(5-6):290–7.

25. Thompson JJ, Robertson KD. Misregulation of DNA Methylation Regulators in Cancer. In: DNA and Histone Methylation as Cancer Targets. Cham: Humana Press, Cham; 2017. pp. 97–124. (Cancer Drug Discovery and Development; vol. 99).

26. Li E, Bestor TH, Jaenisch R. Targeted mutation of the DNA methyltransferase gene results in embryonic lethality. Cell. Cell Press; 1992 Jun 12;69(6):915–26.

27. Bestor TH. The DNA methyltransferases of mammals. Human Molecular Genetics. Oxford University Press; 2000 Oct 1;9(16):2395–402.

28. Fernandez AF, Huidobro C, Fraga MF. De novo DNA methyltransferases: oncogenes, tumor suppressors, or both? Trends Genet. Elsevier; 2012 Oct;28(10):474–9.

29. Tang M, Xu W, Wang Q, Xiao W, Xu R. Potential of DNMT and its Epigenetic Regulation for Lung Cancer Therapy. CG. 2009 Aug 1;10(5):336–52.

30. Fang Q-L, Yin Y-R, Xie C-R, Zhang S, Zhao W-X, Pan C, et al. Mechanistic and biological significance of DNA methyltransferase 1 upregulated by growth factors in human hepatocellular carcinoma. International Journal of Oncology. Spandidos Publications; 2015 Feb 1;46(2):782–90.

31. Mizuno S-I, Chijiwa T, Okamura T, Akashi K, Fukumaki Y, Niho Y, et al. Expression of DNA methyltransferases DNMT1,3A, and 3B in normal hematopoiesis and in acute and chronic myelogenous leukemia. Blood. American Society of Hematology; 2001 Mar 1;97(5):1172–9.

32. Yang J, Wei X, Wu Q, Xu Z, Gu D, Jin Y, et al. Clinical significance of the expression of DNA methyltransferase proteins in gastric cancer. Molecular Medicine Reports. Spandidos Publications; 2011 Nov 1;4(6):1139–43.

33. Agoston AT, Argani P, Yegnasubramanian S, De Marzo AM, Ansari-Lari MA, Hicks JL, et al. Increased protein stability causes DNA methyltransferase 1 dysregulation in breast cancer. J Biol Chem. American Society for Biochemistry and Molecular Biology; 2005 May 6;280(18):18302–10.

34. Kohli RM, Zhang Y. TET enzymes, TDG and the dynamics of DNA demethylation. Nature. 2013 Oct 23;502(7472):472–9.

35. Auclair G, Weber M. Mechanisms of DNA methylation and demethylation in mammals. Biochimie. Elsevier; 2012 Nov 1;94(11):2202–11.

36. Branco MR, Ficz G, Reik W. Uncovering the role of 5-hydroxymethylcytosine in the epigenome. Nature Reviews Genetics. Nature Publishing Group; 2012 Jan 1;13(1):7–13.

37. Wu H, Zhang Y. Mechanisms and functions of Tet protein-mediated 5-methylcytosine oxidation. Genes Dev. Cold Spring Harbor Lab; 2011 Dec 1;25(23):2436–52.

38. He Y-F, Li B-Z, Li Z, Liu P, Wang Y, Tang Q, et al. Tet-Mediated Formation of 5-Carboxylcytosine and Its Excision by TDG in Mammalian DNA. Science. American Association for the Advancement of Science; 2011 Sep 2;333(6047):1303–7.

39. Ito S, Shen L, Dai Q, Wu SC, Collins LB, Swenberg JA, et al. Tet Proteins Can Convert 5-Methylcytosine to 5-Formylcytosine and 5-Carboxylcytosine. Science. American Association for the Advancement of Science; 2011 Sep 2;333(6047):1300–3.

40. Valinluck V, Sowers LC. Endogenous cytosine damage products alter the site selectivity of human DNA maintenance methyltransferase DNMT1. Cancer Res. 2007;67(3):946–50.

41. Cortellino S, Xu J, Sannai M, Moore R, Caretti E, Cigliano A, et al. Thymine DNA Glycosylase Is Essential for Active DNA Demethylation by Linked Deamination-Base Excision Repair. Cell. 2011 Jul;146(1):67–79.

42. Larsen F, Gundersen G, Lopez R, Prydz H. CpG islands as gene markers in the human genome. Genomics. 1992;13(4):1095–107.

43. Newell-Price J, Clark AJL, King P. DNA Methylation and Silencing of Gene Expression. Trends in Endocrinology & Metabolism. Elsevier; 2000 May;11(4):142–8.

44. Keshet I, Lieman-Hurwitz J, Cedar H. DNA methylation affects the formation of active chromatin. Cell. 1986;44(4):535–43.

45. Frommer M, McDonald LE, Millar DS. A genomic sequencing protocol that yields a positive display of 5-methylcytosine residues in individual DNA strands. 1992. pp. 1827–31.

46. Lokk K, Modhukur V, Rajashekar B, Märtens K, Mägi R, Kolde R, et al. DNA methylome profiling of human tissues identifies global and tissue-specific methylation patterns. Genome Biology. BioMed Central; 2014 Apr 1;15(4):3248.

47. Weber M, Hellmann I, Stadler MB, Ramos L, Pääbo S, Rebhan M, et al. Distribution, silencing potential and evolutionary impact of promoter DNA methylation in the human genome. Nature Genetics. Nature Publishing Group; 2007 Apr 1;39(4):457–66.

48. Cotton AM, Price EM, Jones MJ, Balaton BP, Kobor MS, Brown CJ. Landscape of DNA methylation on the X chromosome reflects CpG density, functional chromatin state and X-chromosome inactivation. Human Molecular Genetics. Oxford University Press; 2015 Mar 15;24(6):1528–39.

49. Hanna CW, Kelsey G. Genomic imprinting beyond DNA methylation: a role for maternal histones. Genome Biology. BioMed Central; 2017 Dec 1;18(1):177.

50. Ferguson-Smith AC. Genomic imprinting: the emergence of an epigenetic paradigm. Nature Publishing Group. Nature Publishing Group; 2011 Aug 1;12(8):565–75.

51. Song F, Smith JF, Kimura MT, Morrow AD, Matsuyama T, Nagase H, et al. Association of tissue-specific differentially methylated regions (TDMs) with differential gene expression. PNAS. National Acad Sciences; 2005 Mar 1;102(9):3336–41.

52. Byun H-M, Siegmund KD, Pan F, Weisenberger DJ, Kanel G, Laird PW, et al. Epigenetic profiling of somatic tissues from human autopsy specimens identifies tissue- and individual-specific DNA methylation patterns. Human Molecular Genetics. Oxford University Press; 2009 Dec 15;18(24):4808–17.

53. Holliday R. Epigenetics: a historical overview. Epigenetics. 2006;1(2):76–80.

54. Szyf M. DNA methylation, the early-life social environment and behavioral disorders. J Neurodevelop Disord. 2011 Mar 11;3(3):238–49.

55. Dhingra R, Nwanaji-Enwerem JC, Samet M, Ward-Caviness CK. DNA Methylation Age—Environmental Influences, Health Impacts, and Its Role in Environmental Epidemiology. Current Environmental Health Reports. Current Environmental Health Reports; 2018 Jul 23;5(3):317–27.

56. Hannon E, Knox O, Sugden K, Burrage J, Wong CCY, Belsky DW, et al. Characterizing genetic and environmental influences on variable DNA methylation using monozygotic and dizygotic twins. Greally JM,

editor. PLOS Genet. Public Library of Science; 2018 Aug 9;14(8):e1007544.

57. Fraga MF, Ballestar E, Paz MF, Ropero S, Setien F, Ballestar ML, et al. Epigenetic differences arise during the lifetime of monozygotic twins. PNAS. National Acad Sciences; 2005 Jul 26;102(30):10604–9.

58. Cardno AG, Rijsdijk FV, Sham PC. A twin study of genetic relationships between psychotic symptoms. The American Journal of Psychiatry. 2002;159(4):539–45.

59. Waterland RA, Jirtle RL. Transposable Elements: Targets for Early Nutritional Effects on Epigenetic Gene Regulation. Mol Cell Biol. 2003 Aug 1;23(15):5293–300.

60. Szyf M. The early life environment and the epigenome. Biochimica et Biophysica Acta (BBA) - General Subjects. 2009 Sep;1790(9):878–85.

61. Breitling LP, Yang R, Korn B, Burwinkel B, Brenner H. Tobacco-Smoking-Related Differential DNA Methylation: 27K Discovery and Replication. The American Journal of Human Genetics. Cell Press; 2011 Apr 8;88(4):450–7.

62. Shenker NS, Polidoro S, van Veldhoven K, Sacerdote C, Ricceri F, Birrell MA, et al. Epigenome-wide association study in the European Prospective Investigation into Cancer and Nutrition (EPIC-Turin) identifies novel genetic loci associated with smoking. Human Molecular Genetics. Oxford University Press; 2013 Mar 1;22(5):843–51.

63. Monick MM, Beach SRH, Plume J, Sears R, Gerrard M, Brody GH, et al. Coordinated changes in AHRR methylation in lymphoblasts and

pulmonary macrophages from smokers. American Journal of Medical Genetics Part B: Neuropsychiatric Genetics. Wiley-Blackwell; 2012 Mar 1;159B(2):141–51.

64. Lee KWK, Pausova Z. Cigarette smoking and DNA methylation. Front Genet. 2013;4:132.

65. Opitz CA, Litzenburger UM, Sahm F, Ott M, Tritschler I, Trump S, et al. An endogenous tumour-promoting ligand of the human aryl hydrocarbon receptor. Nature. Nature Publishing Group; 2011 Oct 1;478(7368):197–203.

66. Joubert BR, Håberg SE, Nilsen RM, Wang X, Vollset SE, Murphy SK, et al. 450K Epigenome-Wide Scan Identifies Differential DNA Methylation in Newborns Related to Maternal Smoking during Pregnancy. Environ Health Perspect. National Institute of Environmental Health Science; 2012 Oct 1;120(10):1425–31.

67. Lee HY, Lee SD, Shin K-J. Forensic DNA methylation profiling from evidence material for investigative leads. BMB reports. 2016 Jul 31;49(7):359–69.

68. Sesso HD, Cook NR, Buring JE, Manson J. Alcohol consumption and the risk of hypertension in women and men. Hypertension. Alcohol consumption and the risk of hypertension in women and men; 2008;51(4):1080–7.

69. Philibert R, Plume JM, Gibbons FX, Brody GH, Beach S. The Impact of Recent Alcohol Use on Genome Wide DNA Methylation Signatures. Front Genet. Frontiers; 2012 Apr 10;3(54).

70. Semmler A, Heese P, Stoffel-Wagner B, Muschler M, Heberlein A, Bigler L, et al. Alcohol abuse and cigarette smoking are associated with global DNA hypermethylation: Results from the German Investigation on Neurobiology in Alcoholism (GINA). Alcohol. Elsevier; 2015 Mar 1;49(2):97–101.

71. Bönsch D, Lenz B, Reulbach U, Kornhuber J, Bleich S. Homocysteine associated genomic DNA hypermethylation in patients with chronic alcoholism. J Neural Transm. Springer-Verlag; 2004 Dec;111(12):1611–6.

72. Zhang H, Wang F, Kranzler HR, Zhao H, Gelernter J. Profiling of Childhood Adversity-Associated DNA Methylation Changes in Alcoholic Patients and Healthy Controls. Marsit CJ, editor. PLOS ONE. Public Library of Science; 2013 Jun 14;8(6):e65648.

73. Zhao R, Zhang R, Li W, Liao Y, Tang J, Miao Q, et al. Genome-wide DNA methylation patterns in discordant sib pairs with alcohol dependence. Asia-Pacific Psychiatry. John Wiley & Sons, Ltd; 2013 Mar 1;5(1):39–50.

74. Liu C, Marioni RE, Hedman ÅK, Pfeiffer L, Tsai P-C, Reynolds LM, et al. A DNA methylation biomarker of alcohol consumption. Molecular Psychiatry 2016 23:2. Nature Publishing Group; 2018 Feb 1;23(2):422–33.

75. Illum LRH, Bak ST, Lund S, Nielsen AL. DNA methylation in epigenetic inheritance of metabolic diseases through the male germ line. Journal of Molecular Endocrinology. 2018;60:R39–R56.

76. Morgan HD, Santos F, Green K, Dean W, Reik W. Epigenetic reprogramming in mammals. Human Molecular Genetics. Oxford University Press; 2005 Apr 15;14(suppl_1):R47–R58.

77. Tang WWC, Dietmann S, Irie N, Leitch HG, Floros VI, Bradshaw CR, et al. A Unique Gene Regulatory Network Resets the Human Germline Epigenome for Development. Cell. Cell Press; 2015 Jun 4;161(6):1453–67.

78. Guibert S, Forné T, Weber M. Global profiling of DNA methylation erasure in mouse primordial germ cells. Genome Res. Cold Spring Harbor Lab; 2012 Feb 22;22(4):633–41.

79. Lane N, Dean W, Erhardt S, Hajkova P, Surani A, Walter J, et al. Resistance of IAPs to methylation reprogramming may provide a mechanism for epigenetic inheritance in the mouse. genesis. John Wiley & Sons, Ltd; 2003 Feb 1;35(2):88–93.

80. Laird PW. Principles and challenges of genome-wide DNA methylation analysis. Nature Publishing Group. Nature Publishing Group; 2010 Mar 1;11(3):191–203.

81. Bock C, Tomazou EM, Brinkman AB, Müller F, Simmer F, Gu H, et al. Quantitative comparison of genome-wide DNA methylation mapping technologies. Nature Biotechnology. Nature Publishing Group; 2010 Sep 19;28(10):1106–14.

82. Costello JF, Frühwald MC, Smiraglia DJ, Rush LJ, Robertson GP, Gao X, et al. Aberrant CpG-island methylation has non-random and tumour-type–specific patterns. Nature Genetics. Nature Publishing Group; 2000 Feb 1;24(2):132–8.

83. Susan JC, Harrison J, Paul CL, Frommer M. High sensitivity mapping of methylated cytosines. Nucleic Acids Research. Oxford University Press; 1994 Aug 11;22(15):2990–7.

84. Zuo T, Tycko B, Liu T-M, Lin H-JL, Huang TH-M. Methods in DNA methylation profiling. Epigenomics. 2009 Dec;1(2):331–45.

85. Touleimat N, Tost J. Complete pipeline for Infinium® Human Methylation 450K BeadChip data processing using subset quantile normalization for accurate DNA methylation estimation. http://dxdoiorg/102217/epi1221. Future Medicine Ltd London, UK; 2012 Jun 12;4(3):325–41.

86. Bibikova M, Le J, Barnes B, Saedinia-Melnyk S, 2009. Genome-wide DNA methylation profiling using Infinium® assay. Epigenomics. 2009;1(1):177–200.

87. Weisenberger DJ, Campan M, Long TI, Kim M, Woods C, Fiala E, et al. Analysis of repetitive element DNA methylation by MethyLight. Nucleic Acids Research. Oxford University Press; 2005;33(21):6823–36.

88. Yi SH, Xu LC, Mei K, Yang RZ, Huang DX. Isolation and identification of age-related DNA methylation markers for forensic age-prediction. Forensic Science International: Genetics. Elsevier; 2014 Jul;11:117–25.

89. Giuliani C, Cilli E, Bacalini MG, Pirazzini C, Sazzini M, Gruppioni G, et al. Inferring chronological age from DNA methylation patterns of human teeth. Am J Phys Anthropol. 2015 Dec 15;159:585–95.

90. Claus R, Wilop S, Hielscher T, Sonnet M, Dahl E, Galm O, et al. A systematic comparison of quantitative high-resolution DNA methylation analysis and methylation-specific PCR. Epigenetics. 2014 Oct 27;7(7):772–80.

91. Grunau C, Clark SJ, Rosenthal A. Bisulfite genomic sequencing: systematic investigation of critical experimental parameters. Nucleic Acids Research. Oxford University Press; 2001 Jul 1;29(13):e65–5.

92. Lee HY, Jung S-E, Oh YN, Choi A, Yang WI, Shin K-J. Epigenetic age signatures in the forensically relevant body fluid of semen: a preliminary study. Forensic Science International: Genetics. Elsevier; 2015 Jan 11;19:28–34.

93. Lee HY, An JH, Jung SE, Oh YN, Lee EY, Science ACF, et al. Genome-wide methylation profiling and a multiplex construction for the identification of body fluids using epigenetic markers. Forensic Science International: Genetics. 2015;17:17–24.

94. Koch CM, Wagner W. Epigenetic-aging-signature to determine age in different tissues. Aging (Albany NY). 2011 Oct;3(10):1018–27.

95. Steemers FJ, Chang W, Lee G, Barker DL, Shen R, Gunderson KL. Whole-genome genotyping with the single-base extension assay. Nature Methods. Nature Publishing Group; 2006 Jan 1;3(1):31–3.

96. Michael KL, Taylor LC, Schultz SL, Walt DR. Randomly Ordered Addressable High-Density Optical Sensor Arrays. Anal Chem. American Chemical Society; 1998 Apr;70(7):1242–8.

97. Bibikova M, Lin Z, Zhou L, Chudin E, Garcia EW, Wu B, et al. High-throughput DNA methylation profiling using universal bead arrays. Genome Res. Cold Spring Harbor Lab; 2006 Mar 1;16(3):383–93.

98. Bibikova M, Barnes B, Tsan C, Ho V, Klotzle B, Le JM, et al. High density DNA methylation array with single CpG site resolution. Genomics. 2011 Oct;98(4):288–95.

99. Lizio M, Harshbarger J, Shimoji H, Severin J, Kasukawa T, Sahin S, et al. Gateways to the FANTOM5 promoter level mammalian expression atlas. Genome Biology. BioMed Central; 2015 Jan 5;16(1):22.

100. Siggens L, Ekwall K. Epigenetics, chromatin and genome organization: recent advances from the ENCODE project. Journal of Internal Medicine. 2014 Sep 1;276(3):201–14.

101. Pidsley R, Zotenko E, Peters TJ, Lawrence MG, Risbridger GP, Molloy P, et al. Critical evaluation of the Illumina MethylationEPIC BeadChip microarray for whole-genome DNA methylation profiling. Genome Biology. BioMed Central; 2016 Oct 7;17(1):208.

102. Stirzaker C, Taberlay PC, Statham AL, Clark SJ. Mining cancer methylomes: prospects and challenges. Trends Genet. Elsevier Current Trends; 2014 Feb 1;30(2):75–84.

103. Relton CL, Gaunt T, McArdle W, Ho K, Duggirala A, Shihab H, et al. Data Resource Profile: Accessible Resource for Integrated Epigenomic Studies (ARIES). Int J Epidemiol. Oxford University Press; 2015 Aug 1;44(4):1181–90.

104. Wilhelm-Benartzi CS, Koestler DC, Karagas MR, Flanagan JM, Christensen BC, Kelsey KT, et al. Review of processing and analysis methods for DNA methylation array data. British Journal of Cancer. Nature Publishing Group; 2013 Aug 27;109(6):1394–402.

105. Du P, Zhang X, Huang C-C, Jafari N, Kibbe WA, Hou L, et al. Comparison of Beta-value and M-value methods for quantifying methylation levels by microarray analysis. BMC Bioinformatics 2010 11:1. BioMed Central; 2010 Nov 30;11(1):587.

106. Horvath S. DNA methylation age of human tissues and cell types. Genome Biology. BioMed Central Ltd; 2013 Oct 21;14(10):R115.

107. Dedeurwaerder S, Defrance M, Calonne E, Denis H, Sotiriou C, Fuks F. Evaluation of the Infinium Methylation 450K technology. Epigenomics. Future Medicine Ltd London, UK; 2011 Nov 25;3(6):771–84.

108. Teschendorff AE, Marabita F, Lechner M, Bartlett T, Tegnér J, Gomez-Cabrero D, et al. A beta-mixture quantile normalization method for correcting probe design bias in Illumina Infinium 450 k DNA methylation data. Bioinformatics. Oxford University Press; 2013 Jan 15;29(2):189–96.

109. Maksimovic J, Gordon L, Oshlack A. SWAN: Subset-quantile Within Array Normalization for Illumina Infinium HumanMethylation450 BeadChips. Genome Biology. BioMed Central; 2012 Jun 15;13(6):R44.

110. Marabita F, Almgren M, Lindholm ME, Ruhrmann S, Fagerström-Billai F, Jagodic M, et al. An evaluation of analysis pipelines for DNA

methylation profiling using the Illumina HumanMethylation450 BeadChip platform. Epigenetics. 2014 Oct 27;8(3):333–46.

111. Vidaki A, Ballard D, Aliferi A, Miller TH, Barron LP, Court DS. DNA methylation-based forensic age prediction using artificial neural networks and next generation sequencing. Forensic Science International: Genetics. Elsevier; 2017;0(0):225–36.

112. Jenkins TG, Aston KI, Pfluger C, Cairns BR, Carrell DT. Next generation bisulfite sequencing reveals consistent population-wide regional sperm DNA methylation alterations with age. Fertility and Sterility. Elsevier; 2014 Sep 1;102(3):e100–1.

113. An JH, Choi A, Shin K-J, Yang WI, Lee HY. DNA methylation-specific multiplex assays for body fluid identification. Int J Legal Med. Springer-Verlag; 2012 Jun 1;127(1):35–43.

114. Stowell LI, Sharman LE, Hamel K. An enzyme-linked immunosorbent assay (ELISA) for prostate-specific antigen. Forensic Science International. 1991;50(1):125–38.

115. Itoh Y, Matsuzawa S. Detection of human hemoglobin A (HbA) and human hemoglobin F (HbF) in biological stains by microtiter latex agglutination-inhibition test. Forensic Science International. 1990;47(1):79–89.

116. Khaldi N, Miras A, Botti K, Benali L, Gromb S. Evaluation of Three Rapid Detection Methods for the Forensic Identification of Seminal Fluid in Rape Cases. Journal of Forensic Science. ASTM International; 2004 May 5;49(4):1–5.

117. Zubakov D, Boersma AWM, Choi Y, van Kuijk PF, Wiemer EAC, Kayser M. MicroRNA markers for forensic body fluid identification obtained from microarray screening and quantitative RT-PCR confirmation. Int J Legal Med. Springer-Verlag; 2010;124(3):217–26.

118. Bauer M. RNA in forensic science. Forensic Science International: Genetics. Elsevier; 2007 Mar;1(1):69–74.

119. Park S-M, Park S-Y, Kim J-H, Kang T-W, Park J-L, Woo K-M, et al. Genome-wide mRNA profiling and multiplex quantitative RT-PCR for forensic body fluid identification. Forensic Science International: Genetics. Elsevier Ireland Ltd; 2013 Jan 1;7(1):143–50.

120. Setzer M, Juusola J, Ballantyne J. Recovery and Stability of RNA in Vaginal Swabs and Blood, Semen, and Saliva Stains. Journal of Forensic Science. Blackwell Publishing Ltd; 2008 Mar;53(2):296–305.

121. Ohgane J, Yagi S, K S. Epigenetics: the DNA methylation profile of tissue-dependent and differentially methylated regions in cells. Placenta. 2008;29(SUPPL.):29–35.

122. Lee HY, Jung S-E, Lee EH, Yang WI, Shin K-J. DNA methylation profiling for a confirmatory test for blood, saliva, semen, vaginal fluid and menstrual blood. Forensic Science International: Genetics. Elsevier Ireland Ltd; 2016 Sep 1;24:75–82.

123. Forat S, Huettel B, Reinhardt R, Fimmers R, Haidl G, Denschlag D, et al. Methylation Markers for the Identification of Body Fluids and Tissues from Forensic Trace Evidence. Kim J-W, editor. PLOS ONE. Public Library of Science; 2016 Feb 1;11(5):e0147973–19.

124. Alkass K, Buchholz BA, Ohtani S, Yamamoto T, Druid H. Age Estimation in Forensic Sciences. Molecular Proteomics. 2010 May;9(5):1022–30.

125. Thevissen PW, Galiti D, Willems G. Human dental age estimation combining third molar(s) development and tooth morphological age predictors. Int J Legal Med. Springer-Verlag; 2012 Aug 12;126(6):883–7.

126. Helfman PM, Bada JL. Aspartic acid racemization in tooth enamel from living humans. PNAS. National Acad Sciences; 1975 Aug;72(8):2891–4.

127. Spalding KL, Buchholz BA, Bergman L-E, Druid H, Frisén J. Forensics: Age written in teeth by nuclear tests. Nature. Nature Publishing Group; 2005 Sep 15;437(7057):333–4.

128. Buchholz BA, Spalding KL. Year of birth determination using radiocarbon dating of dental enamel. Bailey M, editor. Surf Interface Anal. John Wiley & Sons, Ltd; 2010 May;42(5):398–401.

129. Uno KT, Quade J, Fisher DC. Bomb-curve radiocarbon measurement of recent biologic tissues and applications to wildlife forensics and stable isotope (paleo) ecology. 2013. pp. 117336–111741.

130. Liu L, Wylie RC, Andrews LG, Tollefsbol TO. Aging, cancer and nutrition: the DNA methylation connection. Mechanisms of Ageing and Development. 2003 Dec;124(10-12):989–98.

131. Wilson VL, Smith RA, Ma S, Cutler RG. Genomic 5-methyldeoxycytidine decreases with age. The Journal of Biological Chemistry. 1987;262(21):9948–51.

132. Drinkwater RD, Blake TJ, Morley AA, Turner DR. Human lymphocytes aged in vivo have reduced levels of methylation in transcriptionally active and inactive DNA. Mutation Research/DNAging. Elsevier; 1989 Jan 1;219(1):29–37.

133. Maegawa S, Hinkal G, Kim HS, Shen L, Zhang L, Zhang J, et al. Widespread and tissue specific age-related DNA methylation changes in mice. Genome Res. Cold Spring Harbor Lab; 2010 Mar 1;20(3):332–40.

134. Teschendorff AE, Menon U, Gentry-Maharaj A, Ramus SJ, Weisenberger DJ, Shen H, et al. Age-dependent DNA methylation of genes that are suppressed in stem cells is a hallmark of cancer. Genome Res. Cold Spring Harbor Lab; 2010 Apr;20(4):440–6.

135. Wang S-C, Oelze B, Schumacher A. Age-Specific Epigenetic Drift in Late-Onset Alzheimer's Disease. Toland AE, editor. PLOS ONE. Public Library of Science; 2008 Jul 16;3(7):e2698.

136. Hannum G, Guinney J, Zhao L, Zhang L, Hughes G, Sadda S, et al. Genome-wide methylation profiles reveal quantitative views of human aging rates. Molecular Cell. Elsevier; 2013 Jan;49(2):359–67.

137. Marcucci G, Metzeler KH. Age-related prognostic impact of different types of DNMT3A mutations in adults with primary cytogenetically normal acute myeloid leukemia. Journal of Clinical Oncology. Journal of …; 2012;30(7):742–50.

138. Wojdacz TK, Hansen LL. Techniques Used in Studies of Age-Related DNA Methylation Changes. Annals of the New York Academy of …. 2006;1067:479–87.

139. Christensen BC, Houseman EA, Marsit CJ, Zheng S, Wrensch MR, Wiemels JL, et al. Aging and Environmental Exposures Alter Tissue-Specific DNA Methylation Dependent upon CpG Island Context. Schübeler D, editor. PLOS Genet. Public Library of Science; 2009 Aug 14;5(8):e1000602.

140. Thompson RF, Atzmon G, Gheorghe C, Liang HQ, Lowes C, Greally JM, et al. Tissue-specific dysregulation of DNA methylation in aging. Aging Cell. Blackwell Publishing Ltd; 2010 Aug 1;9(4):506–18.

141. Day K, Waite LL, Thalacker-Mercer A, West A, Bamman MM, Brooks JD, et al. Differential DNA methylation with age displays both common and dynamic features across human tissues that are influenced by CpG landscape. Genome Biology. BioMed Central Ltd; 2013 Sep 13;14(9):R102.

142. Chu MW, Siegmund KD, Eckstam CL, Kim JY, Yang AS, Kanel GC, et al. Lack of increases in methylation at three CpG-rich genomic loci in non-mitotic adult tissues during aging. BMC Medical Genetics 2007 8:1. BioMed Central; 2007 Jul 31;8(1):50.

143. Bekaert B, Kamalandua A, Zapico SC, Van de Voorde W, Decorte R. A selective set of DNA-methylation markers for age determination of blood, teeth and buccal samples. Forensic Science International: Genetics Supplement Series. 2015 Dec 1;5:e144–5.

144. Huang Y, Yan J, Hou J, Fu X, Li L, Hou Y. Developing a DNA methylation assay for human age prediction in blood and bloodstain. Forensic Science International: Genetics. Elsevier; 2015 Jul;17:129–36.

145. Weidner CI, Lin Q, Koch CM, Eisele L, Beier F, Ziegler P, et al. Aging of blood can be tracked by DNA methylation changes at just three CpG sites. Genome Biology. BioMed Central Ltd; 2014 Feb 3;15(2):R24.

146. Florath I, Butterbach K, Müller H, Bewerunge-Hudler M, Brenner H. Cross-sectional and longitudinal changes in DNA methylation with age: an epigenome-wide analysis revealing over 60 novel age-associated CpG sites. Human Molecular Genetics. Oxford University Press; 2014 Mar 1;23(5):1186–201.

147. Freire-Aradas A, Phillips C, Mosquera-Miguel A, Girón-Santamaría L, Gómez-Tato A, de Cal MC, 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 Science International: Genetics. Elsevier Ireland Ltd; 2016 Sep 1;24:65–74.

148. Zbieć-Piekarska R, Spólnicka M, Kupiec T, Makowska Ż, Spas A, Parys-Proszek A, et al. Examination of DNA methylation status of the ELOVL2 marker may be useful for human age prediction in forensic science. Forensic Science International: Genetics. Elsevier; 2015 Jan;14:161–7.

149. Bekaert B. Improved age determination of blood and teeth samples using a selected set of DNA methylation markers. Epigenetics. 2015 Jan 1;10(10):922–30.

150. Mawlood SK, Dennany L, Watson N, Pickard BS. The EpiTect Methyl qPCR Assay as novel age estimation method in forensic biology. Forensic Science International. Elsevier; 2016 Jul 1;264:132–8.

151. Kim EJ, Kim M-K, Jin X-J, Oh J-H, Kim JE, Chung JH. Skin Aging and Photoaging Alter Fatty Acids Composition, Including 11,14,17-

eicosatrienoic Acid, in the Epidermis of Human Skin. Journal of Korean Medical Science. 2010 Jun 1;25(6):980–3.

152. Garagnani P, Bacalini MG, Pirazzini C, Gori D, Giuliani C, Mari D, et al. Methylation of ELOVL2 gene as a new epigenetic marker of age. Aging Cell. 2012 Dec 1;11(6):1132–4.

153. Johansson Å, Enroth S, Gyllensten U. Continuous Aging of the Human DNA Methylome Throughout the Human Lifespan. Suter CM, editor. PLOS ONE. Public Library of Science; 2013 Jun 27;8(6):e67378.

154. Rönn T, Volkov P, Gillberg L, Kokosar M, Perfilyev A, Jacobsen AL, et al. Impact of age, BMI and HbA1c levels on the genome-wide DNA methylation and mRNA expression patterns in human adipose tissue and identification of epigenetic biomarkers in blood. Human Molecular Genetics. Oxford University Press; 2015 Jul 1;24(13):3792–813.

155. Soares Bispo Santos Silva D, Antunes J, Balamurugan K, Duncan G, Sampaio Alho C, McCord B. Evaluation of DNA methylation markers and their potential to predict human aging. Electrophoresis. 2015 Jul 14;36(15):1775–80.

156. Bocklandt S, Lin W, Sehl ME, one FSP, 2011. Epigenetic predictor of age. PLOS ONE. 2011;6(6):e14821.

157. Hong SR, Jung S-E, Lee EH, Shin K-J, Yang WI, Lee HY. DNA methylation-based age prediction from saliva: High age predictability by combination of 7 CpG markers. Forensic Science International: Genetics. 2017 Jul;29:118–25.

158. R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.URL http://www.R-project.org/. 2013.

159. Gentleman RC, Carey VJ, Bates DM, Ben Bolstad, Dettling M, Dudoit S, et al. Bioconductor: open software development for computational biology and bioinformatics. Genome Biology. BioMed Central; 2004 Sep 1;5(10):1–16.

160. Edgar R, Domrachev M. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. Nucleic Acids Research. 2002 Jan 1;30(1):207–10.

161. Weisenberger DJ. Characterizing DNA methylation alterations from The Cancer Genome Atlas. J Clin Invest. American Society for Clinical Investigation; 2014 Jan 2;124(1):17–23.

162. Athar A, Füllgrabe A, George N, Iqbal H, Huerta L, Ali A, et al. ArrayExpress update – from bulk to single-cell expression data. Nucleic Acids Research. 2019 Jan 8;47(D1).

163. Morris TJ, Butcher LM, Feber A, Teschendorff AE, Chakravarthy AR, Wojdacz TK, et al. ChAMP: 450k Chip Analysis Methylation Pipeline. Bioinformatics. Oxford University Press; 2014 Feb 1;30(3):428–30.

164. Assenov Y, Müller F, Lutsik P, Walter J, Lengauer T, Bock C. Comprehensive analysis of DNA methylation data with RnBeads. Nature Methods. Nature Research; 2014 Nov 1;11(11):1138–40.

165. Pidsley R, Wong CCY, Volta M, Lunnon K, Mill J, Schalkwyk LC. A data-driven approach to preprocessing Illumina 450K methylation

array data. BMC Genomics 2014 15:1. BioMed Central; 2013 Dec 1;14(1):293.

166. Price EM, Cotton AM, Lam LL, Farré P, Emberly E, Brown CJ, et al. Additional annotation enhances potential for biologically-relevant analysis of the Illumina Infinium HumanMethylation450 BeadChip array. Epigenetics & Chromatin. BioMed Central; 2013 Mar 3;6(1):1.

167. Teschendorff AE, Relton CL. Statistical and integrative system-level analysis of DNA methylation data. Nature Publishing Group. Nature Publishing Group; 1AD;:1–19.

168. Sun Z, Chai H, Wu Y, White WM, Donkena KV, Klein CJ, et al. Batch effect correction for genome-wide methylation data with Illumina Infinium platform. BMC Medical Genomics 2011 4:1. BioMed Central; 2011 Dec 16;4(1):1.

169. Jones MJ, Islam SA, Edgar RD, Kobor MS. Adjusting for Cell Type Composition in DNA Methylation Data Using a Regression-Based Approach. In: Population Epigenetics. New York, NY: Humana Press, New York, NY; 2015. pp. 99–106. (Methods and Protocols; vol. 1589).

170. Jaffe AE, Irizarry RA. Accounting for cellular heterogeneity is critical in epigenome-wide association studies. Genome Biology. BioMed Central; 2014 Feb 1;15(2):R31.

171. Tan Q, Heijmans BT, Hjelmborg JVB, Soerensen M, Christensen K, Christiansen L. Epigenetic drift in the aging genome: a ten-year follow-up in an elderly twin cohort. Int J Epidemiol. Narnia; 2016 Aug 1;45(4):1146–58.

172. Houseman EA, Accomando WP, Koestler DC, Christensen BC, Marsit CJ, Nelson HH, et al. DNA methylation arrays as surrogate measures of cell mixture distribution. BMC Bioinformatics 2010 11:1. BioMed Central; 2012 Dec 1;13(1):86.

173. Aryee MJ, Jaffe AE, Corrada-Bravo H, Ladd-Acosta C, Feinberg AP, Hansen KD, et al. Minfi: a flexible and comprehensive Bioconductor package for the analysis of Infinium DNA methylation microarrays. Bioinformatics. Narnia; 2014 May 15;30(10):1363–9.

174. Howitt D, Cramer D. Introduction to Statistics in Psychology. Pearson Education; 2007. 1 p.

175. Hauke J, Kossowski T. Comparison of Values of Pearson“s and Spearman”s Correlation Coefficients on the Same Sets of Data - ProQuest. Quaestiones geographicae. 2011.

176. Dabney A, Storey JD, Warnes GR. Q-value estimation for false discovery rate control. R package version. 2010.

177. Quigen. mini kit and QIAamp DNA blood mini kit handbook. 2016.

178. Krueger F, Andrews SR. Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications. Bioinformatics. Narnia; 2011 Jun 1;27(11):1571–2.

179. Zbieć-Piekarska R, Spólnicka M, Kupiec T, Parys-Proszek A, Makowska Ż, Pałeczka A, et al. Development of a forensically useful age prediction method based on DNA methylation analysis. Forensic Science International: Genetics. Elsevier; 2015 Jul;17:173–9.

180. Gopalan S, Carja O, Fagny M, Patin E, Myrick JW, McEwen LM, et al. Trends in DNA Methylation with Age Replicate Across Diverse Human Populations. Genetics. 2017 Jul 6;206(3):1659–74.

181. Tserel L, Limbach M, Saare M, Kisand K, Metspalu A, Milani L, et al. CpG sites associated with NRP1, NRXN2 and miR-29b-2 are hypomethylated in monocytes during ageing. Immunity & Ageing 2014 11:1. BioMed Central; 2014 Dec 1;11(1):1.

182. Li S, Christiansen L, Christensen K, Kruse TA, Redmond P, Marioni RE, et al. Identification, replication and characterization of epigenetic remodelling in the aging genome: a cross population analysis. Scientific Reports. Nature Publishing Group; 2017 Aug 15;7(1):8183.

183. Baglietto L, Ponzi E, Haycock P, Hodge A, Bianca Assumma M, Jung CH, et al. DNA methylation changes measured in pre-diagnostic peripheral blood samples are associated with smoking and lung cancer risk. International Journal of Cancer. 2017 Jan 1;140(1):50–61.

184. Marttila S, Kananen L, Häyrynen S, Jylhävä J, Nevalainen T, Hervonen A, et al. Ageing-associated changes in the human DNA methylome: genomic locations and effects on gene expression. BMC Genomics 2014 15:1. BioMed Central; 2015 Dec 1;16(1):179.

185. Freire-Aradas A, Phillips C, Lareu MV. Forensic individual age estimation with DNA: From initial approaches to methylation tests. Forensic Science Rev. 2017;29:121.

186. Eipel M, Mayer F, Arent T, Ferreira MRP, Birkhofer C, Gerstenmaier U, et al. Epigenetic age predictions based on buccal swabs are more precise in combination with cell type-specific DNA methylation

signatures. Aging (Albany NY). Impact Journals, LLC; 2016 May 1;8(5):1034–48.

187. Tsuji A, Ishiko A, Takasaki T, International NIFS, 2002. Estimating age of humans based on telomere shortening. Elsevier. 2002 May;126(3):197–9.

188. Cortopassi GA, Shibata D, Soong NW, Arnheim N. A pattern of accumulation of a somatic deletion of mitochondrial DNA in aging human tissues. PNAS. National Academy of Sciences; 1992 Aug 15;89(16):7370–4.

189. Slieker RC, Relton CL, Gaunt TR, Slagboom PE, Heijmans BT. Age-related DNA methylation changes are tissue-specific with ELOVL2 promoter methylation as exception. Epigenetics & Chromatin. BioMed Central; 2018 May 30;:1–11.

190. Rebhan M, Chalifa-Caspi V, Prilusky J, Lancet D. GeneCards: integrating information about genes, proteins and diseases. Trends Genet. Elsevier; 1997 Apr 1;13(4):163.

191. Spiers H, Hannon E, Wells S, Williams B, Fernandes C, Mill J. Age-associated changes in DNA methylation across multiple tissues in an inbred mouse model. Mechanisms of Ageing and Development. Elsevier Ireland Ltd; 2016 Mar 1;154:20–3.

192. Aliferi A, Ballard D, Gallidabino MD, Thurtle H, Barron L, Syndercombe Court D. DNA methylation-based age prediction using massively parallel sequencing data and multiple machine learning models. Forensic Science International: Genetics. Elsevier; 2018 Nov 1;37:215–26.

193. Parson W. Age Estimation with DNA: From Forensic DNA Fingerprinting to Forensic (Epi)Genomics: A Mini-Review. GER. Karger Publishers; 2018;64:326–32.

194. McEwen LM, Jones MJ, Lin DTS, Edgar RD, Husquin LT, MacIsaac JL, et al. Systematic evaluation of DNA methylation age estimation with common preprocessing methods and the Infinium MethylationEPIC BeadChip array. Clinical Epigenetics. BioMed Central; 2018 Dec 1;10(1):123.

195. Dou J, Schmidt RJ, Benke KS, Newschaffer C, Hertz-Picciotto I, Croen LA, et al. Cord blood buffy coat DNA methylation is comparable to whole cord blood methylation. Taylor & Francis; 2018 Feb 21;:1–10.

196. Zaimi I, Pei D, Koestler DC, Marsit CJ, De Vivo I, Tworoger SS, et al. Variation in DNA methylation of human blood over a 1-year period using the Illumina MethylationEPIC array. Epigenetics. Taylor & Francis; 2018 Dec 4;13(10-11):1056–71.

197. Curtis SW, Cobb DO, Kilaru V, Terrell ML, Kennedy EM, Marder ME, et al. Exposure to polybrominated biphenyl (PBB) associates with genome-wide DNA methylation differences in peripheral blood. Epigenetics. Taylor & Francis; 2019 Feb 6;14(1):52–66.

198. Fortin J, Triche T, Hansen K. Preprocessing, normalization and integration of the Illumina HumanMethylationEPIC array with minfi | Bioinformatics | Oxford Academic. Bioinformatics. 2017;33(4):558–60.

199. Johnson WE, Li C, Rabinovic A. Adjusting batch effects in microarray expression data using empirical Bayes methods. Biostatistics. Narnia; 2007 Jan 1;8(1):118–27.

200. Jung S-E, Lim SM, Hong SR, Lee EH, Shin K-J, Lee HY. DNA methylation of the ELOVL2, FHL2, KLF14, C1orf132/MIR29B2C, and TRIM59 genes for age prediction from blood, saliva, and buccal swab samples. Forensic Science International: Genetics. Elsevier; 2019 Jan 1;38:1–8.

201. Davenport EC, Pendolino V, Kontou G, McGee TP, Sheehan DF, López-Doménech G, et al. An Essential Role for the Tetraspanin LHFPL4 in the Cell-Type-Specific Targeting and Clustering of Synaptic GABAA Receptors. Cell Rep. 2017 Oct 3;21(1):70–83.

202. Wang SS, Smiraglia DJ, Wu Y-Z, Ghosh S, Rader JS, Cho KR, et al. Identification of novel methylation markers in cervical cancer using restriction landmark genomic scanning. Cancer Res. American Association for Cancer Research; 2008 Apr 1;68(7):2489–97.

203. Patel N, Itakura T, Gonzalez JM, Schwartz SG, Fini ME. GPR158, an orphan member of G protein-coupled receptor Family C: glucocorticoid-stimulated expression and novel nuclear role. Languino LR, editor. PLOS ONE. Public Library of Science; 2013;8(2):e57843.

204. Musikacharoen T, Bandow K, Kakimoto K, Kusuyama J, Onishi T, Yoshikai Y, et al. Functional involvement of dual specificity phosphatase 16 (DUSP16), a c-Jun N-terminal kinase-specific phosphatase, in the regulation of T helper cell differentiation. The Journal of Biological Chemistry. American Society for Biochemistry and Molecular Biology; 2011 Jul 15;286(28):24896–905.

205. Rastogi D, Suzuki M, Greally JM. Differential epigenome-wide DNA methylation patterns in childhood obesity-associated asthma. Scientific Reports. Nature Publishing Group; 2013;3(1):2164.

206. Baer C, Claus R, Frenzel LP, Zucknick M, Park YJ, Gu L, et al. Extensive promoter DNA hypermethylation and hypomethylation is associated with aberrant microRNA expression in chronic lymphocytic leukemia. Cancer Res. American Association for Cancer Research; 2012 Aug 1;72(15):3775–85.

207. Hoed den M, Eijgelsheim M, Esko T, Brundel BJJM, Peal DS, Evans DM, et al. Identification of heart rate–associated loci and their effects on cardiac conduction and rhythm disorders. Nature Genetics. Nature Publishing Group; 2013 Jun 1;45(6):621–31.

208. Weinhold L, Wahl S, Schmid M. A Statistical Model for the Analysis of Beta Values in DNA Methylation Studies. 2016.

209. Sun W, He T, Qin C, Qiu K, Zhang X, Luo Y, et al. A potential regulatory network underlying distinct fate commitment of myogenic and adipogenic cells in skeletal muscle. Scientific Reports. 2017 Mar 9;7(1):11G.

210. Miller-Delaney SFC, Bryan K, Das S, McKiernan RC, Bray IM, Reynolds JP, et al. Differential DNA methylation profiles of coding and non-coding genes define hippocampal sclerosis in human temporal lobe epilepsy. Brain. 2015 Mar;138(Pt 3):616–31.

211. Verma D, Ekman A-K, Bivik Eding C, Enerbäck C. Genome-Wide DNA Methylation Profiling Identifies Differential Methylation in Uninvolved Psoriatic Epidermis. J Invest Dermatol. 2018 May;138(5):1088–93.

212. Fluhr S, Boerries M, Busch H, Symeonidi A, Witte T, Lipka DB, et al. CREBBP is a target of epigenetic, but not genetic, modification in juvenile myelomonocytic leukemia. Clinical Epigenetics. BioMed Central; 2016;8(1):50.

213. Musialik E, Bujko M, Kober P, Wypych A, Gawle-Krawczyk K, Matysiak M, et al. Promoter methylation and expression levels of selected hematopoietic genes in pediatric B-cell acute lymphoblastic leukemia. Blood Res. 2015 Mar;50(1):26–32.

214. Park J-L, Kim JH, Seo E, Bae DH, Kim S-Y, Lee H-C, et al. Identification and evaluation of age-correlated DNA methylation markers for forensic use. Forensic Science International: Genetics. Elsevier Ireland Ltd; 2016 Jul 1;23:64–70.

215. Naue J, Hoefsloot HCJ, Mook ORF, Rijlaarsdam-Hoekstra L, van der Zwalm MCH, Henneman P, et al. Chronological age prediction based on DNA methylation: Massive parallel sequencing and random forest regression. Forensic Science International: Genetics. Elsevier; 2017 Nov 1;31:19–28.

216. Lee JW, Choung CM, Jung JY, Lee HY, Lim S-K. A validation study of DNA methylation-based age prediction using semen in forensic casework samples. Legal Medicine. Elsevier; 2018 Mar 1;31:74–7.

217. Schroeder JW, Conneely KN, Cubells JF, Kilaru V, Newport DJ, Knight BT, et al. Neonatal DNA methylation patterns associate with gestational age. Epigenetics. Taylor & Francis; 2011 Dec 1;6(12):1498–504.

218. Breitling LP, Saum K-U, Perna L, Ben Schöttker, Holleczek B, Brenner H. Frailty is associated with the epigenetic clock but not with telomere length in a German cohort. Clinical Epigenetics. BioMed Central; 2016 Dec 1;8(1):21.

219. Marioni RE, Shah S, McRae AF, Ritchie SJ, Muniz-Terrera G, Harris SE, et al. The epigenetic clock is correlated with physical and cognitive

fitness in the Lothian Birth Cohort 1936. Int J Epidemiol. Narnia; 2015 Aug 1;44(4):1388–96.

220. Levine ME, Lu AT, Bennett DA, Horvath S. Epigenetic age of the pre-frontal cortex is associated with neuritic plaques, amyloid load, and Alzheimer's disease related cognitive functioning. Aging (Albany NY). Impact Journals, LLC; 2015 Dec 1;7(12):1198–211.

221. Chen BH, Marioni RE, Colicino E, Peters MJ, Ward-Caviness CK, Tsai P-C, et al. DNA methylation-based measures of biological age: meta-analysis predicting time to death. Aging (Albany NY). Impact Journals, LLC; 2016 Sep 1;8(9):1844–65.

Appendix A: Supplemental tables

Amplicon quality control report

An overview of the quality control (QC) tested assays for the 16 targeted regions harboring age related CpG sites. All primers were mixed and diluted to 2 µM each. All primers were then tested using Real-Time PCR with 1 ng of bisulfite-converted control DNA, in duplicate individual reactions. DNA melt analysis was performed to confirm the presence of a specific PCR product.

Primers that passed QC:

- Had average Cp values <40
- Duplicate Cps do not have a Cp difference >1
- Reached the plateau phase before the run ended at cycle 45
- Produced melting curves in the expected range for PCR products
- Product size matched expected values
- Had no amplification in No Template Control (NTC) reactions

1- **Illumina probe ID = cg00573770**

Dark gray = Regions not covered and not tested
Light gray = Regions of interest that were tested and did not pass QC
Light blue = Regions of interest covered by amplicons that passed QC

CG = Targeted CpG site
CG = Adjacent CpG sites

hg19_dna Range=chr2:145278285-145278685 5'pad=200 3'pad=200 strand=+

CTGCCCTTTCTCTTATTGTTATTTTTTTCTTTTTAGGTACCAGAGCCAGAAAAAAAAA
TGCTGCATGGGAGCTGCATCTTAGGGCATGTGTATTAGGGTGTGTGCATGATGAATTTCTG
GACTGGATCCCAATATTAAAAAGTAGTTTGGCATTTTAATAAAGGGTCTCTAAGTAAATTAAA
ATAGAACACT**CG**GTTGGC**CG**ATCTCTGAATCTCTCTACACCT**CG**GGGAGACCTCACTACAA
AGTAAGGGAGAGTGTGTAGGGAGGCAGGAGGAGAAA**CG**AGAAAGGCCATAGAGAAACTTA
GCAGGGAAGGGAGAGCATGATATTAAA**CG**GCATGGGGTCTAAGTTTTGGAGTATATATTTT
TTGATATTACAATTAAAGTAAGATTAAAAAGAAAAGA

2- **Illumina probe ID = cg04875128**

Dark gray = Regions not covered and not tested
Light gray = Regions of interest that were tested and did not pass QC
Light blue = Regions of interest covered by amplicons that passed QC

CG = Targeted CpG site
CG = Adjacent CpG sites

hg19_dna range=chr15:31775595-31776195 5'pad=300 3'pad=300 strand=-

cCGCGCAGACCGGAGACGGAGGGCGTGCCGGTCCCGGAGCGCGCCTCTCCGGG
CCCACCCACGCAGCTGGTGCTCAAGCTCAAGGAgCGgcCGagcccCGggccCGCGgcagggCGt
gCGgCGCGggCGgCGgCGggCGgcaCGgcctcccCGgggggaggCGCGCGgCGTGCGAGCGCCAG
CGGACCAGTGCCTGGCCGCAGCCCCC**CG**G**CG**CCAG**CGCG**CCAGAG**CG**TCATCCA**CG**TGC
AGG**CG**T**CG**GG**CGCGCG**GGA**CG**AGG**CG**TG**CGCG**C**CG**GC**CG**TGGGGG**CG**CTG**CG**GC**CG**T
G**CG**CCA**CG**TACC**CG**CAGCAGAAC**CG**CT**CG**CTGT**CG**T**CG**CAGAGCTACAgcc**CG**g**CGCGCG**
c**CG**c**CG**cccTG**CG**CAC**CG**TCAACA**CG**GT**CG**AGT**CG**CTgg**CGCGCGCG**gtgccCGgggccctacCG
ggCGCGgCGgggaCGgCGggggCGgcCGAGCACAAGTCGCAGACCTACACCAACGGCTTCGG
CGCCCTGCGCGACGGCCTGGAGTTCGCCGACGCCGACGCGCCGACCGCGCGCTCGAAC
GGTGAGTGCGGCCGTGGCGGCCCGGGGCCGGTGCAGCGGCGCTGCCAGCGCGAGAAC

3- **Illumina probe ID = cg06279276**

Dark gray = Regions not covered and not tested
Light gray = Regions of interest that were tested and did not pass QC
Light blue = Regions of interest covered by amplicons that passed QC

CG = Targeted CpG site
CG = Adjacent CpG sites

hg19_dna range=chr16:67183864-67184464 5'pad=300 3'pad=300 strand=-

GGGGAGGGCGGGGCTCGCTGCCCCCTGCTGCCGACTGCGACCCTTACAGGGGAG
GGAGGGCGCAGGCCGCGCGGAGATGAGGAGGAGGCTGCGCCTACGCAGGGACGCATTG
CTCACGCTGCTCCTTGGCGCCTCCCTGGGCCTCTTACTCTATGCGCAGCGCGACGGCGCG
GCCCCGACGGCGAGCGCGCCGCGAGGGCGAGGGAGGGCGGCACCGAGGCCCACCCCC
GGACCCCGCGCGTTCCAGTTACCCGACGCGGGTGCAGCCCCGCCGGCCTACGAAGGGGA
CACACCGGCGCCCCACGCCTACGGGACCCTTTGACTTCGCCGCTATTTGCGCGCCAA
GGACCAGCGGCGGTTTCCACTGCTCATTAACCAGCCGCACAAGTGCCGCGGCGACGGCG
CACCCGGTGGCCGCCCGGACCTGCTTATTGCTGTCAAGTCGGTGGCAGAGGACTTCGAGC
GGCGCCAAGCCGTGCGCCAGACGTGGGGCGCGGAGGGTCGCGTGCAGGGGGCGCTGGT
GCGCCGCGTGTTCTTGCTGGGCGTGCCCAGGGGCGCAGGCTCGGGCGGGGCCGACGAA
GTTGGGGAGGGCGCGCG

4- **Illumina probe ID = cg07365960**

Dark gray = Regions not covered and not tested
Light gray = Regions of interest that were tested and did not pass QC
Light blue = Regions of interest covered by amplicons that passed QC

CG = Targeted CpG site
CG = Adjacent CpG sites

hg19_dna range=chr17:72848235-72848835 5'pad=300 3'pad=300 strand=+

CGCCAGTAGCTGTGGGCGCCCAGGGCCAGAATGGCCACGCCGTCGCGCACCTTC
TGGCGCAGGCTGAGGCGCCAGCTCTCGGTGACGACGCTGATGAGGCCCACGGGGAAGGT
GGCGGGGGGCGCATCGGTGCTGCCCAGCGCCAGGTTGGGCACCAGCCACA**CG**TGGC**CG**
GGCCCCACCAGAC**CG**GCCTG**CG**C**CG**CCT**CG**G**CG**AAGAGCACCT**CG**GCCTCCT**CGCGCG**A
GCAGTAGGCCACAAACA**CG**GG**CGCG**T**CG**AGCTGG**CG**CAGCAGG**CG**CTG**CG**TG**CGCGCG**
CGCGGCCCTCC**CG**GCCCAGCTCCAGCGTGACCACGTCCAGCAGCCGCCAACTCACGTGG
CTGGCGTCGGCGACGGCGCGCACGCCCTCCAGGAAGAGCGCGTGGCCCGGGTGCAGGC
TGGTGATGACGGCGAAGGCGCTCCAGTCGTACTCTTCCAGCACCTTGAACAGCACCTGCA
GCTGCTGCTCCAGGGACACGCCCAGCTGCAGGAAGGCGGAGCCCGGCTCCTGGGGGCG
GGCGGGGCCTGAGCGGGGCGGGAGGGCCGAGCCCCTCCTCCCGCCCCTTCCCCGACCT
CGGCCCCTCCATCAGCTC

5- **Illumina probe ID = cg10501210**

Dark gray = Regions not covered and not tested
Light gray = Regions of interest that were tested and did not pass QC
Light blue = Regions of interest covered by amplicons that passed QC

CG = Targeted CpG site
CG = Adjacent CpG sites

hg19_dna range=chr1:207996820-207997220 5'pad=200 3'pad=200 strand=+

CAGCTGACACTAGGGAAAAGAAATTAAAGTGGGAAAAAACCCTCCCTCAGAGAAAT
AAATAGCAAAAATCGAGAAAGAAGGTGAGAAAGACAGAGCACCCACATACACAGAGACAGC
GCCCCTGATCCCAGCAAATACATACGTGGGGGAAGAAGGGGGTTACGCCATCAAGTCCTG
AAGCC**CG**T**CG**GACCACCCAT**CG**C**CG**CCTG**CG**CAGACCCAAATCTTGGTCC**CG**C**CG**TAAGG
TGC**CG**CAGTCC**CG**AATGTTCCAGAATTTGGTCCCATCAAACCCTCCACCGTCGCCCCACAA
CCTCTTGCTCCCACCCCTGCCCCCCACCACCACCCCACCTCCTCCCCACGGGAACCGCCC
GTGCCACCTTGCGTGTCATCTCCTAGCCTAGGCTACCCAGAGG

6- **Illumina probe ID = cg10804656**

Dark gray = Regions not covered and not tested
Light gray = Regions of interest that were tested and did not pass QC
Light blue = Regions of interest covered by amplicons that passed QC

CG = Targeted CpG site
CG = Adjacent CpG sites

hg19_dna range=chr10:22623260-22623660 5'pad=200 3'pad=200 strand=+

AAGAAACAGCCTCTCTCCTTTTCCTTATTTTCTAATTAGCATCTTACAGAGGAGTGG
AAACAGCTACAGCCCAGTCCCCTGCTCAAAACTGCGCCACCCCAGTTCGGCCCTGCTGGG
CGCGCGAGCCAAGGC**CGCG**GGGCAC**CG**GGAGGCCATTTTG**CGCG**TG**CG**CTGCT**CG**CCT**C**
GCGC**CG**CCCT**CG**GCTCTG**CG**GACT**CG**GATCC**CG**CCAAATTTGAA**CGCG**AGATTGTCAGGC
CCTGAGGGGCTTGAGGGGCGGGGGAACGACGCCGCTCTCCAAAGTTGGACCCCGTGGCG
AGCGGCGGCGACAGCCGGGTGCTCGCTGCCTCCCGAGGTGCTCCCTTTTCCCGCCGAAG
CCCTCCACAGCGGCAGGCCGAGGCGCAGCGACGTGTCCCTGTACCCC

7- **Illumina probe ID = cg16867657**

Dark gray = Regions not covered and not tested
Light gray = Regions of interest that were tested and did not pass QC
Light blue = Regions of interest covered by amplicons that passed QC

CG = Targeted CpG site
CG = Adjacent CpG sites

hg19_dna range=chr6:11044677-11045177 5'pad=200 3'pad=300 strand=+

CGccccctctcccacAGGGGCCTCGCCGGCCGCCGCGCCAGGAGGGCGCGCGGGGGA
GGGGCGCAGGGCAAGTGAggCGgCGccccCGcccctgCGgcctCGCGCGcccccTCCTGGGCGA
CCGACCTCGCCCTCGCGTCCGCGGCGTCCCCTGCCGGCCGGGCGGCGATTTGCAGGTCC
AGC**CG**G**CG**C**CG**GTTT**CGCG**C**G**G**CG**GCTCAA**CG**TCCA**CG**GAGCCCCAGGAATACCCACCC
GCTGCCCAGATCGGCAGCCGCTGCTGCGGGGAGAAGCAGTATCGTGCAGGGCGGGCACG
CTGGTCTTGCTTACAGTTGGGCTTCGGTGGGTTTGAAGCACACATTAGGGGGAAATGGCTC
TGTTCCTGCAGGTTTGCGCAGTCTGGGTTTCTTAGGTTTAGGGGGTTGGGTGGGTTTCTCT
GGGGGTGCGGTGGGAAGCGGATCAGTTCGGATAACGGCCCTGAGCAAGAGTCTCTGTCC
CGCTCCCGGCCTGACGCGGGG

8- **Illumina probe ID = cg23606718**

Dark gray = Regions not covered and not tested
Light gray = Regions of interest that were tested and did not pass QC
Light blue = Regions of interest covered by amplicons that passed QC

CG = Targeted CpG site
CG = Adjacent CpG sites

hg19_dna range=chr2:131513627-131514227 5'pad=300 3'pad=300 strand=+

AAAGCCGGCTCGAGCGCGGCCCGAGCGGCGGCGCAGGGAGGCGGCAGCCTGGC
GGAGCAGCCCCCAcCGCGgcCGgCGCGcccccctCGccacccctgCGCGccagCGcCGgccCGccccTC
CCGCTGGCCTCTGGCGGCTTAACCCT**CG**CCTGCC**CG**ACCC**CGCG**GGGCCCCTGGAGC**CG**
CGGTGGGTGGGTg**CG**gg**CG**c**CG**ggcctcccctcc**CG**gc**CGCG**g**CGCGCG**g**CG**c**CG**ggAGCTGAC
CGTGGTGCTGAG**CGCG**GCT**CGCG**CTC**CG**A**CGCG**GTGCC**CG**AGCCTGT**CGCG**GC**CG**C**G**C
CCTGCTGCACTGCGGGCCCCCAGCGGTAAGTCGCCAAGGCCCCGAGAGGCTGCGTTGGT
CCTGCCCCGCGGATGTGGACCCCCGGGGAGGGCAAGGATTGGGGAATTTGGTGCATTCT
CTGGCGCGATGGACGGGCTGAGGGCAGGAAGCAGGGTCGACACCCTCCACCCAGCGGTT
CCCGGCGCAGTCAGGGGCCTGGGAGCGGTGACCCTTTTAGGTCTGGGCTAGGGAAGCGG
AGAAACCCTCCGCCGGCGGATGCAGGAGCTGAGGGAGAGCCAATATCGCTGTGAGGCCC
AT

9- Illumina probe ID = cg25124276

Dark gray = Regions not covered and not tested
Light gray = Regions of interest that were tested and did not pass QC
Light blue = Regions of interest covered by amplicons that passed QC
CG = Targeted CpG site
CG = Adjacent CpG sites

hg19_dna range=chr10:25463808-25464208 5'pad=200 3'pad=200 strand=-

TAAGAATGTTTCGAGATTGGCCCCGGGTGGCTAAGCTACCGTGCCCGCTGCGCGA
GGGGCCGGCTGGGGATTACGCACCTCGCAGCCTGGAGCCGAGCGGGTTACATGGCCTCG
CGTCGCAGAAATCAAGTCACCTGTGGCAGCGCTGGCCGCTCCCAGCGGCTGGAGTCAGC
C**CG**AGTC**CG**TCTCT**CG**GCC**CG**GCTGC**CG**CCAC**CG**C**CG**CTCTCATTGAGG**CGCG**TTCAGAA
GCTGCTGCTGCTGCTGCCGCCGCGGCCGCCGAAGACGCTGCTCCATAGTCTCACCCGCC
GCAGGTCGTTCCCGCCTCCCGGCACTGGCTCCTCCTCGACCGCTTCGCCCGGAGGCGCG
GGGCGCGCGCCCAGACCCGCACCGCGAGCGCAGCCGCGACTCCCGAGGAG

10- Illumina probe ID = cg18384097

Dark gray = Regions not covered and not tested
Light gray = Regions of interest that were tested and did not pass QC
Light blue = Regions of interest covered by amplicons that passed QC
CG = Targeted CpG site
CG = Adjacent CpG sites

hg19_dna range=chr1:202129366-202129766 5'pad=200 3'pad=200 strand=-

AGCTGGCTTCCTGGAGCCTTCTCAGCCCTCAAagacagacCGacagacagacagacagCTG
GCAAGAGGCAGCCTGGGGGCCACAGCTGCTTCAGTAAGTATCTGAAGGGGGGACTGGGA
GTCCTGTGGCCC**CG**GGGGGTG**CG**AACTC**CG**GGGATATAAGAGGGCATCTCTAGGAGGGA
GTG**CG**GGAGGG**CG**AGTGGGG**CG**CCACAGTGCCTGGCTGGGGTATGGGTGCTCACAGACC
TGATGTCCCCAAGACGGGGGTGAGCAGGGAAGCCACAGGGAGCTACAAGGAGAGCAGAG
GCTGAAGGGACCTTTTCTGCTACCAGAGACCCTCGCTCTACCACTCACACCTGTGCCAGGC
CCATTCTGTCCCCTCACCTCCGTCCCTGCTGCCTTGGTGATTC

11- Illumina probe ID = cg00481951

Dark gray = Regions not covered and not tested
Light gray = Regions of interest that were tested and did not pass QC
Light blue = Regions of interest covered by amplicons that passed QC
CG = Targeted CpG site
CG = Adjacent CpG sites

hg19_dna range=chr3:187387450-187387850 5'pad=200 3'pad=200 strand=-

CCTTCTAAAGAGTTTTGGTGCTTTTCTGGGTCCCTCAGCTCCCGAAGCTCTTGAGA
AAACTATCAAAGGCTAGAATCCCCTTCTAACTCTTTTTTTCCCCCATGATAAGCGCAGTCGG
TCACAGTTCAGGTGAGTTCTTACTTGGCATTCAAGAAAATTACAAAATCTGGGTAGTTGTCT
GGGCACGAAGCGACAATGGCGTCTATCCCTGGTGCTGACCCTGGGAAGCGCTGACCCAG
GTGCTGAAACGCAGACCTCTGAAGCTGCTACCTCTTAGCGTACCTCACTTCCAAACGTCGG
GACTAGGGCAAAGGGGCAATCTAAAGACCGAACGCCGTATGTTTGAGATTGTGAGAAGTCT
CGTTCCCCTACAGTTTACTTGGTAAAAATGGTAAAACAAT

12- Illumina probe ID = cg19671120

Dark gray = Regions not covered and not tested
Light gray = Regions of interest that were tested and did not pass QC
Light blue = Regions of interest covered by amplicons that passed QC
CG = Targeted CpG site
CG = Adjacent CpG sites

hg19_dna range=chr2:98962774-98963174 5'pad=200 3'pad=200 strand=+

TGTGAACTAATGACTTCTGCCTCTTTCGGACCGCCTTTGAGGCTCCAGAGCCTCAC
CTTACTTTCCCACGGAGAGGGAGGCCACAGGCTCCTTCAGCAGTCGCCGAGCAGAGTCCT
GGGCCGGGAGCGCGGGGGAGGGAGCGAGCGGAACTGCGCCTAGGAGGCCGAGGGAGG
AGGCGCTCCGCAGACCCTGGCGCGCCGGAGAAGCTCAAACTTTGGCAGGGTAAGGAT
TTTTAGGGGCTCTTGAGCTGGAATTTTTTGGGGGGCGCCGGGAGGTGTGCTGGGGCCGCA
GACCCCATACAGGAGGTAAGTTAGAGAACCACACGCAGGGGAGGGATGCTGCTGCTTCCA
GGGGCGGGCGCGGCGCTGTCCGCAGCCCCCGGTGCTGAAACGGGCCGCG

13- Illumina probe ID = cg14361627

Dark gray = Regions not covered and not tested
Light gray = Regions of interest that were tested and did not pass QC
Light blue = Regions of interest covered by amplicons that passed QC
CG = Targeted CpG site
CG = Adjacent CpG sites

hg19_dna range=chr7:130418916-130419316 5'pad=200 3'pad=200 strand=-

GCTACAGACTTGGGGATATTTGCAATTTAGTCTGAGAGCGGCGAACGGGGAGTGA
AATGGAATTCGAGACCACTTCGCTAACAATCGCAATTATGAACCGAAAGACATGTCAGGTAT
TAGCAATTTTTTTCCTTAAAAAAAAAAAAAACTTTCTGGGACTCCGCGGGACACCCAGCTGG
CGACGGACCAGCGGGCGGCGGGCTGCGGGGAGGGGGGGCGAGGCTGCTGCAACCCAGA
AGTTCCGACTGGGGAGTTTCGCTCTGTTACCATTACCTGGCTCGCCGGCAGAAGAAAGAACG
CGGGAGACAAGATAATTTCTGAGGCTGTTAAACATGACTTAGCCGGGGGGCCGCGCGTTC
CGAGGGGGTGTCCTGCGGGCCGGGGCGGGTCTCTGCCGCCCCCGCGGGCTCCGGTGCG
TCAGGGGCGCGTCAGGCGGGGCGGGCTCCGCGCGggCGgCGgCGgcagCGgCGgctgCGgC
GgCGgCGgCGgcagcaggCGgcaggCGgCGAGCACCCGGCCTCCTGCTTCTCGCTCGCCGGCG
GCCGGGCGGTCCCAGCATGTCGGCCGCCGTGGCGTGCCTGGACTACTTCGCCGCCGAGT
G

14- Illumina probe ID = cg08928145

Dark gray = Regions not covered and not tested
Light gray = Regions of interest that were tested and did not pass QC
Light blue = Regions of interest covered by amplicons that passed QC
CG = Targeted CpG site
CG = Adjacent CpG sites

hg19_dna range=chr19:19625164-19625564 5'pad=200 3'pad=200 strand=+

CTCGGGGCCTGGTCGCAGGGAATCGCGGATGCGCATGCGCCTCCAGCTCCCGGG
ATCGCGGGGAACGTGGATTCCGAACAAGGGCAACTGCGGACTCCCCCGTGGGAAGAAAG
GGAGGGAAGCGGAAGGGAAAAAGCGCATGTGCAGCAGCACGCGGCAGCTTCGCATATTC
CCTCGAAGCGCGCCTCTTGCGCGTGCGCCGCCCTGGCCCAGTGCCCTTGGCTGCAGGAA
TGGCTGGAACCACCCGGCTTCTAGCCGGAGTCCCCGGCGCGCAGCCAGCAGTTGCGCGC
TACCTGGCCCGCGGAGGGCCTGGCGGACGGGCTGAACTGCAGCAGCTCGGCGATCAGGG
CCTTGCAGCGCTCGGACAGCTCGAGGCCTTCGGGATAGAGCACGCCGCGTTT

15- Illumina probe ID = cg12757011

Dark gray = Regions not covered and not tested
Light gray = Regions of interest that were tested and did not pass QC
Light blue = Regions of interest covered by amplicons that passed QC
CG = Targeted CpG site
CG = Adjacent CpG sites

hg19_dna range=chr2:162280911-162281311 5'pad=200 3'pad=200 strand=-

AAACTAGCCATTCGTTACAATAAATTAACACTATGTACAATCATTTCACAGGCTTTGT
TCCACTAAAATTATTAACATCCCTAccatccatccatccatccatccaCTCGGGGATAAGAAAGGACGC
GCTGGGTGGGTCCAGGTTCTAGAGTCAACATCCTATGGTTAATGTGGAGGCCGAGACTTG
GCGGGGTCGGAATCGCTGCTGCCTGACAGGACTGCCCAGGTCTCCTAAAGGTGAAGAGTTT
CATTACAAGAAAGAGAAGTAGGAGGTAAGAGAAAGACGAGGGGGAGGGGAATTGTAGGAG
GATAACTCCACAGAAGAAGAGTAAGTAGGAAAACCAAAAGGTTTACAGCAAAGTTGAAGGT
CGGTGAGCTAATTGCAGAGACTGCTGTGCAAAA

16- Illumina probe ID = cg07547549

Dark gray = Regions not covered and not tested
Light gray = Regions of interest that were tested and did not pass QC
Light blue = Regions of interest covered by amplicons that passed QC
CG = Targeted CpG site
CG = Adjacent CpG sites

hg19_dna range=chr20:44658025-44658425 5'pad=200 3'pad=200 strand=-

CACGGGTGCCTGCTCTGCGCCAGGACGCCGGGGTCCCCTGTCGGGGAAAGGAGC
CCTCCTCCCGCCGTCGAGCTCCACATCAGCCCATTCTAGGTCTTCTATCCCCTTCCCACCG
CCTCCTCGGTTTAGCTAACCCAAGTCAGCCCGAAGCCGTGGGCAGCGATAATCCCCCGGC
CCGCAGCTCTGCCCCCGGGCGCGGCGCATTCCAGCACCTTGGACAGCGCCCGGAGCATT
CCAATGGAGCTGAGCTCCAGGCTGCAACTGGCGCCCACCGCCCGGGTCAGTCCCAGCCT
CCTCTCTGAAGGAGGTGGCCCCCGCTCCGCCTCTAACCCCAGAGCTGCCTCCTCTCCTCG
CTGCCCCCTCCCCCGCCGCCCCCGGACCACAGCTTACCCGGGTTGGCT

Table A1 Age-related CpG sites selected by elastic net regression. The training data contained 527 samples assayed on the Illumina EPIC BeadChip.

Probe ID	UCSC[1] Ref. Gene name	UCSC Ref. Gene group	Probe type	Chr.[2]	Pos.[3]
cg12642568	*CALML6*	5'UTR;1stExon	EPIC	chr1	1846648
cg10397932	*SKI*	Body	HM450K	chr1	2166155
cg21919596	*CHD5*	Body	HM450K	chr1	6201851
cg05808226	*unknown*	*unknown*	HM450K	chr1	11837939
cg01832549	*CAPZB*	Body	HM450K	chr1	19774989
cg12646386	*unknown*	*unknown*	HM450K	chr1	23902856
cg08876437	*RUNX3*	Body	HM450K	chr1	25228690
cg03987199	*OPRD1*	Body	HM450K	chr1	29189655
cg17957967	*AZIN*	TSS1500	EPIC	chr1	33546437
cg05336094	*POU3F1*	1stExon	HM450K	chr1	38512238
cg09547767	*BMP8A*	1stExon;5'UTR	HM450K	chr1	39957387
cg24248329	*NFYC*	1stExon;5'UTR	EPIC	chr1	41175132
cg07368443	*PLK3*	TSS1500	HM450K	chr1	45265337
cg04501188	*FOXD2*	1stExon	HM450K	chr1	47904171
cg00600454	*unknown*	*unknown*	HM450K	chr1	65468017
cg08179881	*DNAJC6*	5'UTR;Body	EPIC	chr1	65735532
cg03140521	*GNG12*	TSS1500	HM450K	chr1	68299388
cg13697378	*DIRAS3*	Body	HM450K	chr1	68512845
cg05654164	*C1orf52*	TSS1500;TSS1500	HM450K	chr1	85725892
cg10631373	*RBMXL1;CCBL2*	5'UTR;5'UTR;5'UTR;5'UTR	HM450K	chr1	89457642
cg25402655	*GFI1*	5'UTR;TSS1500;1stExon	EPIC	chr1	92952297
cg14375944	*unknown*	*unknown*	HM450K	chr1	94316420
cg24150153	*ARHGAP29*	TSS200	HM450K	chr1	94703463
cg01527715	*ARHGAP29*	TSS1500	EPIC	chr1	94704159
cg03373796	*AGL*	TSS1500	EPIC	chr1	100315483
cg03502767	*NTNG1*	3'UTR	EPIC	chr1	108023486
cg12100751	*C1orf59*	1stExon;5'UTR;5'UTR	HM450K	chr1	109203672
cg13673164	*SYPL2*	1stExon	HM450K	chr1	110009509
cg18933331	*unknown*	*unknown*	HM450K	chr1	110186418

cg05132925	*unknown*	*unknown*	HM450K	chr1	110438827
cg01715299	*unknown*	*unknown*	EPIC	chr1	111535465
cg00796661	*unknown*	*unknown*	EPIC	chr1	116856120
cg20760394	*LINC01525*	Body	EPIC	chr1	117855183
cg00701051	*unknown*	*unknown*	HM450K	chr1	145385391
cg01427957	*TNRC4*	TSS1500	HM450K	chr1	151689979
cg06400319	*SPRR2B*	TSS1500	HM450K	chr1	153045274
cg08684904	*SPRR2B*	TSS1500	EPIC	chr1	153045541
cg16256492	*ZBTB7B*	3'UTR	HM450K	chr1	154989843
cg06208270	*MEX3A*	3'UTR	HM450K	chr1	156046344
cg16126393	*MEX3A*	Body	EPIC	chr1	156046778
cg18593717	*HDGF*	TSS200	HM450K	chr1	156722068
cg16761098	*POU2F1*	Body	EPIC	chr1	167302476
cg08745595	*F5*	TSS1500	EPIC	chr1	169556012
cg02053850	*SYT2*	5'UTR	HM450K	chr1	202612633
cg12586428	*BTG2*	TSS1500	HM450K	chr1	203274421
cg11980944	*unknown*	*unknown*	HM450K	chr1	205399731
cg16966520	*LGTN*	TSS1500	HM450K	chr1	206786174
cg10501210	*unknown*	*unknown*	HM450K	chr1	207997020
cg11206148	*unknown*	*unknown*	EPIC	chr1	213779248
cg09209787	*PROX1-AS1*	Body	EPIC	chr1	214151107
cg27530209	*H3F3A;LOC440926*	TSS200;TSS200	HM450K	chr1	226250384
cg16565196	*RYR2*	Body	EPIC	chr1	237395924
cg15123428	*KLF11*	Body	HM450K	chr2	10186139
cg16832267	*KCNS3*	5'UTR	HM450K	chr2	18060102
cg15712057	*unknown*	*unknown*	HM450K	chr2	19661795
cg26427115	*C2orf70*	Body	HM450K	chr2	26785891
cg09153930	*DPYSL5*	TSS1500	HM450K	chr2	27070303
cg06952880	*FAM98A*	TSS200	HM450K	chr2	33824362
cg00881168	*QPCT*	TSS1500	HM450K	chr2	37570604
cg23859635	*MTA3*	TSS1500	HM450K	chr2	42795262
cg20707764	*MTA3*	Body	HM450K	chr2	42796178
cg02538752	*ACYP2*	TSS1500	EPIC	chr2	54341965
cg11807280	*unknown*	*unknown*	HM450K	chr2	66654644
cg22112832	*unknown*	*unknown*	HM450K	chr2	71098542
cg26090072	*RTKN*	TSS1500	HM450K	chr2	74669387

cg25040225	*LBX2*	TSS200	EPIC	chr2	74726734
cg06639320	*FHL2*	TSS200	HM450K	chr2	106015739
cg22454769	*FHL2*	TSS200	HM450K	chr2	106015767
cg02872546	*unknown*	*unknown*	EPIC	chr2	109741578
cg16532938	*FIGN*	Body	EPIC	chr2	164584635
cg11027822	*ITGA6*	Body	HM450K	chr2	173298550
cg25343589	*FSIP2;LOC101927196*	TSS200;Body	EPIC	chr2	186603311
cg23506322	*FSIP2;LOC101927196*	TSS200;Body	EPIC	chr2	186603317
cg11928668	*unknown*	*unknown*	EPIC	chr2	206754192
cg14947101	*DYTN*	Body	EPIC	chr2	207578484
cg09410607	*4-Mar*	TSS1500	HM450K	chr2	217237040
cg03708443	*unknown*	*unknown*	EPIC	chr3	5347184
cg11084334	*LHFPL4*	Body	HM450K	chr3	9594264
cg07279842	*unknown*	*unknown*	EPIC	chr3	10653031
cg12899747	*unknown*	*unknown*	HM450K	chr3	25391527
cg11870261	*TRANK1*	TSS200	HM450K	chr3	36986679
cg27440986	*VILL*	Body	EPIC	chr3	38045545
cg06911020	*MYRIP*	TSS200	HM450K	chr3	39851123
cg06957788	*SNRK*	5'UTR;5'UTR	EPIC	chr3	43343256
cg26614073	*SCAP*	TSS1500	HM450K	chr3	47517819
cg00664406	*GRM2*	TSS1500	HM450K	chr3	51740875
cg04453050	*GRM2*	TSS200	HM450K	chr3	51740896
cg03607117	*SFMBT1*	TSS1500	HM450K	chr3	53080440
cg15051960	*WNT5A*	TSS200	HM450K	chr3	55521376
cg13722120	*ERC2*	5'UTR	EPIC	chr3	56487225
cg16135090	*unknown*	*unknown*	HM450K	chr3	62934644
cg00232107	*MAGI1*	Body	EPIC	chr3	65390065
cg07941411	*CD80*	5'UTR	EPIC	chr3	119276686
cg27345757	*MYLK*	5'UTR	HM450K	chr3	123602795
cg25707924	*unknown*	*unknown*	HM450K	chr3	127174943
cg12943155	*PODXL2*	TSS200	HM450K	chr3	127347978
cg24607783	*PLSCR2*	5'UTR	HM450K	chr3	146187037
cg16181396	*ZIC1*	TSS1500	HM450K	chr3	147126206
cg06306198	*ZIC1*	Body	HM450K	chr3	147128998
cg21196581	*GPR160*	5'UTR	HM450K	chr3	169780131
cg13473356	*PEX5L*	TSS200	HM450K	chr3	179754613

cg13923516	*unknown*	*unknown*	EPIC	chr3	182389621
cg04612016	*unknown*	*unknown*	EPIC	chr3	197386201
cg09414241	*IQCG*	Body	EPIC	chr3	197665448
cg06110081	*UVSSA*	Body	EPIC	chr4	1350335
cg21678540	*LINC01182*	Body	EPIC	chr4	13929921
cg06839255	*unknown*	*unknown*	EPIC	chr4	25704841
cg24570371	*LOC439933*	TSS200	EPIC	chr4	36258130
cg26777800	*UGDH*	TSS1500	EPIC	chr4	39530705
cg01245787	*DCK*	1stExon	HM450K	chr4	71859630
cg15540044	*RCHY1*	5'UTR;TSS200;1stExon;5'UTR	HM450K	chr4	76439595
cg25428494	*HPSE*	Body	HM450K	chr4	84255411
cg17071446	*unknown*	*unknown*	HM450K	chr4	85402497
cg16789776	*unknown*	*unknown*	EPIC	chr4	99700028
cg22118416	*MGST2*	Body	HM450K	chr4	140621943
cg06633413	*unknown*	*unknown*	EPIC	chr4	152913773
cg05791548	*unknown*	*unknown*	EPIC	chr4	180033854
cg26703534	*AHRR*	Body	HM450K	chr5	377358
cg06430753	*unknown*	*unknown*	HM450K	chr5	958808
cg21788281	*unknown*	*unknown*	EPIC	chr5	2038539
cg06393703	*unknown*	*unknown*	EPIC	chr5	3643101
cg24087669	*unknown*	*unknown*	HM450K	chr5	5419515
cg16076065	*LOC100505625*	Body	EPIC	chr5	6706190
cg11584519	*11-Mar*	Body	HM450K	chr5	16175867
cg00602326	*RNASEN*	Body	HM450K	chr5	31427700
cg13039251	*PDZD2*	Body	HM450K	chr5	32018601
cg03230469	*GDNF*	5'UTR	EPIC	chr5	37837704
cg14558406	*unknown*	*unknown*	EPIC	chr5	55556553
cg17621438	*RNF180*	TSS1500	HM450K	chr5	63461216
cg07850154	*RNF180*	TSS1500	HM450K	chr5	63461232
cg13793354	*MCCC2*	TSS1500	HM450K	chr5	70882903
cg11423680	*unknown*	*unknown*	HM450K	chr5	72712836
cg25231948	*CAMK4*	TSS1500	EPIC	chr5	110559299
cg04991447	*SEMA6A*	1stExon;5'UTR	EPIC	chr5	115910058
cg08790036	*unknown*	*unknown*	HM450K	chr5	118732897

cg11623339	*MGC29506*	1stExon	HM450K	chr5	138725482
cg19505546	*unknown*	*unknown*	HM450K	chr5	139017263
cg24648119	*PCDHA6*	Body;	HM450K	chr5	140242639
cg21548029	*PCDHB5*	1stExon	HM450K	chr5	140515675
cg15389519	*PCDHGB4*	1stExon;Body	HM450K	chr5	140769585
cg01224715	*PCDHGA4*	Body	HM450K	chr5	140811520
cg12145907	*PCDHGA4*	Body	HM450K	chr5	140864834
cg02760293	*PPP2R2B*	5'UTR	HM450K	chr5	146258785
cg04117508	*unknown*	*unknown*	EPIC	chr5	155297843
cg01485938	*UNC5A*	Body	HM450K	chr5	176304609
cg02895588	*EXOC2*	3'UTR	HM450K	chr6	485915
cg09781987	*CDYL*	Body	EPIC	chr6	4828434
cg17619993	*BMP6*	Body	HM450K	chr6	7728888
cg16867657	*ELOVL2*	TSS1500	HM450K	chr6	11044877
cg22736354	*NHLRC1*	1stExon	HM450K	chr6	18122719
cg06493994	*SCGN*	1stExon;5'UTR	HM450K	chr6	25652602
cg01078434	*MAS1L*	1stExon	HM450K	chr6	29455532
cg23061027	*PRRT1*	3'UTR	HM450K	chr6	32116207
cg14027333	*PRRT1*	3'UTR	HM450K	chr6	32116317
cg20245641	*unknown*	*unknown*	EPIC	chr6	32668561
cg18824596	*HLA-DOA*	3'UTR	HM450K	chr6	32972970
cg18468088	*unknown*	*unknown*	HM450K	chr6	35490818
cg26129310	*MDGA1*	Body	HM450K	chr6	37664451
cg08125215	*unknown*	*unknown*	EPIC	chr6	43373300
cg11947985	*TMEM63B*	Body	HM450K	chr6	44119668
cg13523038	*TNFRSF21*	Body	EPIC	chr6	47255889
cg02344735	*B3GAT2*	Body	HM450K	chr6	71664559
cg17852588	*FILIP1*	Body	HM450K	chr6	76059756
cg00230815	*unknown*	*unknown*	EPIC	chr6	85304889
cg01393985	*GABRR1*	TSS200	HM450K	chr6	89927651
cg25838080	*NR2E1*	TSS200	EPIC	chr6	108489026
cg02799448	*OLIG3*	TSS1500	HM450K	chr6	137817008
cg26413501	*unknown*	*unknown*	EPIC	chr6	140762703
cg18418460	*ULBP3*	Body	HM450K	chr6	150390014
cg07979390	*LOC102723831*	Body	EPIC	chr6	151529514
cg02383285	*unknown*	*unknown*	HM450K	chr6	169286348
cg26224354	*C7orf50*	Body	HM450K	chr7	1096374

cg00880477	*unknown*	*unknown*	EPIC	chr7	7963432
cg20238678	*HDAC9*	TSS1500	EPIC	chr7	18548578
cg07522171	*JAZF1-AS1*	TSS1500	EPIC	chr7	28218686
cg02120774	*IGFBP3*	TSS1500	HM450K	chr7	45961473
cg12038684	*VOPP1*	TSS1500	EPIC	chr7	55640726
cg21040230	*unknown*	*unknown*	EPIC	chr7	56243786
cg07286216	*LOC650226*	Body	HM450K	chr7	56515846
cg21005510	*unknown*	*unknown*	HM450K	chr7	63353570
cg01487661	*unknown*	*unknown*	HM450K	chr7	63643277
cg12464817	*CDK14*	TSS1500	EPIC	chr7	90225378
cg25235205	*unknown*	*unknown*	HM450K	chr7	98970792
cg18446045	*VGF*	TSS1500	EPIC	chr7	100809055
cg02298479	*CPED1*	Body	EPIC	chr7	120661848
cg14175438	*FAM3C*	TSS1500	HM450K	chr7	121036729
cg21184711	*CADPS2*	Body	HM450K	chr7	122488330
cg02383785	*unknown*	*unknown*	HM450K	chr7	127808848
cg08097417	*KLF14*	TSS1500	HM450K	chr7	130419133
cg07955995	*KLF14*	TSS1500	HM450K	chr7	130419159
cg03473532	*MKLN1*	Body	HM450K	chr7	131008743
cg05245329	*TMEM178B*	Body	EPIC	chr7	141126001
cg05076820	*LOC100124692*	Body	HM450K	chr7	141871176
cg09910601	*EZH2*	Body	EPIC	chr7	148517771
cg21714581	*ZNF862*	TSS1500	EPIC	chr7	149535134
cg09215510	*ABCB8*	TSS1500	EPIC	chr7	150724046
cg08144358	*unknown*	*unknown*	HM450K	chr7	156889781
cg24216326	*unknown*	*unknown*	EPIC	chr7	158799053
cg07872300	*DLGAP2*	Body	HM450K	chr8	1553395
cg24214068	*NEFM*	TSS1500	EPIC	chr8	24771265
cg21088983	*CLVS1*	TSS200	HM450K	chr8	62200463
cg26290632	*CALB1*	1stExon	HM450K	chr8	91094847
cg14758256	*unknown*	*unknown*	EPIC	chr8	94834635
cg04517323	*LAPTM4B*	Body	HM450K	chr8	98788873
cg08081156	*SNX31*	TSS200	HM450K	chr8	101661976
cg16835398	*unknown*	*unknown*	EPIC	chr8	116837434
cg00415665	*ZHX2*	5'UTR	HM450K	chr8	123875036
cg10302505	*MYC*	TSS1500	HM450K	chr8	128748092

cg07118556	*unknown*	*unknown*	EPIC	chr8	135816995
cg18632612	*unknown*	*unknown*	EPIC	chr8	140116347
cg02290284	*unknown*	*unknown*	EPIC	chr8	144275267
cg01951863	*PLGRKT*	Body	EPIC	chr9	5376323
cg08709434	*ACER2*	TSS1500	EPIC	chr9	19408334
cg02616710	*ACER2*	TSS1500	HM450K	chr9	19408584
cg03462868	*UBE2R2*	TSS1500	EPIC	chr9	33816424
cg13566023	*unknown*	*unknown*	EPIC	chr9	93732908
cg17759214	*unknown*	*unknown*	EPIC	chr9	94137295
cg00590602	*SPTLC1*	TSS1500	EPIC	chr9	94878253
cg25514301	*GRIN3A*	TSS1500	EPIC	chr9	104501812
cg05330471	*unknown*	*unknown*	EPIC	chr9	121771590
cg04517263	*TRAF1*	TSS200	HM450K	chr9	123689193
cg13871695	*unknown*	*unknown*	HM450K	chr9	126101872
cg24765394	*FAM125B;FAM125B*	TSS1500;TSS1500	HM450K	chr9	129088647
cg20313295	*AK1*	TSS1500	EPIC	chr9	130640786
cg25073708	*C9orf106*	5'UTR	HM450K	chr9	132083538
cg08637691	*unknown*	*unknown*	EPIC	chr9	134989631
cg14295611	*unknown*	*unknown*	HM450K	chr9	136876366
cg14004197	*EXD3*	Body	HM450K	chr9	140216230
cg25438730	*CACNA1B;CACNA1B*	Body;Body	EPIC	chr9	141001373
cg06557316	*unknown*	*unknown*	EPIC	chr10	2434356
cg10542514	*unknown*	*unknown*	EPIC	chr10	3879115
cg14093395	*LINC00702*	TSS1500	EPIC	chr10	4286917
cg04968761	*unknown*	*unknown*	EPIC	chr10	18971200
cg17782713	*BMI1*	5'UTR	HM450K	chr10	22613360
cg14532839	*SPAG6*	ExonBnd	EPIC	chr10	22680650
cg13206721	*GPR158;GPR158-AS1*	TSS1500;Body	EPIC	chr10	25463350
cg13612317	*KIF5B*	TSS1500	HM450K	chr10	32345864
cg10381520	*unknown*	*unknown*	HM450K	chr10	33396942
cg06559368	*RET;RET*	Body;Body	HM450K	chr10	43573319
cg00418663	*C10orf105;CDH23*	5'UTR;Body	HM450K	chr10	73491922
cg10558013	*ZNF503*	Body	EPIC	chr10	77043088

cg15767361	*unknown*	*unknown*	EPIC	chr10	94837967
cg15298486	*BLOC1S2;BLOC1S2*	TSS1500;TSS1500	HM450K	chr10	102046690
cg04984663	*unknown*	*unknown*	HM450K	chr10	102632048
cg25546535	*GFRA1;GFRA1;GFRA1*	Body;Body;Body	EPIC	chr10	117898052
cg11705975	*PRLHR*	Body	HM450K	chr10	120354248
cg19979225	*unknown*	*unknown*	EPIC	chr10	132580986
cg22424845	*unknown*	*unknown*	EPIC	chr11	1404022
cg10043090	*HCCA2*	Body	HM450K	chr11	1536810
cg25601886	*INS;INS-IGF2;INS-IGF2*	TSS1500;TSS1500;TSS1500	HM450K	chr11	2183420
cg03811319	*unknown*	*unknown*	HM450K	chr11	2884118
cg02462487	*SLC22A18AS;SLC22A18*	Body;TSS1500	HM450K	chr11	2920350
cg09920974	*OSBPL5;OSBPL5;OSBPL5*	5'UTR;5'UTR;5'UTR	EPIC	chr11	3155992
cg23044178	*MICAL2*	5'UTR	HM450K	chr11	12136405
cg02452500	*unknown*	*unknown*	HM450K	chr11	13161927
cg25160605	*NELL1*	Body	HM450K	chr11	21087846
cg06731443	*LGR4*	TSS1500	HM450K	chr11	27494710
cg16343483	*unknown*	*unknown*	EPIC	chr11	34023105
cg12068553	*DGKZ*	TSS1500	EPIC	chr11	46367725
cg12189835	*SYT7*	Body	HM450K	chr11	61335071
cg13205113	*MACROD1*	Body	EPIC	chr11	63766918
cg20495962	*CATSPER1*	Body	HM450K	chr11	65789003
cg00630018	*MAP6;MAP6*	TSS1500;TSS1500	EPIC	chr11	75379680
cg23461714	*TTC12*	TSS1500	HM450K	chr11	113184990
cg16208682	*unknown*	*unknown*	EPIC	chr11	114127631
cg06038490	*unknown*	*unknown*	EPIC	chr11	123324673
cg05526578	*SIAE;SIAE*	Body;Body	EPIC	chr11	124514971
cg02046143	*IGSF9B*	Body	HM450K	chr11	133797911
cg02925805	*unknown*	*unknown*	HM450K	chr12	298484
cg06167456	*CACNA2D4*	Body	EPIC	chr12	1906837
cg25049091	*CACNA2D4*	TSS1500	EPIC	chr12	2028156

cg21666867	*DCP1B*	TSS1500	EPIC	chr12	2114282
cg19056004	*LRRC23;ENO2;LRRC23;LRRC23*	3'UTR;TSS1500;3'UTR;3'UTR	HM450K	chr12	7023262
cg14834260	*ENO2;LRRC23;LRRC23;LRRC23*	TSS1500;3'UTR;3'UTR;3'UTR	EPIC	chr12	7023269
cg18071806	*AEBP2;AEBP2;AEBP2*	TSS1500;TSS1500;TSS1500	EPIC	chr12	19592058
cg22083892	*KCNJ8*	TSS1500	EPIC	chr12	21928661
cg19722847	*IPO8*	TSS1500	HM450K	chr12	30849114
cg24221490	*DENND5B*	TSS200	EPIC	chr12	31743443
cg10695848	*HDAC7;HDAC7*	Body;Body	HM450K	chr12	48206783
cg19354681	*unknown*	*unknown*	EPIC	chr12	48223130
cg04614625	*unknown*	*unknown*	HM450K	chr12	52262736
cg17436656	*RARG*	TSS1500	HM450K	chr12	53627106
cg10409297	*unknown*	*unknown*	EPIC	chr12	57621737
cg11649376	*ACSS3*	Body	HM450K	chr12	81473234
cg07975200	*ANKS1B*	Body	EPIC	chr12	99525034
cg19598685	*ACACB*	Body	HM450K	chr12	109592525
cg10778288	*unknown*	*unknown*	HM450K	chr12	113917994
cg26682900	*HIP1R*	Body	HM450K	chr12	123344689
cg24891133	*C13orf33;C13orf33*	1stExon;5'UTR	HM450K	chr13	31480335
cg15782451	*unknown*	*unknown*	EPIC	chr13	34820887
cg18138898	*STK24*	Body	EPIC	chr13	99141075
cg01185345	*STK24*	Body	HM450K	chr13	99218287
cg13921483	*unknown*	*unknown*	HM450K	chr13	108657826
cg00593462	*unknown*	*unknown*	HM450K	chr13	110768493
cg10400227	*COL4A1;COL4A1;COL4A2*	TSS1500;TSS1500;Body	EPIC	chr13	110960772
cg20235099	*CARKD*	Body	HM450K	chr13	111288438
cg05183386	*TEX29;TEX29*	Body;Body	EPIC	chr13	111995758
cg22784964	*ANG;RNASE4*	TSS1500;TSS1500	HM450K	chr14	21151036
cg13138043	*unknown*	*unknown*	HM450K	chr14	54202939
cg03032497	*unknown*	*unknown*	HM450K	chr14	61108227

cg01894498	*unknown*	*unknown*	HM450K	chr14	61655848
cg17740900	*unknown*	*unknown*	EPIC	chr14	64266659
cg22820364	*unknown*	*unknown*	HM450K	chr14	81901540
cg14334310	*unknown*	*unknown*	HM450K	chr14	103558502
cg02210934	*unknown*	*unknown*	HM450K	chr14	105511982
cg13836627	*TJP1;TJP1*	Body;Body	HM450K	chr15	30113723
cg12400336	*TJP1;TJP1*	TSS200;TSS200	HM450K	chr15	30114871
cg13412433	*TJP1;TJP1;TJP1;TJP1*	TSS1500;TSS1500;TSS1500;Body	EPIC	chr15	30115098
cg26736154	*C15orf41;C15orf41*	Body;Body	HM450K	chr15	37020592
cg21220286	*DISP2*	TSS1500	HM450K	chr15	40650133
cg01770755	*unknown*	*unknown*	HM450K	chr15	41914122
cg01166932	*CGNL1;CGNL1*	5'UTR;5'UTR	EPIC	chr15	57698073
cg27167601	*RORA*	TSS1500	HM450K	chr15	61521923
cg27099280	*CELF6;CELF6*	1stExon;1stExon	EPIC	chr15	72612204
cg20809087	*BRUNOL6;BRUNOL6*	5'UTR;1stExon	HM450K	chr15	72612221
cg25049597	*CPLX3;CPLX3*	1stExon;5'UTR	HM450K	chr15	75119018
cg18110140	*unknown*	*unknown*	EPIC	chr15	75350380
cg08329821	*TMC3-AS1*	Body	EPIC	chr15	81623305
cg05792169	*unknown*	*unknown*	HM450K	chr15	85874227
cg01351822	*UNC45A;UNC45A*	5'UTR;1stExon	HM450K	chr15	91473475
cg09150269	*RAB11FIP3*	1stExon	EPIC	chr16	476337
cg01960979	*IFT140*	Body	HM450K	chr16	1611342
cg08331960	*SLC9A3R2;SLC9A3R2*	TSS1500;TSS1500	HM450K	chr16	2076597
cg04130886	*unknown*	*unknown*	HM450K	chr16	2723694
cg20106684	*CACNG3*	Body	EPIC	chr16	24269504
cg27151362	*DOC2A*	TSS1500	HM450K	chr16	30023515
cg13901319	*unknown*	*unknown*	HM450K	chr16	33319362
cg12641578	*ITFG1-AS1;ITFG1-*	TSS1500;TSS1500;1stExon;	EPIC	chr16	47177648

	AS1;NETO2;NETO2;NETO2;NETO2	1stExon;5'UTR;5'UTR			
cg00094898	*unknown*	*unknown*	HM450K	chr16	55365950
cg07280206	*BBS2*	TSS1500	HM450K	chr16	56554249
cg08787607	*ADGRG1*	5'UTR	HM450K	chr16	57684303
cg17650822	*NDRG4;NDRG4;NDRG4*	TSS1500;TSS200;5'UTR	EPIC	chr16	58497795
cg06320982	*SLC38A7*	TSS200	HM450K	chr16	58718767
cg06285333	*AGRP;ATP6V0D1;AGRP*	Body;TSS1500;Body	HM450K	chr16	67516546
cg05915866	*ZFHX3*	5'UTR	HM450K	chr16	73090838
cg12959488	*unknown*	*unknown*	EPIC	chr16	73094343
cg03743982	*LDHD;LDHD*	TSS200;TSS200	HM450K	chr16	75150833
cg03576805	*unknown*	*unknown*	HM450K	chr16	82259458
cg22979810	*unknown*	*unknown*	EPIC	chr16	85148525
cg27430293	*unknown*	*unknown*	HM450K	chr16	89069935
cg03172657	*ACSF3;ACSF3;ACSF3*	Body;5'UTR;5'UTR	HM450K	chr16	89163625
cg02228185	*ASPA;ASPA*	1stExon;Body	HM450K	chr17	3379567
cg09451903	*TRPV1;TRPV1*	TSS1500;5'UTR	HM450K	chr17	3501338
cg20717792	*unknown*	*unknown*	EPIC	chr17	17111358
cg03259243	*unknown*	*unknown*	HM450K	chr17	21356007
cg23536675	*WSB1;WSB1*	TSS1500;TSS1500	HM450K	chr17	25620638
cg13029847	*SEZ6;SEZ6*	TSS200;TSS200	HM450K	chr17	27333273
cg05000339	*RAB11FIP4*	Body	HM450K	chr17	29817128
cg26314066	*CNTD1;COA3*	TSS1500;Body	EPIC	chr17	40950091
cg16931499	*DBF4B;DBF4B*	Body;Body	HM450K	chr17	42786676
cg22704696	*PHOSPHO1;PHOSPHO1*	Body;Body	HM450K	chr17	47302476
cg11071401	*CACNA1G*	TSS1500	HM450K	chr17	48637194
cg16987606	*GPRC5C*	TSS1500	HM450K	chr17	72426469

cg07030794	*CD300LD;C17orf77*	TSS1500;3'UTR	HM450K	chr17	72589110
cg17697835	*SEPT9;SEPT9*	TSS200;Body	HM450K	chr17	75283832
cg19715771	*CBX4*	Body	HM450K	chr17	77810912
cg22353329	*CBX4*	TSS1500	HM450K	chr17	77814357
cg06163904	*CHMP6*	TSS1500	HM450K	chr17	78964779
cg24870966	*C17orf55*	5'UTR	HM450K	chr17	79282893
cg01419914	*BAHCC1*	Body	HM450K	chr17	79374691
cg12194745	*BAHCC1*	Body	HM450K	chr17	79423649
cg07497327	*EPB41L3*	TSS200	EPIC	chr18	5629057
cg02314019	*PTPRM*	Body	EPIC	chr18	7575762
cg15820059	*unknown*	*unknown*	EPIC	chr18	36512980
cg24217948	*SETBP1;SETBP1*	5'UTR;5'UTR	HM450K	chr18	42261980
cg17243289	*SMAD2;SMAD2;SMAD2*	TSS1500;TSS1500;TSS1500	HM450K	chr18	45458021
cg12929062	*unknown*	*unknown*	EPIC	chr18	65965879
cg23540632	*DOK6*	Body	EPIC	chr18	67253448
cg27159585	*CTDP1;CTDP1;CTDP1;CTDP1*	5'UTR;5'UTR;1stExon;1stExon	EPIC	chr18	77439862
cg07384708	*THEG;THEG*	Body;Body	HM450K	chr19	372718
cg13393785	*LASS4*	Body	HM450K	chr19	8317932
cg00339281	*ADAMTS10*	TSS1500	EPIC	chr19	8676470
cg15013019	*LYL1;LYL1*	5'UTR;1stExon	HM450K	chr19	13213451
cg26842596	*CCDC105*	TSS1500	HM450K	chr19	15121297
cg13640414	*AKAP8L*	TSS1500	HM450K	chr19	15530870
cg20119148	*PDE4C*	5'UTR	HM450K	chr19	18344195
ch.19.21460585R	*unknown*	*unknown*	HM450K	chr19	21668745
cg15761414	*MAG;MAG*	Body;Body	HM450K	chr19	35801014
cg22891287	*HSPB6;C19orf55*	TSS1500;TSS200	HM450K	chr19	36248992
cg09083721	*unknown*	*unknown*	EPIC	chr19	36420908
cg09474229	*RINL*	Body	HM450K	chr19	39360330
cg09431525	*DLL3;DLL3*	Body;Body	HM450K	chr19	39993313
cg09255748	*SELV*	TSS1500	HM450K	chr19	40004995

cg24726064	*HNRNPUL1;HNRNPUL1;HNRNPUL1*	5'UTR;TSS1500;TSS200	EPIC	chr19	41770190
cg00841035	*TMEM91;TMEM91;TMEM91;TMEM91;TMEM91*	TSS1500;TSS1500;TSS1500;TSS1500;5'UTR	EPIC	chr19	41880968
cg15904523	*ZNF233*	TSS200	HM450K	chr19	44763979
cg20579054	*ZNF233;ZNF233;ZNF233*	TSS200;1stExon;5'UTR	EPIC	chr19	44764048
cg19351603	*unknown*	*unknown*	HM450K	chr19	45943663
cg21632975	*NOVA2*	Body	HM450K	chr19	46456210
cg02807849	*GRIN2D*	Body	HM450K	chr19	48908102
cg04731544	*LMTK3*	Body	HM450K	chr19	49004834
cg27064907	*unknown*	*unknown*	EPIC	chr19	50594033
cg24481841	*NCRNA00085*	Body	HM450K	chr19	52203721
cg03071580	*KIR3DX1;KIR3DX1;KIR3DX1;KIR3DX1*	Body;Body;Body;Body	EPIC	chr19	55045144
cg00149708	*SBK2*	TSS1500	HM450K	chr19	56047884
cg04126816	*CCDC106*	5'UTR	HM450K	chr19	56159710
cg13341864	*TGM6*	Body	HM450K	chr20	2384435
cg03738669	*SLC4A11*	TSS200	HM450K	chr20	3218476
cg02949067	*PCSK2*	3'UTR	HM450K	chr20	17463831
cg13727122	*NINL*	TSS200	HM450K	chr20	25566180
cg17261529	*unknown*	*unknown*	EPIC	chr20	32778060
cg09409865	*PIGU*	Body	EPIC	chr20	33169887
cg18660345	*MYL9;MYL9*	TSS1500;TSS1500	HM450K	chr20	35169539
cg23640862	*unknown*	*unknown*	EPIC	chr20	39389353
cg08058894	*TOX2;TOX2;TOX2*	TSS1500;Body;TSS1500	HM450K	chr20	42544178
cg09753064	*JPH2*	Body	HM450K	chr20	42788303
cg07547549	*SLC12A5;SLC12A5*	Body;Body	HM450K	chr20	44658225
cg06881858	*unknown*	*unknown*	EPIC	chr20	47139682

cg00387658	*CASS4;CASS4;CASS4;CASS4*	TSS1500;TSS1500;TSS1500;TSS1500	HM450K	chr20	54986793
cg06626338	*unknown*	*unknown*	EPIC	chr20	56678844
cg22961457	*LIME1;SLC2A4RG*	3'UTR;TSS1500	HM450K	chr20	62370310
cg20943769	*unknown*	*unknown*	EPIC	chr21	25098318
cg01573121	*DNAJC28;DNAJC28*	5'UTR;5'UTR	HM450K	chr21	34863117
cg18635497	*unknown*	*unknown*	EPIC	chr21	35349128
cg18074297	*CLIC6*	TSS200	HM450K	chr21	36041612
cg25992321	*TMPRSS3;TMPRSS3;TMPRSS3;TMPRSS3;TMPRSS3*	TSS1500;Body;Body;Body;Body	EPIC	chr21	43809689
cg16501323	*AIRE*	TSS200	HM450K	chr21	45705618
cg03208198	*COL18A1;COL18A1;COL18A1*	Body;Body;Body	HM450K	chr21	46898048
cg27060381	*TBC1D10A*	TSS1500	HM450K	chr22	30723363
cg21669271	*unknown*	*unknown*	EPIC	chr22	35850810
cg21737444	*LGALS1*	TSS200	HM450K	chr22	38071591
cg19853760	*LGALS1;LGALS1*	1stExon;5'UTR	HM450K	chr22	38071677
cg19855470	*CACNA1I;CACNA1I*	Body	HM450K	chr22	40060836
cg00058879	*CACNA1I;CACNA1I*	Body	HM450K	chr22	40082173
cg09875523	*unknown*	*unknown*	HM450K	chr22	46280123

[1] Based on UCSC Genome Browser database

[2] Chromosome

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

Table A2 267 AR CpG sites selected by elastic net regression across all tissues in the training data set. The chromosome coordinate is based on human genome assembly hg18.

Probe's ID	Chr.	Coordinate	Gene
cg00236832	17	35719015	RARA
cg00398048	4	178601336	AGA
cg00417297	3	40493746	ZNF619
cg00431549	12	14930292	MGP
cg00540544	1	199742920	CSRP1
cg00577167	9	103237496	ALDOB
cg00718748	17	16497596	ZNF624
cg00864867	12	78609399	PAWR
cg00945507	7	54795171	SEC61G
cg01027739	9	130882559	DOLPP1
cg01027805	14	20636703	ZNF219
cg01137065	17	78070299	FOXK2
cg01234063	11	125731217	ST3GAL4
cg01353448	7	31693437	C7orf16
cg01459453	1	167865836	SELP
cg01485645	17	34115725	MLLT6
cg01511567	11	56860207	SSRP1
cg01626227	7	99355225	TRIM4
cg01632825	22	17658672	CLTCL1
cg01644850	19	62885043	ZNF551
cg01988129	8	67507490	ADHFE1
cg02007844	12	21819257	KCNJ8
cg02016419	17	15185962	TEKT3
cg02049180	1	155094826	INSRR
cg02317907	1	68288376	DIRAS3
cg02331561	16	2331082	ABCA3
cg02388150	8	41284856	SFRP1
cg02477931	11	131037931	C11orf39
cg02780295	5	140835647	PCDHGC3
cg02810134	20	43889329	TNNC2
cg02827112	4	95348426	SMARCAD1

cg02828104	20	48204207	Kua
cg03019000	3	51679391	TEX264
cg03103192	4	52612028	SPATA18
cg03294619	5	172594409	NKX2-5
cg03305230	17	17526063	RAI1
cg03330058	3	128875093	ABTB1
cg03464689	4	103641461	NFKB1
cg03565323	17	16413591	ZNF287
cg03640148	3	49019956	WDR6
cg03641225	1	68285127	DIRAS3
cg03843852	15	53398391	PIGB
cg03975694	19	42734312	ZNF540
cg03991512	16	73707957	LDHD
cg04084157	7	100595769	VGF
cg04119538	20	39090865	TOP1
cg04240200	20	3399292	ATRN
cg04452713	6	56815646	DST
cg04464446	11	68209421	GAL
cg04474832	3	51983527	ABHD14A
cg04528819	7	130068855	KLF14
cg04833845	19	48977916	KCNN4
cg04836038	13	98537383	DOCK9
cg04999691	7	149657983	C7orf29
cg05294243	19	56260918	KLK13
cg05442902	22	19699010	P2RXL1
cg05467458	19	38052872	SLC7A9
cg05600717	13	51276745	FLJ13639
cg05624932	8	76059865	CRISPLD1
cg05675373	1	110555780	KCNC4
cg05965402	4	24844514	PI4K2B
cg06220958	17	10393576	MYH2
cg06222851	10	50640364	OGDHL
cg06493994	6	25760581	SCGN
cg06533629	7	130069910	KLF14
cg06580318	2	169455365	SPBC25

cg06597861	8	144171947	LY6E
cg06615861	1	10193778	KIF1B
cg06780358	11	47959641	PTPRJ
cg06810647	16	1605095	CRAMP1L
cg06836772	1	56882991	PRKAA2
cg06952310	19	19188990	CSPG3
cg07034561	21	42789228	TSGA2
cg07071881	22	31201139	FBXO7
cg07158339	9	70840057	FXN
cg07360076	4	1764437	FGFR3
cg07388493	1	39264046	NDUFS5
cg07441272	3	188340169	RPL39L
cg07590705	7	132417050	CHCHD3
cg08089301	17	44010560	HOXB4
cg08209724	17	27701251	ZNF207
cg08331960	16	2016598	SLC9A3R2
cg08413469	1	68735528	DEPDC1
cg08521225	18	27425975	TTR
cg08537652	1	2976222	PRDM16
cg08725962	5	175725099	ARL10
cg08785215	3	57969019	FLNB
cg08965235	11	65081734	LTBP3
cg09084200	11	133601073	hCAP-D3
cg09150232	7	50817133	GRB10
cg09310112	19	4920989	JMJD2B
cg09462576	1	226364496	MRPL55
cg09497789	1	118529735	SPAG17
cg09646392	13	107719053	TNFSF13B
cg09809672	1	234624305	EDARADD
cg10046620	6	27883021	HIST1H2AI
cg10294836	19	45016586	DYRK1B
cg10377274	11	125122098	PATE
cg10486998	18	73090775	GALR1
cg10523019	2	227408702	RHBDD1
cg10588377	10	124211789	HTRA1

cg11025793	19	13123015	STX10
cg11126134	13	30378304	FLJ14834
cg11170796	19	1601224	TCF3
cg11299964	9	127509604	MAPKAP1
cg11377136	22	45037624	PKDREJ
cg11673969	7	100262881	EPHB4
cg12118011	1	177529918	SOAT1
cg12351433	2	48836461	LHCGR
cg12373771	22	15981381	CECR6
cg12447832	2	3362264	TTC15
cg12478185	17	10542726	SCO1
cg12686016	7	27102002	HOXA1
cg12688670	16	29709103	KIF22
cg12774845	14	73556065	ENTPD5
cg12903171	7	50818058	GRB10
cg12946225	19	3524751	HMG20B
cg13164537	18	65775051	CD226
cg13382694	17	44795202	ZNF652
cg13460409	21	37301440	DSCR6
cg13526007	14	41146460	LRFN5
cg13547237	11	65444453	Bles03
cg13552869	16	29817425	SEZ6L2
cg13672791	19	59107977	CACNG7
cg13682722	14	89868321	C14orf102
cg13697378	1	68285433	DIRAS3
cg13870494	9	71848178	MAMDC2
cg13899108	19	18205322	PDE4C
cg13975369	7	129867789	TSGA14
cg14121103	3	43706977	ABHD5
cg14155397	15	64465836	MAP2K1
cg14163776	3	196645869	CENTB2
cg14258236	6	29431309	OR5V1
cg14313310	1	182273439	GLT25D2
cg14407667	1	51757797	EPS15
cg14797887	15	54544604	MNS1

cg14892066	4	184076257	DCTD
cg14894144	18	19524552	LAMA3
cg14925024	1	226358328	C1orf35
cg15032239	15	20443395	CYFIP1
cg15188491	1	145110730	PRKAB2
cg15201877	1	71285561	PTGER3
cg15341340	19	12853237	DNASE2
cg15343119	18	73090773	GALR1
cg15377518	2	144994583	ZFHX1B
cg15379633	22	21817586	RAB36
cg15473868	16	55274122	MT1X
cg15648905	9	129252778	RPL12
cg15701111	12	50971601	KRTHB1
cg15804973	6	137156206	MAP3K5
cg15974053	19	54031601	DHRS10
cg16034652	14	92868062	KIAA1409
cg16168311	1	154828571	APOA1BP
cg16254764	3	13496178	HDAC11
cg16273597	6	14225459	CD83
cg16319578	14	64077265	HSPA2
cg16338035	3	145173959	C3orf58
cg16421589	5	167939339	PANK3
cg16547529	11	74818329	FLJ33790
cg16731240	19	57083062	ZNF577
cg16744741	4	82345049	PRKG2
cg16832407	9	72926359	TRPM3
cg16984944	3	101462115	TBC1D23
cg17096191	1	160305848	NOS1AP
cg17324128	10	44775506	RASSF4
cg17403875	14	54666109	LGALS3
cg17575811	11	2422985	KCNQ1
cg17589175	11	63841913	HSPC152
cg17655614	16	67328445	CDH1
cg17688525	18	6404978	L3MBTL4
cg17861230	19	18204901	PDE4C

cg17940013	1	9111275	GPR157
cg17945001	1	18306705	IGSF21
cg17966192	2	108360548	SULT1C2
cg18031008	1	148532935	MRPS21
cg18440048	22	22423826	ZNF70
cg18441959	22	20929332	VPREB1
cg18573383	12	73889668	KCNC2
cg18674980	8	86537833	CA3
cg18740800	11	27871510	HSPCAL3
cg18992688	1	204389864	AVPR1B
cg19023700	7	130662889	MKLN1
cg19029220	2	63921326	UGP2
cg19055231	3	36397401	STAC
cg19237753	20	1824215	PTPNS1
cg19464016	6	106640651	PRDM1
cg19523029	9	81376704	TLE4
cg19526626	19	11770408	ZNF491
cg19709625	3	144090671	PCOLCE2
cg19722847	12	30740381	IPO8
cg19761273	17	77825385	CSNK1D
cg19941758	11	82675116	MDS025
cg19945840	1	1157899	B3GALT6
cg20240860	11	44043999	PHACS
cg20300246	9	138236568	LHX3
cg20557567	6	33347672	RPS18
cg20630655	15	73705755	RNUT1
cg20692569	7	72486417	FZD9
cg20695562	1	77998224	USP33
cg20702327	20	472484	CSNK2A1
cg20716064	10	11692956	USP6NL
cg20761322	15	76210619	CIB2
cg20828084	15	78857906	KIAA1199
cg20925811	20	44071924	MMP9
cg21081971	13	37342765	TRPC4
cg21341271	7	39572208	C7orf36

cg21389884	11	119104635	PVRL1
cg21418052	13	30671806	B3GTL
cg21801378	15	70399179	BRUNOL6
cg21808053	1	68285651	DIRAS3
cg21948783	12	47658741	WNT1
cg22143352	14	74967594	JDP2
cg22171829	7	95063456	PDK4
cg22289837	8	86537530	CA3
cg22395019	2	31215196	GALNT14
cg22541143	4	84596316	HEL308
cg22637507	11	43858983	ALKBH3
cg22679120	7	2319928	SNX8
cg22723026	14	20222292	ANG
cg22736354	6	18230698	NHLRC1
cg22809047	2	100984693	RPL31
cg22920873	7	138675693	HSPC268
cg22947000	16	79829782	BCMO1
cg23081213	2	219404716	PRKAG3
cg23089840	21	44699314	LRRC3
cg23124451	22	37878077	CBX7
cg23325242	22	41374732	CYB5R3
cg23808301	17	4656972	PLD2
cg23837897	17	50401269	COX11
cg23873703	3	157321319	KCNAB1
cg24058132	14	87529619	GALC
cg24081819	8	27404857	EPHX2
cg24127874	2	238814598	HES6
cg24254120	13	33290869	RFC3
cg24384676	14	36711715	SLC25A21
cg24638647	17	39447773	TMEM101
cg24860534	1	225573491	CDC42BPA
cg24888049	15	89227671	FES
cg24958765	19	45975507	RAB4B
cg25101936	11	113434374	ZBTB16
cg25141720	9	88751646	GAS1

cg25148589	4	158361386	GRIA2
cg25420583	2	74539110	WBP1
cg25475443	10	125796378	GALNAC4S-6ST
cg25655096	12	6615553	GPR92
cg25736482	19	59369044	TMC4
cg25771195	16	56721315	GTL3
cg25809905	17	39823254	ITGA2B
cg25836301	14	100362059	MEG3
cg25894551	11	117205061	FXYD2
cg25915982	7	50816909	GRB10
cg26005082	19	4720660	C19orf30
cg26069745	7	27108725	HOXA2
cg26356176	5	53849137	SNAG1
cg26372517	1	35811746	TFAP2E
cg26374101	9	139620634	ARRDC1
cg26394940	22	44828125	FLJ10945
cg26581729	9	139059613	NPDC1
cg26614073	3	47492823	SCAP
cg26738080	3	52462773	TNNC1
cg26842024	19	16297122	KLF2
cg27015931	16	21919905	MGC50721
cg27016307	19	54350725	HRC
cg27169020	15	81745233	BNC1
cg27303880	1	64012776	ROR1
cg27486427	3	25444923	RARB
cg27544190	21	32707305	C21orf63

Appendix B: Participant information sheet and consent forms

B1. Participant information sheet

Name of department: Department of Pure and Applied Chemistry

Title of the study: Identifying age related DNA methylation markers in saliva.

Introduction

Chief Investigator:
Name: Dr Penny Haddrill
Status: Teaching Fellow
Tel.: 0141 548 4337
E-mail: penny.haddrill@strath.ac.uk

Co-Investigator:
Name: Hussain Alsaleh
Status: PhD Student
Tel.: 01415485992
E-mail: hussain-alsaleh@strath.ac.uk

What is the purpose of this investigation?

At present, there is ongoing research trying to find age related DNA methylation markers for different types of body fluids for reliable and accurate age estimation. Methylation is a chemical modification found in everybody's DNA that changes over the course of a person's life. Such markers would be invaluable for forensic practices, allowing investigators to estimate the age an individual – for example a perpetrator who left biological sample (saliva) at a crime scene. The aim of this study is to develop a method for generating intelligence information from biological evidence, especially saliva, by estimating the age of the individual who left their sample at a crime scene.

Does your child have to take part?

This study involves participants donating biological samples in the form of saliva. Participation in this research is entirely voluntary and you have the right to refuse to give consent for your child to participate without giving a reason; refusing to allow your child participate will not negatively affect you or your child in any way. You also have the right to withdraw your child's participation from this research at any time up to the completion of the project (estimated to be around the 10th January 2019) without detriment and without giving a reason, and ask for their data to be destroyed.

What will your child do in the project?

Your child will be required to provide the following type of biological samples:

- A saliva sample (less than 5 mL), collected by the participant themselves into a sterile tube.

DNA will be extracted from the sample and methylation profiling will be carried out on the DNA. This will involve measuring the level of DNA methylation present at different positions in the genome, and then relating the level of methylation to the chronological age of the sample donors using statistical analysis. Your child will therefore also be required to provide their age when you agree that they can donate a sample.

The sample processing will take place at both the Centre for Forensic Science, University of Strathclyde, Royal College Building, 204 George Street, Glasgow, G1 1XW, and the Zymo Research's services labs in U.S.A. No payments will be provided for taking part in this research.

Why has your child been invited to take part?

In order to study the ageing of saliva, it is essential that these samples are collected from volunteers; these samples cannot be simulated. All participants aged 12 and over in age are welcomed to participate, in order to provide the necessary biological samples.

What are the potential risks to you in taking part?

Handling a biological sample such as saliva may carry a small risk of infection, but participants will be asked only to handle their own biological samples.

What happens to the information in the project?

All participants' information will be kept confidential. In order to preserve anonymity, samples will be labelled with a code that does not contain any information allowing the participant to be identified, except for by the investigators, to allow a participant's sample and data to be identified, removed and destroyed should they subsequently wish to withdraw. All data outputs will be stored electronically on password-protected computers only accessible by the investigators. Electronic data may be retained indefinitely in this form, but no information will be put onto any databases. All biological samples (e.g. DNA samples) will be securely disposed of within 1 month of the conclusion of the study, which is estimated to be around the 10th January 2019. Once the project is completed, the codes linking identity with samples will be deleted and after this point participants will no longer be able to withdraw from the study.

The outcomes of this study will be written into a PhD thesis by co-investigator Hussain Alsaleh. In addition, it is envisaged that the outcomes of this study will be written into journal/conference publication(s). In neither form of publication will any information be included that could allow the participants to be identified.

The University of Strathclyde is registered with the Information Commissioner's Office who implements the Data Protection Act 1998. All personal data on participants will be processed in accordance with the provisions of the Data Protection Act 1998.

Thank you for reading this information – please ask any questions if you are unsure about what is written here.

What happens next?

If you are happy for your child to be involved in the project after explaining the project to them and are happy for them to provide a sample, please sign the consent form provided to confirm this. The researcher will also explain the project to your child and ask them to verbally consent to say they are willing to participate in the project. Please note that participants will not be informed of the specific results of the tests.

If you do not want to be involved in the project, we would like to thank you for your attention.

Researcher contact details:

Hussain Alsaleh, PhD Student
Centre for Forensic Science, Department of Pure and Applied Chemistry
University of Strathclyde
Royal College, 204 George Street, Glasgow, G1 1XW
Telephone: 0141 958 5992
E-mail: hussain-alsaleh@strath.ac.uk

Chief Investigator details:

Dr Penny Haddrill, Teaching Fellow
Centre for Forensic Science, Department of Pure and Applied Chemistry
University of Strathclyde
Royal College, 204 George Street, Glasgow, G1 1XW
Telephone: 0141 548 4377
E-mail: penny.haddrill@strath.ac.uk

This investigation was granted ethical approval by the Department of Pure and Applied Chemistry Ethics Committee.

If you have any questions/concerns, during or after the investigation, or wish to contact an independent person to whom any questions may be directed or further information may be sought from, please contact:

Secretary to the University Ethics Committee
Research & Knowledge Exchange Services
University of Strathclyde
Graham Hills Building
50 George Street
Glasgow
G1 1QE
Telephone: 0141 548 3707
Email: ethics@strath.ac.uk

B2. Consent Form

Name of department: Department of Pure and Applied Chemistry

Title of the study: Identifying age related DNA methylation markers in saliva.

- I confirm that I have read and understood the information sheet for the above project and the researcher has answered any queries to my satisfaction.
- I understand that my participation is voluntary and that I am free to withdraw from the project at any time, up to the point of completion of the study, without having to give a reason and without any consequences. If I exercise my right to withdraw and I don't want my data to be used, any data which have been collected from me will be destroyed. However, I understand that upon the completion of the study I will no longer be able to withdraw.
- I understand that any information recorded in the investigation will remain confidential and no information that identifies me will be made publicly available.
- I consent to being a participant in the project.
- I understand that I will be asked to donate a saliva sample, collected by myself into a tube.
- I consent to the DNA in my samples being analysed.
- I consent to the taking of biological samples from me, and understand that they will be the property of the University of Strathclyde.
- I understand that all of my biological samples will be securely destroyed within one month of the end of the project, which is estimated to be on the 10th of January 2019.

Please provide the information below:

Age :	
Gender : (optional)	

(PRINT NAME)	Date:
Signature of Participant:	

B3. Parent's Consent Form

Name of department: Department of Pure and Applied Chemistry

Title of the study: Identifying age related DNA methylation markers in saliva.

- I confirm that I have read and understood the information sheet for the above project and the researcher has answered any queries to my satisfaction.
- I understand that my child's participation is voluntary and that I am free to withdraw their participation from the project at any time, up to the point of completion of the study, without having to give a reason and without any consequences. If I or my child exercise my child's right to withdraw and don't want the data to be used, any data which have been collected from my child will be destroyed. However, I understand that upon the completion of the study it will no longer be possible to withdraw.
- I understand that any information recorded in the investigation will remain confidential and no information that identifies my child will be made publicly available.
- I consent to my child being a participant in the project.
- I understand that my child will be asked to donate a saliva sample, collected by the child themselves into a tube.
- I consent to the DNA in my child's samples being analysed.
- I consent to the taking of biological samples from my child, and understand that they will be the property of the University of Strathclyde.
- I understand that all of my child's biological samples will be securely destroyed within one month of the end of the project, which is estimated to be on the 10th of January 2019.

Please provide the information below:

Age :	
Gender : (optional)	

(PRINT NAME) (PARENT)	(PRINT NAME)(CHILD)
Signature of Parent:	Date:

Appendix C: R codes used for DNA methylation analysis

Appendix C1: R codes used for Chapter 3

This section provides the R codes used in Chapter 3, to identify the optimum method for identifying age related CpG sites and build a saliva-specific HM450k model.

```
library(GEOquery)

 GSE59509 <- getGEO("GSE59509",getGPL = FALSE,AnnotGPL = FALSE,GSEMatrix = TRU
E)

# Storing expression data, phenotype data, and feature data
GSE59509exprs <- as.data.frame(exprs(GSE59509[[1]]))
GSE59509pheno<-pData(phenoData(GSE59509[[1]]))
GSE59509feature <- read.table("/Users/husainalsaleh/AgePrediction_cache/GSE59509fe
ature.txt", header = TRUE)

sum(is.na(GSE59509exprs))
# Removing NA values
GSE59509exprs_NoNa=GSE59509exprs[complete.cases(GSE59509exprs),]
# Check
sum(is.na(GSE59509exprs_NoNa))
# The number of remaining CpG sites
nrow(GSE59509exprs)

colors <- terrain.colors(42, alpha = 1)
plot(density(GSE59509exprs_NoNa[,1]), col="red",ylim=c(0,5.4) ,xlab = "DNA Methylation Le
vel -Beta values", main="Beta values in all samples")
for (i in 2:42) {lines(density(GSE59509exprs_NoNa[,i]),col=colors[i]) }
# or plot(density(na.omit(GSE59509exprs[,1])), col="red",ylim=c(0,5.4) ,xlab = "DNA Methylation
Level -Beta values", main="Beta values in all samples")

# Removing outlier
knitr::kable(GSE59509pheno[1,c(4,10,6,13)])

GSE59509exprs_NoOut=GSE59509exprs_NoNa[,-c(1)]

# New dimension of the dataset
dim(GSE59509exprs_NoOut)
```

```
probeII= subset(GSE59509feature, Infinium_Design_Type == "II" , select=ID )
probeI= subset(GSE59509feature, Infinium_Design_Type == "I" , select=ID )

probeI=as.character(probeI$ID)
probeII=as.character(probeII$ID)

#setting type I and II infinium
GSE59509exprsI=GSE59509exprs_NoOut[rownames(GSE59509exprs_NoOut) %in% probeI,]
GSE59509exprsII=GSE59509exprs_NoOut[rownames(GSE59509exprs_NoOut) %in% probeII,
]

plot(density(GSE59509exprsI[,1],from=0, to=1), col="red",ylim=c(0,7) ,xlab = "DNA Methylati
on Level (Beta values)", main=" Raw Data",cex.main=2.3, font.lab=2,cex.lab=1.5)
for (i in 2:41) {lines(density(GSE59509exprsI[,i],from=0, to=1), col="red")}
for (i in 1:41) { lines(density(GSE59509exprsII[,i],from=0, to=1), col="blue")}
legend(x=0.3,y=7,c("Infinium I","Infinium II"),cex=1.6,col=c("Red","blue"),lwd=3.5,bty="n")
```

```
NewGSE59509feature<-GSE59509feature[ rownames(GSE59509feature) %in% rownames(G
SE59509exprs_NoOut), ]
Probe_design=as.character(NewGSE59509feature$Infinium_Design_Type)
pro<-gsub("I", "1", Probe_design)
pro<-gsub("11", "2", pro)
pro =as.integer(pro)
# Checking if the length of probe design is equal to number of probes in the data
length(pro) == nrow(GSE59509exprs_NoOut)
```

```
# Normalisation
library(wateRmelon)
GSE59509exprsList=list()
for (i in 1:41) {GSE59509exprsList[[i]]<-BMIQ(GSE59509exprs_NoOut[,i], pro, doH = TRUE, nfit
= 50000, th1.v = c(0.2, 0.75), th2.v = NULL, niter = 5, tol = 0.001)}
# Joining the data togather in dataframe
NormalGSE59509<-data.frame("1"= GSE59509exprsList[[1]]$nbeta, "2"= GSE59509exprsList[[
2]]$nbeta, "3"= GSE59509exprsList[[3]]$nbeta,"4" = GSE59509exprsList[[4]]$nbeta, "5"= GSE5
9509exprsList[[5]]$nbeta, "6"= GSE59509exprsList[[6]]$nbeta,  "7"= GSE59509exprsList[[7]]$nb
eta, "8"= GSE59509exprsList[[8]]$nbeta, "9"= GSE59509exprsList[[9]]$nbeta, "10"= GSE59509
exprsList[[10]]$nbeta,"11"= GSE59509exprsList[[11]]$nbeta,"12"=GSE59509exprsList[[12]]$nb
eta,"13"=GSE59509exprsList[[13]]$nbeta,"14"=GSE59509exprsList[[14]]$nbeta,"15"=GSE5950
9exprsList[[15]]$nbeta,"16"=GSE59509exprsList[[16]]$nbeta,"17"=GSE59509exprsList[[17]]$nb
eta,"18"=GSE59509exprsList[[18]]$nbeta,"19"=GSE59509exprsList[[19]]$nbeta,"20"=GSE5950
9exprsList[[20]]$nbeta,"21"=GSE59509exprsList[[21]]$nbeta,"22"=GSE59509exprsList[[22]]$nb
eta,"23"=GSE59509exprsList[[23]]$nbeta,"24"=GSE59509exprsList[[24]]$nbeta,"25"=GSE5950
9exprsList[[25]]$nbeta,"26"=GSE59509exprsList[[26]]$nbeta,"27"=GSE59509exprsList[[27]]$nb
eta,"28"=GSE59509exprsList[[28]]$nbeta,"29"=GSE59509exprsList[[29]]$nbeta,"30"=GSE5950
9exprsList[[30]]$nbeta,"31"=GSE59509exprsList[[31]]$nbeta,"32"=GSE59509exprsList[[32]]$nb
eta,"33"=GSE59509exprsList[[33]]$nbeta,"34"=GSE59509exprsList[[34]]$nbeta,"35"=GSE5950
9exprsList[[35]]$nbeta,"36"=GSE59509exprsList[[36]]$nbeta,"37"=GSE59509exprsList[[37]]$nb
```

```
eta,"38"=GSE59509exprsList[[38]]$nbeta,"39"=GSE59509exprsList[[39]]$nbeta,"40"=GSE5950
9exprsList[[40]]$nbeta,"41"=GSE59509exprsList[[41]]$nbeta, row.names=(row.names(GSE595
09exprs_NoOut)))
# Naming the samples with their original accession ID
colnames(NormalGSE59509)<-colnames(GSE59509exprs)
```

```
# Setting vectors for type I and II infinium probes for plotting
NormalGSE59509typeI=NormalGSE59509[rownames(NormalGSE59509) %in% probeI,]
NormalGSE59509typeII=NormalGSE59509[rownames(NormalGSE59509) %in% probeII,]
# Ploting
plot(density(NormalGSE59509typeI[,1],from=0,to=1), col="red",ylim=c(0,7) ,xlab = "DNA Meth
ylation Level (Beta values)", main=" Normalised Data",cex.main=2.2, font.lab=2,cex.lab=1.5
)
for (i in 2:41) {lines(density(NormalGSE59509typeI[,i],from=0,to=1), col="red")}
for (i in 1:41) { lines(density(NormalGSE59509typeII[,i],from=0,to=1), col="blue")}
legend(x=0.3,y=6.5,c("Infinium I","Infinium II"),cex=1.6,col=c("Red","blue"),lwd=3.5,bty="n")

palette(rainbow(6))
tissues<-(as.character(GSE59509pheno[-1,13]))
tissues<-sub("tissue:", "", tissues)
boxplot(NormalGSE59509[,c(1,2,3,4,5,28,29,30,31,38,39,6,7,8,9,10,11,41,40,35,34,33,32,12,1
3,14,15,16,17,18,19,20,21,22,23,24,25,26,27,37,36)],ylab = "DNA methylation level", notch=F
ALSE,las=2, main="Distribution of Beta-values Across Tissues",col=c(rep(2,11),rep(3,12),r
ep(4,3),rep(1,3),rep(5,12)), cex.axis = 0.7, names=tissues[c(1,2,3,4,5,28,29,30,31,38,39,6,7,8,9
,10,11,41,40,35,34,33,32,12,13,14,15,16,17,18,19,20,21,22,23,24,25,26,27,37,36)])
```

```
# Probe filtration

chr<-c("X","Y")
t<-subset(GSE59509feature, CHR %in% chr) # where chr is character containing "X" "Y"
sexprobes<-as.character(t$ID)
length(sexprobes)
xfiltered_NormalGSE59509=NormalGSE59509[!rownames(NormalGSE59509) %in% sexprobe
s, ]
dim(xfiltered_NormalGSE59509)

PriceAnno<-read.csv("/Users/husainalsaleh/AgePrediction_cache/PriceAno.csv", header
= TRUE)
names(PriceAnno)
length(which(PriceAnno$XY_Hits == "XY_YES"))
XY_hits<-which(PriceAnno$XY_Hits == "XY_YES")
XY_hits_ProbeID<-as.character(PriceAnno[XY_hits,1])
xfiltered_XYHit_NormalGSE59509=xfiltered_NormalGSE59509[! rownames(xfiltered_NormalG
SE59509) %in% XY_hits_ProbeID, ]
length(which(PriceAnno$Autosomal_Hits == "A_YES"))
Autosomal_hits<-which(PriceAnno$Autosomal_Hits == "A_YES")
Autosomal_hits_ProbeID<-as.character(PriceAnno[Autosomal_hits,1])
```

```
xfiltered_XYHit_Auto_NormalGSE59509<-xfiltered_XYHit_NormalGSE59509[! rownames(xfilter
ed_XYHit_NormalGSE59509) %in% Autosomal_hits_ProbeID,]

SNPprobes=as.character(PriceAnno$SNPprobe)
SNPprobes=sub("","nosnp", SNPprobes) # To fecilitate the imputation of probes with SNP
noSNP_position<-which(SNPprobes == "nosnp")
length(noSNP_position)
Probe_Name_No_SNP<-as.character(PriceAnno[ noSNP_position,1])
xfiltered_XYHit_Auto_SNP_NormalGSE59509<-xfiltered_XYHit_Auto_NormalGSE59509[rowna
mes(xfiltered_XYHit_Auto_NormalGSE59509) %in% Probe_Name_No_SNP,]

# Rename the filtered expression data
# NorFilter_GSE59509=xfiltered_XYHit_Auto_SNP_NormalGSE59509

NorFilter_GSE59509 <- read.table("/Users/husainalsaleh/AgePrediction_cache/NorFilter_
GSE59509_Original.txt", header = TRUE)
```

```
# SVD

Row_Mean<-rowMeans(NorFilter_GSE59509)
NorFilter_GSE59509_Centered<-NorFilter_GSE59509 - Row_Mean
# Then preform svd analysis
svd1=svd(NorFilter_GSE59509_Centered)

plot(svd1$d^2/sum(svd1$d^2),main="The Precent Variance Explained by Singular Values
",ylab="Precent Explained", col="blue", cex=2, pch=19, cex.main=1.5, cex.lab=1.4,cex.axis=.9
, xlab="No. of Singular values")

ggplot(as.data.frame(svd1$v), aes(svd1$v[,1],svd1$v[,2],color=tissues)) + geom_point(size=7
,alpha=0.7)+ggtitle("Singular Value Decomposition (SVD)") + xlab("SV1") +ylab("SV2")+la
bs(color="Tissue Type",face="bold",size=25)+ theme(plot.title = element_text(size=21,face="
bold"))+ theme(axis.title.x = element_text(size=15,face="bold"))+theme(axis.title.y=element_t
ext(size=15,face="bold"))

ggplot(as.data.frame(svd1$v), aes(svd1$v[,2],svd1$v[,3],color=tissues)) + geom_point(size=7
,alpha=0.7)+ggtitle("Singular Value Decomposition (SVD)") + xlab("SV2") +ylab("SV3")+la
bs(color="Tissue Type",face="bold",size=25)+ theme(plot.title = element_text(size=21,face="
bold"))+ theme(axis.title.x = element_text(size=15,face="bold"))+theme(axis.title.y=element_t
ext(size=15,face="bold"))

# Required packages for Cluster Analysis

library(devtools)
library(Biobase)
library(dendextend)

# First we add tissue type as sample's name
NorFilter_GSE59509_Tissuenames = NorFilter_GSE59509
colnames(NorFilter_GSE59509_Tissuenames) <- tissues
hclust1 <- hclust(dist(t(NorFilter_GSE59509_Tissuenames)))
dend <- as.dendrogram(hclust1)
```

```
dend <- color_labels(hclust1, 3, col = 1:4)
par(mar = c(5, 5, 4, 9))
plot( dend, xlab = "Distance", main = "Dendrogram", horiz = T, cex.main = 2.5, cex.lab = 1.2, f
ont.lab = 2)



# Spearman's rank correlation
cl <- makeCluster(2)
registerDoParallel(cl)
Cor.rvalues <- foreach (i = 1 : (ncol(Train_data_transp)-1))  %dopar% { cor(Train_data_transp[,
ncol(Train_data_transp)], Train_data_transp[,i], method = "spearman") } # Spearman
Cor.r<-unlist(Cor.rvalues)

cl<-makeCluster(2)
registerDoParallel(cl)
Cor.pvalues<- foreach (i= 1:(ncol(Train_data_transp)-1) ) %dopar% {cor.test(Train_data_trans
p[,ncol(Train_data_transp)], Train_data_transp[,i], method = "spearman")$p.value }
Cor.p<-unlist(Cor.pvalues)

# The qqunit.plot function has been loaded but not shown here in the report
qqunif.plot(Cor.p)

 objM<- qvalue(Cor.p,  pi0.method = "smoother", fdr.level= 0.05)
summary(objM)

plot(objM)

# the significant probes under the fdr.level 0.05 will have TURE in obj$significant
ProbesSigTFM<-objM$significant
sigProbes_Cor<-(which(ProbesSigTFM == TRUE))
SigPvalues_Cor1<-Cor.p[sigProbes_Cor]
SigPvalues_Cor<-unlist(SigPvalues_Cor1)

# Spearman's with M value
cl <- makeCluster(2)
registerDoParallel(cl)
Cor.rvaluesM <- foreach (i = 1 : (ncol(Train_data_M_age)-1))  %dopar% { cor(Train_data_M_a
ge[,ncol(Train_data_M_age)], Train_data_M_age[,i], method = "spearman") } # Spearman
Cor.rM<-unlist(Cor.rvaluesM)

cl<-makeCluster(2)
registerDoParallel(cl)
Cor.pvaluesM<- foreach (i= 1:(ncol(Train_data_M_age)-1) ) %dopar% {cor.test(Train_data_M_
age[,ncol(Train_data_M_age)], Train_data_M_age[,i], method = "spearman")$p.value }
Cor.pM<-unlist(Cor.pvaluesM)

# The qqunit.plot function has been loaded but not shown here in the report
qqunif.plot(Cor.pM)

 objMM<- qvalue(Cor.pM,  pi0.method = "smoother", fdr.level= 0.05)
summary(objMM)

plot(objMM)
```

```
# the significant probes under the fdr.level 0.05 will have TURE in obj$significant
ProbesSigTFMM<-objMM$significant
sigProbes_CorM<-(which(ProbesSigTFMM == TRUE))
SigPvalues_Cor1M<-Cor.pM[sigProbes_CorM]
SigPvalues_CorM<-unlist(SigPvalues_Cor1M)



# Pearson's correlation with Beta value
Cor.rvaluesP <- foreach (i = 1 : (ncol(Train_data_transp)-1))  %dopar% { cor(Train_data_trans
p[,ncol(Train_data_transp)], Train_data_transp[,i], method = "pearson") } # pearson
Cor.rP<-unlist(Cor.rvaluesP)



Cor.pvaluesP<- foreach (i= 1:(ncol(Train_data_transp)-1) ) %dopar% {cor.test(Train_data_tran
sp[,ncol(Train_data_transp)], Train_data_transp[,i], method = "pearson")$p.value }
Cor.pP<-unlist(Cor.pvaluesP)

# The qqunit.plot function has been loaded but not shown here in the report
qqunif.plot(Cor.pP)

objMP<- qvalue(Cor.pP,  pi0.method = "smoother", fdr.level= 0.05)
summary(objMP)

# the significant probes under the fdr.level 0.05 will have TURE in obj$significant
ProbesSigTFMP<-objMP$significant
sigProbes_CorP<-(which(ProbesSigTFMP == TRUE))
SigPvalues_Cor1P<-Cor.pP[sigProbes_CorP]
SigPvalues_CorP<-unlist(SigPvalues_Cor1P)


# Pearson's with M value

Cor.rvaluesMP <- foreach (i = 1 : (ncol(Train_data_M_age)-1))  %dopar% { cor(Train_data_M_
age[,ncol(Train_data_M_age)], Train_data_M_age[,i], method = "pearson") } # Spearman
Cor.rMP<-unlist(Cor.rvaluesMP)

Cor.pvaluesMP<- foreach (i= 1:(ncol(Train_data_M_age)-1) ) %dopar% {cor.test(Train_data_
M_age[,ncol(Train_data_M_age)], Train_data_M_age[,i], method = "pearson")$p.value }
Cor.pMP<-unlist(Cor.pvaluesMP)

# The qqunit.plot function has been loaded but not shown here in the report
qqunif.plot(Cor.pMP)

objMMP<- qvalue(Cor.pMP,  pi0.method = "smoother", fdr.level= 0.05)
summary(objMMP)

# the significant probes under the fdr.level 0.05 will have TURE in obj$significant
ProbesSigTFMMP<-objMMP$significant
sigProbes_CorMP<-(which(ProbesSigTFMMP == TRUE))
SigPvalues_Cor1MP<-Cor.pMP[sigProbes_CorMP]
SigPvalues_CorMP<-unlist(SigPvalues_Cor1MP)
```

```
# Simple linear regression

CoefEsts.for = matrix(NA, nrow = (ncol(Train_data_transp)-1), ncol = 3) # this is to design matri
x for inputs
rownames(CoefEsts.for) = colnames(Train_data_transp)[1:(ncol(Train_data_transp)-1)]
colnames(CoefEsts.for) = c("Intercept", "Coef","Pvalue")
for (i in 1:(ncol(Train_data_transp)-1)) { fit = lm(Train_data_transp[,ncol(Train_data_transp)]~Tr
ain_data_transp[,i]) ;CoefEsts.for[i,] = summary(fit)[[4]][c(1,2,8)] }

# Making a vector of pvalues
Reg.p=CoefEsts.for[,3]
# Q-Q plot
qqunif.plot(Reg.p)

Reg_objM<- qvalue(Reg.p,  pi0.method = "smoother", fdr.level= 0.05)
summary(Reg_objM)

# the significant probes under the fdr.level 0.05 will have TURE in obj$significant
ProbesSigReg <- Reg_objM$significant
sigProbes_Reg <- (which(ProbesSigReg == TRUE))
SigPvalues_Reg1 <- Reg.p[sigProbes_Reg]
SigPvalues_Reg <- unlist(SigPvalues_Reg1)

ggplot(data = as.data.frame(Train_data_transp_Cor_Sig), aes(x = Train_data_transp[, ncol(Tr
ain_data_transp)], y = Train_data_transp_Cor_Sig[, names(SigPvalues_Reg)])) + geom_point(
size=1.5) + geom_smooth(method=lm, fill="pink",se=TRUE) + xlab("Chronological Age") +yl
ab("Methylation level") +ggtitle(paste(names(SigPvalues_Reg),"marker Vs. Age"))+theme(
plot.title = element_text(size=20,face="bold")) +theme(axis.title.x = element_text(size=15,face
="bold"))+theme(axis.title.y=element_text(size=15,face="bold"))

# Converting DNA methylation level from beta to M-values for simple regression
Train_data_No_age <- as.matrix(Train_data_age[-nrow(Train_data_age), ])
Train_data_No_age_Mvalue <- Beta2M(Train_data_No_age)
Train_data_age_Mvalue <- rbind(Train_data_No_age_Mvalue, Train_data_age[nrow(Train_dat
a_age), ])
Train_data_age_M_Transp <- t(Train_data_age_Mvalue)
# Simple linear regression code
CoefEsts.forM = matrix(NA, nrow = (ncol(Train_data_age_M_Transp)-1), ncol = 3) # this is to d
esign matrix for inputs
rownames(CoefEsts.forM) = colnames(Train_data_age_M_Transp)[1:(ncol(Train_data_age_M
_Transp)-1)]
colnames(CoefEsts.forM) = c("Intercept", "Coef","Pvalue")
for (i in 1:(ncol(Train_data_age_M_Transp)-1)) { fit = lm(Train_data_age_M_Transp[,ncol(Train
_data_age_M_Transp)]~Train_data_age_M_Transp[,i]) ;CoefEsts.forM[i,] = summary(fit)[[4]][c(
1,2,8)] }

# Making a vector of pvalues
RegM.p <- CoefEsts.forM[, 3]
# Q-Q plot
qqunif.plot(RegM.p)

RegM_objM<- qvalue(RegM.p,  pi0.method = "smoother", fdr.level= 0.05)
summary(RegM_objM)
```

```
# the significant probes under the fdr.level 0.05 will have TURE in obj$significant
ProbesSigRegM <- RegM_objM$significant
sigProbes_RegM <- (which(ProbesSigRegM == TRUE))
SigPvalues_RegM1 <- RegM.p[sigProbes_RegM]
SigPvalues_RegM <- unlist(SigPvalues_RegM1)
names(SigPvalues_RegM)

GSE92767 <- getGEO("GSE92767",getGPL = FALSE,filename = "GSE92767_series_matrix
.txt.gz")

# Storing expression data, phenotype data
GSE92767exprs <- as.data.frame(exprs(GSE92767))
GSE92767pheno<-pData(phenoData(GSE92767))

# Feature data

GSE59509feature <-read.table("/Users/husainalsaleh/AgePrediction_cache/GSE59509fe
ature.txt", header = TRUE)

Training_AgeDist  <-as.integer(sub("age: ","",pData(phenoData(GSE92767))[,10],fixed = TRU
E))
Training_AgeDist <- as.data.frame(Training_AgeDist)
colnames(Training_AgeDist)[1]<-"Chronological age"

par(mar=c(5.1 ,6.5 ,4.1 ,2.1))
par(bg = "gray99")

hist(Training_AgeDist$`Chronological age`,
     main="",
     xlab="Chronological Age", ylab="Sample Size",
     freq=TRUE, border="black"
 , ylim=c(0,13),cex.lab=2,cex.main=3, font=2, font.lab=2)

# Looking for any missing data (NA)
sum(is.na(GSE92767exprs))

dat1 <- cbind(rownames(GSE92767exprs),GSE92767exprs)
colnames(dat1)[1] <-"ProbeID"

fastImputation= FALSE
nSamples=dim(dat1)[[2]]-1
nProbes= dim(dat1)[[1]]

meanMethBySample =as.numeric(apply(as.matrix(dat1[,-1]),2,mean,na.rm=TRUE))
minMethBySample   =as.numeric(apply(as.matrix(dat1[,-1]),2,min,na.rm=TRUE))
maxMethBySample  =as.numeric(apply(as.matrix(dat1[,-1]),2,max,na.rm=TRUE))

datMethUsed= t(dat1[,-1])
colnames(datMethUsed)=as.character(dat1[,1])


noMissingPerSample=apply(as.matrix(is.na(datMethUsed)),1,sum)
```

```
table(noMissingPerSample)


#STEP 2:
if (! fastImputation & nSamples>1 & max(noMissingPerSample,na.rm=TRUE)<3000 ){

  # run the following code if there is at least one missing
  if ( max(noMissingPerSample,na.rm=TRUE)>0 ){
    dimnames1=dimnames(datMethUsed)
    datMethUsed= data.frame(t(impute.knn(t(datMethUsed))$data))
    dimnames(datMethUsed)=dimnames1
  } # end of if
} # end of if (! fastImputation )

if ( max(noMissingPerSample,na.rm=TRUE)>=3000 ) fastImputation=TRUE


if ( fastImputation | nSamples==1 ){
  noMissingPerSample=apply(as.matrix(is.na(datMethUsed)),1,sum)
  table(noMissingPerSample)
  if ( max(noMissingPerSample,na.rm=TRUE)>0 & max(noMissingPerSample,na.rm=TRUE) >=
3000 ) {normalizeData=FALSE}

  # run the following code if there is at least one missing
  if ( max(noMissingPerSample,na.rm=TRUE)>0 & max(noMissingPerSample,na.rm=TRUE) < 3
000 ){
    dimnames1=dimnames(datMethUsed)
    for (i in which(noMissingPerSample>0) ){
      selectMissing1=is.na(datMethUsed[i,])
      datMethUsed[i,selectMissing1] = as.numeric(probeAnnotation21kdatMethUsed$goldstandar
d2[selectMissing1])
    } # end of for loop
    dimnames(datMethUsed)=dimnames1
  } # end of if
} # end of if (! fastImputation )


GSE92767exprs_NoNa <- t(datMethUsed)

colors <- terrain.colors(54, alpha = 1)
plot(density(GSE92767exprs_NoNa[,1]), col="orange",ylim=c(0,5.4) ,xlab = "DNA Methylatio
n Level -Beta values", main="Beta values in all samples")
for (i in 2:54) {lines(density(GSE92767exprs_NoNa[,i]),col="orange") }


library(dplyr)

probeII= filter(GSE59509feature, Infinium_Design_Type=="II") %>% select(ID)
probeI= filter(GSE59509feature, Infinium_Design_Type=="I") %>% select(ID)

probeI=as.character(probeI$ID)
probeII=as.character(probeII$ID)
```

```
#setting type I and II infinium
GSE92767exprsI=GSE92767exprs_NoNa[rownames(GSE92767exprs_NoNa) %in% probeI,]
GSE92767exprsII=GSE92767exprs_NoNa[rownames(GSE92767exprs_NoNa) %in% probeII,]

plot(density(as.matrix(rowMeans(GSE92767exprsI))[,1]), col="red",ylim=c(0,7) ,xlab = "Beta
values", main="",cex.main=2.3, font.lab=2,cex.lab=1.5, lwd=5)
lines(density(as.matrix(rowMeans(GSE92767exprsII))[,1]), lwd=4,col="Blue")

legend(x=0.3,y=7,c("Infinium I","Infinium II"),cex=1.6,col=c("Red","blue"),lwd=3.5,bty="n")

NewGSE59509feature<-GSE59509feature[ rownames(GSE59509feature) %in% rownames(G
SE92767exprs_NoNa), ]
Probe_design=as.character(NewGSE59509feature$Infinium_Design_Type)
pro<-gsub("I", "1", Probe_design)
pro<-gsub("11", "2", pro)
pro =as.integer(pro)
# Checking if the length of probe design is equal to number of probes in the data
length(pro) == nrow(GSE92767exprs_NoNa)

plot(density(as.matrix(rowMeans(GSE92767exprsI))[,1]), col="red",ylim=c(0,7) ,xlab = "Beta
values", main="",cex.main=2.3, font.lab=2,cex.lab=1.5, lwd=5)
lines(density(as.matrix(rowMeans(NormalGSE92767typeII))[,1]), col="blue", lwd=5)
lines(density(as.matrix(rowMeans(GSE92767exprsII))[,1]), col="blue", lwd=5, lty="dotted")
legend(x=0.1,y=8.2,c("Type I","Type II (before normaliztion)", "Type II (after normaliztion)
"),cex=1.2,col=c("Red","blue", "blue"),lwd=3.5,bty="n", lty=c("solid", "dotted","solid"))

par(mfrow=c(1,2))

boxplot(GSE92767exprs_NoNa[,1:54],names=1:54,ylab = "DNA methylation level", notch=F
ALSE,las=2, main="Distribution of Beta-values Across Tissues", cex.axis = 0.7)

boxplot(NormalGSE92767[,1:54],names=1:54,ylab = "DNA methylation level", notch=FALSE,
las=2, main="Distribution of Beta-values Across Tissues", cex.axis = 0.7)

# SVD

Row_Mean<-rowMeans(NormalGSE92767)
Nor_GSE92767_Centered<-NormalGSE92767 - Row_Mean
# Then preform svd analysis
svd1=svd(Nor_GSE92767_Centered)

plot(svd1$d^2/sum(svd1$d^2),main="The Precent Variance Explained by Singular Values
",ylab="Precent Explained", col="blue", cex=2, pch=19, cex.main=1.5, cex.lab=1.4,cex.axis=.9
, xlab="No. of Singular values")

ggplot(as.data.frame(svd1$v), aes(svd1$v[,1],svd1$v[,2],color="pink")) + geom_point(size=7,
alpha=0.7)+ggtitle("Singular Value Decomposition (SVD)") + xlab("SV1") +ylab("SV2")+la
bs(color="Tissue Type",face="bold",size=25)+ theme(plot.title = element_text(size=21,face="
bold"))+ theme(axis.title.x = element_text(size=15,face="bold"))+theme(axis.title.y=element_t
ext(size=15,face="bold"))+geom_text(aes(label=GSE92767pheno[,2]),hjust=0.9, vjust=0)
```

```
ggplot(as.data.frame(svd1$v), aes(svd1$v[,2],svd1$v[,3],color="pink")) + geom_point(size=7,
alpha=0.7)+ggtitle("Singular Value Decomposition (SVD)") + xlab("SV2") +ylab("SV3")+la
bs(color="Tissue Type",face="bold",size=25)+ theme(plot.title = element_text(size=21,face="
bold"))+ theme(axis.title.x = element_text(size=15,face="bold"))+theme(axis.title.y=element_t
ext(size=15,face="bold"))+geom_text(aes(label=GSE92767pheno[,2]),hjust=0.9, vjust=0)

newGSE92767feature <- GSE59509feature[ rownames(GSE59509feature) %in% rownames(
NormalGSE92767), ]
SNPprobes_10=as.character(newGSE92767feature$Probe_SNPs_10)
SNPprobes=sub("","nosnp", SNPprobes_10, fixed = FALSE) # To fecilitate the imputation of pr
obes with SNP
noSNP_position<-which(SNPprobes == "nosnp")
length(noSNP_position)

GSE92767exprs_Nosnp <- NormalGSE92767[noSNP_position,]
```

```
library(doParallel)

# Extracting age covariate from phenotype data
Train_Age1<-gsub("age:","",as.vector(GSE92767pheno[,10]),fixed = TRUE)
Train_Age<-as.integer(Train_Age1)
# Preparing training dataset for correlation test
Train_data_age=rbind(GSE92767exprs_Nosnp,Train_Age)
Train_data_transp=t(Train_data_age)
colnames(Train_data_transp)[ncol(Train_data_transp)]<-"age"
# Correlation test - recording r values in Cor.rvalues for each marker
# First parallelize the analyses
cl <- makeCluster(2)
registerDoParallel(cl)
Cor.rvalues <- foreach (i = 1 : (ncol(Train_data_transp)-1))  %dopar% { cor(Train_data_transp[
,ncol(Train_data_transp)], Train_data_transp[,i], method = "spearman") } # Spearman
Cor.r<-unlist(Cor.rvalues)

hist(Cor.r, main=" Distribution of Spearman coefficients",cex.main=2,ylim = c(0,1.9) ,xlim=c(
-0.7,0.7),xlab=expression("r"[s]),font.lab=2 ,ylab="Density" ,breaks=80, border="pink", col="co
rnflowerblue", las=1, prob = TRUE, cex.lab=1.5)

Cor.pvalues<- foreach (i= 1:(ncol(Train_data_transp)-1) ) %dopar% {cor.test(Train_data_trans
p[,ncol(Train_data_transp)], Train_data_transp[,i], method = "spearman")$p.value }
Cor.p<-unlist(Cor.pvalues)

# The qqunit.plot function has been loaded but not shown here in the report
qqunif.plot(Cor.p)

library(qvalue)
objM<- qvalue(Cor.p,  pi0.method = "smoother", fdr.level= 0.00000001)
summary(objM)

# the significant probes under the fdr.level will have TURE in obj$significant
ProbesSigFDR <- objM$significant
sigProbes_FDR <- (which(ProbesSigFDR == TRUE))
```

```
SigPvalues_FDR1 <- Cor.p[sigProbes_FDR]
SigPvalues_FDR <- unlist(SigPvalues_FDR1)
length(SigPvalues_FDR)
```

```
# preparing train data with only probes passed FDR and then calculate their rho
Train_data_FDR=t(Train_data_transp[,sigProbes_FDR])
Train_data_FDR_age <-rbind(Train_data_FDR, Train_data_transp[,ncol(Train_data_transp)] )
rownames(Train_data_FDR_age)[nrow(Train_data_FDR_age)] <-"age"
Train_data_FDR_age_trans <- t(Train_data_FDR_age)

# script to calculate rho
cl <- makeCluster(2)
registerDoParallel(cl)
Cor.rvaluesFDR <- foreach (i = 1 : (ncol(Train_data_FDR_age_trans)-1))  %dopar% { cor(Trai
n_data_FDR_age_trans[,ncol(Train_data_FDR_age_trans)], Train_data_FDR_age_trans[,i], me
thod = "spearman") } # Spearman
Cor.rFDR<-unlist(Cor.rvaluesFDR)

length(which(Cor.rFDR >= 0.6))
length(which(Cor.rFDR <= - 0.6))
length(which(Cor.rFDR > 0.6)) + length(which(Cor.rFDR < - 0.6))

Train_data_No_age <- as.matrix(Train_data_FDR_age_trans[,which(Cor.rFDR > 0.6 | Cor.rFD
R < - 0.6)])

# writing script to remove the markers with less than 0.1 difference between min and max methyl
ation level
difereFDR <-vector()
for (i in 1:ncol(Train_data_No_age)) { difereFDR[i]<-max(Train_data_No_age[,i]) - min(Train_d
ata_No_age[,i])}
which(difereFDR < 0.1)

Train_data_No_age_Dif <- Train_data_No_age[,-c(which(difereFDR < 0.1))]
Train_data_No_age_Mvalue <- Beta2M(Train_data_No_age_Dif)

Train_data_age_Mvalue <- cbind(Train_data_No_age_Mvalue,Train_data_FDR_age_trans[,nc
ol(Train_data_FDR_age_trans)])
colnames(Train_data_age_Mvalue)[ncol(Train_data_age_Mvalue)]<-"age"

#Step-wise Regression
regfitfull_FDR=regsubsets(age~., data = as.data.frame(Train_data_age_Mvalue), nvmax = 10,
nbest=1, method = "exhaustive", really.big = T)
reg.summary_FDR=summary(regfitfull_FDR)
plot(reg.summary_FDR$bic, xlab="Number of variables", cex.lab=1.3,ylab="BIC stats",cex=2
, col="red",pch=20,main=" Bayesian information criterion (BIC) ",font.lab=2,cex.main=2)
abline(v=which.min(reg.summary_FDR$bic),  lwd=2, col="blue")
```

The names of the CpG sites that are included in the best model are:

```
names(coef(regfitfull_FDR,(which.min(reg.summary_FDR$bic))))[-1]
```

```
# Prepare the data for multivariate linear regression
Train_data_transp1<-Train_data_transp[,-449043]
Train_data_transp2<-Beta2M(Train_data_transp1)
Train_data_transp_M_age <- cbind(Train_data_transp2, Train_data_transp[,449043])
colnames(Train_data_transp_M_age)[449043]<- "age"

# Regression
Model_M_SigFDR <- lm(age~cg00573770+cg04875128+cg06279276+cg07365960+cg10501210+cg10804656+cg16867657+cg23606718+cg25124276, data= as.data.frame(Train_data_transp_M_age))

summary(Model_M_SigFDR)

mean(abs(residuals(Model_M_SigFDR)))
```

```
#Residual sum of squares:

RSS_FDR <- sum(abs(Model_M_SigFDR$residuals))

#Mean squared error:

MSE_FDR <- RSS_FDR / length(Model_M_SigFDR$residuals)

# Root MSE:

RMSE_FDR <- sqrt(MSE_FDR)
print(RMSE_FDR)

# Ploting predicted vs actual

ggplot(as.data.frame(Train_data_transp_M_age), aes(x = Train_data_transp_M_age[,449043], y = predict(Model_M_SigFDR), color="Saliva")) + geom_point(size=4.5,alpha=0.9, color="red")+ xlab("Actual Age") +ylab("Predicted Age")+labs(color="Tissue Type",face="bold",size=3)+ geom_abline(slope=1, intercept=0, size=1, color="black")+theme(plot.title = element_text(size=21,face="bold"))+ theme(axis.title.x = element_text(size=18,face="bold"))+theme(axis.title.y=element_text(size=18,face="bold"))+coord_fixed(ratio=3/4) +xlim(15, 75) +ylim(15, 75)
```

```
# downloading GSE99029 for validation

GSE99029 <- getGEO("GSE99029",getGPL = FALSE,filename = "GSE99029_series_matrix.txt.gz")
# Storing expression data, phenotype data, and feature data
GSE99029exprs <- as.data.frame(exprs(GSE99029))
GSE99029pheno<-pData(phenoData(GSE99029))
```

```
dim(GSE99029exprs)
# Showing 5 samples x 5 CpG probes
GSE99029exprs[1:5,1:5]
# Discriptive stats for the first four samples
summary(GSE99029exprs[,1:4])

# Looking for any missing data (NA)
sum(is.na(GSE99029exprs))

dat1 <- cbind(rownames(GSE99029exprs),GSE99029exprs)
colnames(dat1)[1] <-"ProbeID"

fastImputation= FALSE
nSamples=dim(dat1)[[2]]-1
nProbes= dim(dat1)[[1]]

meanMethBySample =as.numeric(apply(as.matrix(dat1[,-1]),2,mean,na.rm=TRUE))
minMethBySample   =as.numeric(apply(as.matrix(dat1[,-1]),2,min,na.rm=TRUE))
maxMethBySample  =as.numeric(apply(as.matrix(dat1[,-1]),2,max,na.rm=TRUE))

datMethUsed= t(dat1[,-1])
colnames(datMethUsed)=as.character(dat1[,1])


noMissingPerSample=apply(as.matrix(is.na(datMethUsed)),1,sum)
table(noMissingPerSample)


#STEP 2: Imputing
if (! fastImputation & nSamples>1 & max(noMissingPerSample,na.rm=TRUE)<3000 ){

  # run the following code if there is at least one missing
  if ( max(noMissingPerSample,na.rm=TRUE)>0 ){
    dimnames1=dimnames(datMethUsed)
    datMethUsed= data.frame(t(impute.knn(t(datMethUsed))$data))
    dimnames(datMethUsed)=dimnames1
  } # end of if
} # end of if (! fastImputation )

if ( max(noMissingPerSample,na.rm=TRUE)>=3000 ) fastImputation=TRUE


if ( fastImputation | nSamples==1 ){
  noMissingPerSample=apply(as.matrix(is.na(datMethUsed)),1,sum)
  table(noMissingPerSample)
  if ( max(noMissingPerSample,na.rm=TRUE)>0 & max(noMissingPerSample,na.rm=TRUE) >=
3000 ) {normalizeData=FALSE}

  # run the following code if there is at least one missing
  if ( max(noMissingPerSample,na.rm=TRUE)>0 & max(noMissingPerSample,na.rm=TRUE) < 3
000 ){
    dimnames1=dimnames(datMethUsed)
    for (i in which(noMissingPerSample>0) ){
```

```
    selectMissing1=is.na(datMethUsed[i,])
    datMethUsed[i,selectMissing1] = as.numeric(probeAnnotation21kdatMethUsed$goldstandard2[selectMissing1])
   } # end of for loop
   dimnames(datMethUsed)=dimnames1
  } # end of if
} # end of if (! fastImputation )


GSE99029exprs_NoNa <- t(datMethUsed)

colors <- terrain.colors(57, alpha = 1)
plot(density(GSE99029exprs_NoNa[,1]), col="red",ylim=c(0,5.4) ,xlab = "DNA Methylation Level -Beta values", main="Beta values in all samples")
for (i in 2:57) {lines(density(GSE99029exprs_NoNa[,i]),col="red") }

Testing_AgeDist<-as.data.frame(as.integer(gsub("age:","",as.character(GSE99029pheno[,11]))))
colnames(Testing_AgeDist)[1]<-"Chronological age"
rownames(Testing_AgeDist)<- rownames(GSE99029pheno)

par(mar=c(5.1 ,6.5 ,4.1 ,2.1))
par(bg = "gray99")

hist(Testing_AgeDist$`Chronological age`,
     main="",
     xlab="Chronological Age", ylab="Sample Size",
     freq=TRUE, border="black"
 , ylim=c(0,20),cex.lab=2,cex.main=3, font=2, font.lab=2)

Testing_age <-as.integer(gsub("age:","",as.character(GSE99029pheno[,11])))

Gender_Distribution <- as.character(gsub("gender:","",as.character(GSE99029pheno[,10])))
Gender_Distribution<-gsub(" female", "female", Gender_Distribution)
Gender_Distribution<-gsub(" male", "male", Gender_Distribution)

length(which(Gender_Distribution == "female"))

length(which(Gender_Distribution == "male"))
```

```
probeII= subset(GSE59509feature, Infinium_Design_Type == "II" , select=ID )
probeI= subset(GSE59509feature, Infinium_Design_Type == "I" , select=ID )

probeI=as.character(probeI$ID)
probeII=as.character(probeII$ID)

#setting type I and II infinium
GSE99029exprsI=GSE99029exprs_NoNa[rownames(GSE99029exprs_NoNa) %in% probeI,]
GSE99029exprsII=GSE99029exprs_NoNa[rownames(GSE99029exprs_NoNa) %in% probeII,]
```

```
plot(density(as.matrix(rowMeans(GSE99029exprsI))[,1]), col="red",ylim=c(0,7) ,xlab = "Beta values", main="",cex.main=2.3, font.lab=2,cex.lab=1.5, lwd=5)
lines(density(as.matrix(rowMeans(GSE99029exprsII))[,1]), lwd=4,col="Blue")

legend(x=0.3,y=7,c("Infinium I","Infinium II"),cex=1.6,col=c("Red","blue"),lwd=3.5,bty="n")

NewGSE59509feature<-GSE59509feature[ rownames(GSE59509feature) %in% rownames(GSE99029exprs_NoNa), ]
Probe_design=as.character(NewGSE59509feature$Infinium_Design_Type)
pro<-gsub("I", "1", Probe_design)
pro<-gsub("11", "2", pro)
pro =as.integer(pro)
# Checking if the length of probe design is equal to number of probes in the data
length(pro) == nrow(GSE99029exprs_NoNa)
```

```
# Normalisation
GSE99029exprsList<-list()
for (i in 1:57) {GSE99029exprsList[[i]]<-BMIQ(GSE99029exprs_NoNa[,i], pro, doH = TRUE, nfit = 50000, th1.v = c(0.2, 0.75), th2.v = NULL, niter = 5, tol = 0.001)}
# Joining the data togather in dataframe
NormalGSE99029<-data.frame("1"= GSE99029exprsList[[1]]$nbeta)
for (i in 2:57) {NormalGSE99029[,i]<-c(GSE99029exprsList[[i]]$nbeta)  }

# Naming the samples with their original accession ID and CpG markers
dimnames(NormalGSE99029) <-dimnames(GSE99029exprs_NoNa)
```

```
# Setting vectors for type I and II infinium probes for plotting
NormalGSE99029typeI=NormalGSE99029[rownames(NormalGSE99029) %in% probeI,]
NormalGSE99029typeII=NormalGSE99029[rownames(NormalGSE99029) %in% probeII,]
# Ploting

plot(density(as.matrix(rowMeans(NormalGSE99029typeI))[,1]), col="red",ylim=c(0,7) ,xlab = "Beta values", main="",cex.main=2.3, font.lab=2,cex.lab=1.5, lwd=5)
lines(density(as.matrix(rowMeans(NormalGSE99029typeII))[,1]), lwd=4,col="Blue")
lines(density(as.matrix(rowMeans(GSE99029exprsII))[,1]), col="blue", lwd=5, lty="dotted")
legend(x=0.1,y=8.2,c("Type I","Type II (before normalistion)", "Type II (after normalistion)"),cex=1.2,col=c("Red","blue", "blue"),lwd=3.5,bty="n", lty=c("solid", "dotted","solid"))

colors <- terrain.colors(57, alpha = 1)
plot(density(NormalGSE99029[,1]), col="red",ylim=c(0,5.4) ,xlab = "DNA Methylation Level - Beta values", main="Beta values in all samples")
for (i in 2:57) {lines(density(NormalGSE99029[,i]),col=colors[i]) }
# or plot(density(na.omit(GSE99029exprs[,1])), col="red",ylim=c(0,5.4) ,xlab = "DNA Methylation Level -Beta values", main="Beta values in all samples")

boxplot(NormalGSE99029[,1:57],ylab = "DNA methylation level", notch=FALSE,las=2, main= "Distribution of Beta-values Across Tissue", cex.axis = 0.7)
```

```
Row_Mean1<-rowMeans(NormalGSE99029)
Nor_GSE99029_Centered<-NormalGSE99029 - Row_Mean1
# Then preform svd analysis
svd11=svd(Nor_GSE99029_Centered)

plot(svd11$d^2/sum(svd11$d^2),main="The Precent Variance Explained by Singular Values",ylab="Precent Explained", col="blue", cex=2, pch=19, cex.main=1.5, cex.lab=1.4,cex.axis=.9, xlab="No. of Singular values")

ggplot(as.data.frame(svd11$v), aes(svd11$v[,1],svd11$v[,2],color=GSE99029pheno[,12])) + geom_point(size=7,alpha=0.7)+ggtitle("Singular Value Decomposition (SVD)") + xlab("SV1") +ylab("SV2")+labs(color="Tissue Type",face="bold",size=25)+ theme(plot.title = element_text(size=21,face="bold"))+ theme(axis.title.x = element_text(size=15,face="bold"))+theme(axis.title.y=element_text(size=15,face="bold"))+geom_text(aes(label=GSE99029pheno[,2]),hjust=0.9, vjust=0)

ggplot(as.data.frame(svd11$v), aes(svd11$v[,1],svd11$v[,2],color=GSE99029pheno[,10])) + geom_point(size=7,alpha=0.7)+ggtitle("Singular Value Decomposition (SVD)") + xlab("SV1") +ylab("SV2")+labs(color="Tissue Type",face="bold",size=25)+ theme(plot.title = element_text(size=21,face="bold"))+ theme(axis.title.x = element_text(size=15,face="bold"))+theme(axis.title.y=element_text(size=15,face="bold"))+geom_text(aes(label=GSE99029pheno[,2]),hjust=0.9, vjust=0)

require(sva)

mod = model.matrix(~GSE99029pheno$characteristics_ch1.2,data=GSE99029pheno) # two models one with covariates and the other (below) no covariates
mod0 = model.matrix(~1, data=GSE99029pheno) # has no covariates
sva1 = sva(as.matrix(NormalGSE99029),mod,mod0,n.sv=1)
summary(lm(sva1$sv ~ mod[,2]))

# Converting Beta to M value

NormalGSE99029_M <- Beta2M(NormalGSE99029)

# Predicting the age using Model_Sig_M
mean(abs(as.integer(gsub("age:","",as.character(GSE99029pheno[,11]))) - predict(Model_M_SigFDR, as.data.frame(t(NormalGSE99029_M)))))

Predicted_M_Sig<-predict(Model_M_SigFDR, as.data.frame(t(NormalGSE99029_M)))

Testing_age <- as.integer(gsub("age:","",as.character(GSE99029pheno[,11])))

# Ploting predicted vs actual

ggplot(as.data.frame(t(NormalGSE99029_M)), aes(x = Testing_age, y = Predicted_M_Sig, color="Saliva")) + geom_point(size=4.5,alpha=0.9, color="blue")+ xlab("Actual Age") +ylab("Predicted Age")+labs(color="Tissue Type",face="bold",size=3)+ geom_abline(slope=1, intercept=0, size=1, color="black")+theme(plot.title = element_text(size=21,face="bold"))+ theme(axis.title.x = element_text(size=18,face="bold"))+theme(axis.title.y=element_text(size=18,face="bold"))+coord_fixed(ratio=3/4) +xlim(15, 100) +ylim(15, 100)
```

```r
cor.test(Testing_age , Predicted_M_Sig)

Gender_Distribution <- as.character(gsub("gender:","",as.character(GSE99029pheno[,10])))
Gender_Distribution<-gsub(" female", "female", Gender_Distribution)
Gender_Distribution<-gsub(" male", "male", Gender_Distribution)

length(which(Gender_Distribution == "female"))

length(which(Gender_Distribution == "male"))

# female
mean(abs( Testing_age[which(Gender_Distribution == "female")] - predict(Model_M_SigFDR,
as.data.frame(t(NormalGSE99029_M)[which(Gender_Distribution == "female"),])))

# male

mean(abs( Testing_age[which(Gender_Distribution == "male")] - predict(Model_M_SigFDR, a
s.data.frame(t(NormalGSE99029_M)[which(Gender_Distribution == "male"),])))

# t.test between male and female

absPredicted_male<-abs( Testing_age[which(Gender_Distribution == "male")] - predict(Model
_M_SigFDR, as.data.frame(t(NormalGSE99029_M)[which(Gender_Distribution == "male"),]))

absPredicted_female<-abs( Testing_age[which(Gender_Distribution == "female")] - predict(M
odel_M_SigFDR, as.data.frame(t(NormalGSE99029_M)[which(Gender_Distribution == "femal
e"),]))

t.test(absPredicted_female, absPredicted_male)

# cor.test

cor.test(Testing_age[which(Gender_Distribution == "male")], predict(Model_M_SigFDR, as.d
ata.frame(t(NormalGSE99029_M)[which(Gender_Distribution == "male"),])))

cor.test(Testing_age[which(Gender_Distribution == "female")], predict(Model_M_SigFDR, as.
data.frame(t(NormalGSE99029_M)[which(Gender_Distribution == "female"),])))


mgsub <- function(pattern, replacement, x, ...) {
  if (length(pattern)!=length(replacement)) {
    stop("pattern and replacement do not have the same length.")
  }
  result <- x
  for (i in 1:length(pattern)) {
    result <- gsub(pattern[i], replacement[i], result, ...)
  }
  result
}

#### Bootstrap analysis for the prediction model
```

```
NormalGSE99029_model <- t(NormalGSE99029_M[The_model,])
names(Testing_age) <- rownames(GSE99029pheno)

# Bootstrap function
ss <-vector()
for (i in 1:10000 ) {
  n <- NormalGSE99029_model[sample(rownames(NormalGSE99029_model),57, replace = TR
UE),]
  p<-predict(Model_M_SigFDR, as.data.frame(n));
  trt<-as.numeric(p)
  names(trt)<-names(p)
  pree <- Testing_age[mgsub( c(".1",".2",".3",".4",".5"), c("","","","","",""), x = names(trt), fixed = T
RUE)]
  ss[i] <-mean(abs(pree - trt))
}


t.test(ss)

# Plot hist of MAD estimates

par(mar=c(5, 6, 4, 2))

hist(ss,
     main="MAD estimation by bootstrap analysis",
     cex.main=2.5,
     xlab="MAD values",
     border="blue",
     col="pink",
     las=1,
     cex.axis=1,
     cex.lab=2,
     breaks=50,
     prob = TRUE)
lines(density(na.omit(ss)), lwd=4)

Model_Hwan_M <- lm(age~cg18384097+cg00481951+cg19671120+cg14361627+cg08928145
+cg12757011+cg07547549,data= as.data.frame(Train_data_transp_M_age))

summary(Model_Hwan_M)

## The prediction accuracy (MAD value) of this model is equal:

mean(abs(residuals(Model_Hwan_M)))


#Residual sum of squares:

RSS <- sum(abs(Model_Hwan_M$residuals))

#Mean squared error:
```

```
MSE <- RSS / length(Model_Hwan_M$residuals)

# Root MSE:

RMSE <- sqrt(MSE)
print(RMSE)

# Predicting the age

mean(abs(as.integer(gsub("age:","",as.character(GSE99029pheno[,11]))) - predict(Model_H
wan_M, as.data.frame(t(NormalGSE99029_M)))))

Predicted_Hwan_M_Sig <- predict(Model_Hwan_M, as.data.frame(t(NormalGSE99029_M)))

# Ploting predicted vs actual
ggplot(as.data.frame(t(NormalGSE99029_M)), aes(x = Testing_age, y = Predicted_Hwan_M_
Sig, color="Saliva")) + geom_point(size=4.5,alpha=0.9, color="orange")+ xlab("Actual Age")
+ylab("Predicted Age")+labs(color="Tissue Type",face="bold",size=3)+ geom_abline(slope
=1, intercept=0, size=1, color="black")+theme(plot.title = element_text(size=21,face="bold"))+
theme(axis.title.x = element_text(size=18,face="bold"))+theme(axis.title.y=element_text(size=
18,face="bold"))+coord_fixed(ratio=3/4) +xlim(15, 100) +ylim(15, 100)

cor.test(Testing_age , Predicted_Hwan_M_Sig)

Model_Hwan_beta <- lm(age~cg18384097+cg00481951+cg19671120+cg14361627+cg089281
45+cg12757011+cg07547549,data= as.data.frame(Train_data_transp))

summary(Model_Hwan_beta)

## The prediction accuracy (MAD value) of this model is equal:

mean(abs(residuals(Model_Hwan_beta)))

#Residual sum of squares:

RSS <- sum(abs(Model_Hwan_beta$residuals))

#Mean squared error:

MSE <- RSS / length(Model_Hwan_beta$residuals)

# Root MSE:

RMSE <- sqrt(MSE)
print(RMSE)

mean(abs(as.integer(gsub("age:","",as.character(GSE99029pheno[,11]))) - predict(Model_H
wan_beta, as.data.frame(t(NormalGSE99029))))
```

Appendix C2: R codes used in Chapter 5

This section provides the R codes used in Chapter 5, for the identification of blood-specific age-related DNA methylation markers on the Illumina MethylationEPIC® array

```
library(knitr)
library(ggplot2)
library(wateRmelon)
library(GEOquery)
library(modes)
library(impute)
library(minfi)
library(ChAMP)
library(RColorBrewer)
library("IlluminaHumanMethylationEPICmanifest")


#GSE103189

dataDirectory <- "/Users/husainalsaleh/GSE103189"
rgSet <- read.metharray.exp(dataDirectory)
rgSet
sampleNames(rgSet) <- substr(sampleNames(rgSet),1,10)

GSE103189 <- getGEO("GSE103189",getGPL = FALSE,AnnotGPL = FALSE,GSEMatrix = TR
UE)
pD.all <- pData(GSE103189[[1]])
rm(GSE103189)
data.frame(Samples=pD.all$geo_accession,Cell_type=pD.all$characteristics_ch1.1,Sex=pD.all
$`Sex:ch1`)

pD.all <- data.frame(Sample=pD.all$geo_accession,Age=rep(0,16),Sex= pD.all$`Sex:ch1`,Sour
ce=pD.all$`sample type:ch1`)
pD_GSE103189 <-pD.all[which(pD.all$Source == "WholeBloodDNA"),]
rownames(pD_GSE103189) <- pD_GSE103189$Sample
pD_GSE103189

rgSet <- rgSet[,which(pD.all$Source == "WholeBloodDNA")]


test_match_order <- function(x,y) {

  if (all(x==y)) print('Perfect match in same order')

  if (!all(x==y) && all(sort(x)==sort(y))) print('Perfect match in wrong order')

  if (!all(x==y) && !all(sort(x)==sort(y))) print('No match')
}

test_match_order(sampleNames(rgSet), pD_GSE103189$Sample)
```

```
pData(rgSet)[,1:3] <- pD_GSE103189[,1:3]
rgSet

#  Poor quality samples
which(colMeans(detP) > 0.05)
rm(detP)

# QC plot based on median Meth and Unmeth signals
rgSetGenoMethSet <- mapToGenome(rgSet)
rgSetQC <-getQC(rgSetGenoMethSet)
plotQC(rgSetQC, badSampleCutoff = 10.5)
rm(rgSetGenoMethSet)
rm(rgSetQC)

mSetSq <- preprocessQuantile(rgSet, removeBadSamples = TRUE,badSampleCutoff = 10.5)

mSetRaw <- preprocessRaw(rgSet)

par(mfrow=c(1,2))
densityPlot(mSetRaw,main="Raw", legend=FALSE,pal = "red")

densityPlot(getBeta(mSetSq),
            main="Normalised", legend=FALSE,pal = "red")
rm(mSetRaw)

# Box plot for the blood samples
rgSetNormBeta <- getBeta(mSetSq)

boxplot(rgSetNormBeta[,1:ncol(rgSetNormBeta)], notch=TRUE,las=2, main="Distribution of B
eta-values in Blood Samples", cex.axis = 0.7)

# Singular Value Decomposition
rgSetNormBeta <- getBeta(mSetSq)
Row_Mean<-rowMeans(rgSetNormBeta)
rgSetNormBeta_Centered<-rgSetNormBeta - Row_Mean

svd1=svd(rgSetNormBeta_Centered)

plot(svd1$d^2/sum(svd1$d^2),main="The Precent Variance Explained by Singular Values
",
     ylab="Precent Explained", col="blue", cex=2, pch=19, cex.main=1.5, cex.lab=1.4,cex.axis=
.9,
     xlab="No. of Singular values")

ggplot(as.data.frame(svd1$v), aes(svd1$v[,1],svd1$v[,3],color=colData(mSetSq)[1:8,3])) +
  geom_point(size=7,alpha=0.7)+ ggtitle("Singular Value Decomposition (SVD)") + xlab("S
V1") +
  ylab("SV2")+labs(color="Sex",face="bold",size=25)+ theme(plot.title = element_text(size=2
1,face="bold"))+
  theme(axis.title.x = element_text(size=15,face="bold"))+theme(axis.title.y=element_text(size
=15,
```

```
 face="bold"))

rm(Row_Mean)
rm(rgSetNormBeta)
rm(rgSetNormBeta_Centered)
rm(svd1)


# preparing the data
# reading IDAT files using minfi package
rgSet <- read.metharray.exp(dataDirectory)
rgSet
sampleNames(rgSet) <- substr(sampleNames(rgSet),1,10)

rgSet <- rgSet[,sampleNames(mSetSq)]

cell_counts <- estimateCellCounts(rgSet, compositeCellType = "Blood",processMethod = "aut
o")
head(cell_counts)
colData(mSetSq) <- DataFrame(Sample=pD_GSE103189$Sample,Age=pD_GSE103189$Age,
Sex=pD_GSE103189$Sex,cell_counts)
colnames(mSetSq) <- sampleNames(rgSet)
rgSet_GSE103189 <-dropLociWithSnps(mSetSq)
pD_GSE103189_age <- pD_GSE103189[colnames(rgSet_GSE103189),2]

 nrow(mSetSq) - nrow(rgSet_GSE103189)
 rm(mSetSq)
 rm(rgSet)
 rm(cell_counts)

#GSE123914

dataDirectory <- "/Users/husainalsaleh/GSE123914/GSE123914_RAW"

# reading IDAT files using minfi package
rgSet <- read.metharray.exp(dataDirectory)
rgSet

sampleNames(rgSet) <- substr(sampleNames(rgSet),1,10)

GSE123914 <- getGEO("GSE123914",getGPL = FALSE,AnnotGPL = FALSE,GSEMatrix = TR
UE)
pD.all <- pData(GSE123914[[1]])
rm(GSE123914)
pD.all$`age:ch1`[which(pD.all$characteristics_ch1.1 == "year of collection: 2014")] <- as.inte
ger (pD.all$`age:ch1`)[which(pD.all$characteristics_ch1.1 == "year of collection: 2014")] +1

pD_GSE123914<-data.frame(Sample=rownames(pD.all),Age=pD.all$`age:ch1`,Sex=rep("Fe
male",69))
rm(pD.all)

test_match_order(sampleNames(rgSet), pD_GSE123914$Sample)
```

```
colData(rgSet)[,1:3] <- DataFrame(pD_GSE123914)
rgSet

which(colMeans(detP) > 0.05)
rm(detP)

rgSetGenoMethSet <- mapToGenome(rgSet)
rgSetQC <-getQC(rgSetGenoMethSet)
plotQC(rgSetQC, badSampleCutoff = 10.5)
rm(rgSetGenoMethSet)
rm(rgSetQC)

mSetSq <- preprocessQuantile(rgSet, removeBadSamples = TRUE,badSampleCutoff = 10.5)

mSetRaw <- preprocessRaw(rgSet)

par(mfrow=c(1,2))
densityPlot(mSetRaw,main="Raw", legend=FALSE,pal = "red")

densityPlot(getBeta(mSetSq),
            main="Normalized", legend=FALSE,pal = "red")
rm(mSetRaw)

rgSetNormBeta <- getBeta(mSetSq)

boxplot(rgSetNormBeta[,1:ncol(rgSetNormBeta)], notch=TRUE,las=2, main="Distribution of B
eta-values in Blood Samples", cex.axis = 0.7)

Row_Mean<-rowMeans(rgSetNormBeta)
rgSetNormBeta_Centered<-rgSetNormBeta - Row_Mean

svd1=svd(rgSetNormBeta_Centered)

plot(svd1$d^2/sum(svd1$d^2),main="The Precent Variance Explained by Singular Values
",
     ylab="Precent Explained", col="blue", cex=2, pch=19, cex.main=1.5, cex.lab=1.4,cex.axis=
.9,
     xlab="No. of Singular values")

ggplot(as.data.frame(svd1$v), aes(svd1$v[,1],svd1$v[,3],color=colData(mSetSq)[1:69,3])) +
  geom_point(size=7,alpha=0.7)+ ggtitle("Singular Value Decomposition (SVD)") + xlab("S
V1") +
  ylab("SV2")+labs(color="Sex",face="bold",size=25)+ theme(plot.title = element_text(size=2
1,face="bold"))+
  theme(axis.title.x = element_text(size=15,face="bold"))+theme(axis.title.y=element_text(size
=15,
  face="bold"))

rm(Row_Mean)
rm(rgSetNormBeta)
rm(rgSetNormBeta_Centered)
rm(svd1)
```

```
rgSet <- read.metharray.exp(dataDirectory)
rgSet
sampleNames(rgSet) <- substr(sampleNames(rgSet),1,10)

rgSet <- rgSet[,sampleNames(mSetSq)]

cell_counts <- estimateCellCounts(rgSet, compositeCellType = "Blood",processMethod = "auto")
head(cell_counts)
colData(mSetSq) <- DataFrame(Sample=pD_GSE123914$Sample,Age=pD_GSE123914$Age,Sex=pD_GSE123914$Sex,cell_counts)
colnames(mSetSq) <- sampleNames(rgSet)

rgSet_GSE123914 <-dropLociWithSnps(mSetSq)
rownames(pD_GSE123914) <- pD_GSE123914$Sample

pD_GSE123914_age <- as.numeric(levels(pD_GSE123914$Age))[pD_GSE123914$Age]

# the number of probes that has been dropped is:
 nrow(mSetSq) - nrow(rgSet_GSE123914)
 rm(mSetSq)
 rm(rgSet)
 rm(cell_counts)

#GSE116339

dataDirectory <- "/Users/husainalsaleh/GSE116339/GSE116339_RAW"

rgSet3_1 <- read.metharray.exp(dataDirectory)
rgSet3_1

sampleNames(rgSet3_1) <- substr(sampleNames(rgSet3_1),1,10)

# extracting meta data
GSE116339 <- getGEO("GSE116339",getGPL = FALSE,AnnotGPL = FALSE,GSEMatrix = TRUE)
pD.all <- pData(GSE116339[[1]])
rm(GSE116339)
pD.all_1<-pD.all[1:100,]

gender_GSE116339 <- data.frame(Sample=pD.all$geo_accession, Gender=pD.all$`gender:ch1`)
gender_GSE116339[,1] <- as.character(gender_GSE116339[,1])

pD_1 <- data.frame(Sample=pD.all_1$geo_accession,Age=pD.all_1$`age:ch1`, Sex=pD.all_1$`gender:ch1`)
rm(pD.all_1)

test_match_order(sampleNames(rgSet3_1), pD_1$Sample)

colData(rgSet3_1)[,1:3] <- DataFrame(pD_1)
rgSet3_1
```

```
which(colMeans(detP) > 0.05)
rm(detP)

rgSetGenoMethSet <- mapToGenome(rgSet3_1)
rgSetQC <-getQC(rgSetGenoMethSet)
plotQC(rgSetQC, badSampleCutoff = 10.5)

mSetSq3_1 <- preprocessQuantile(rgSet3_1, removeBadSamples = TRUE,badSampleCutoff
= 10.5)

mSetRaw3 <- preprocessRaw(rgSet3_1)

# visualise what the data looks like before and after normalisation
par(mfrow=c(1,2))
densityPlot(mSetRaw3,main="Raw", legend=FALSE,pal = "red")

densityPlot(getBeta(mSetSq3_1),
            main="Normalized", legend=FALSE,pal = "red")
rm(mSetRaw3)

mSetSq3_1 <- mSetSq3_1[,-21]


# preparing the data
# reading IDAT files using minfi package
rgSet <- read.metharray.exp(dataDirectory)
rgSet
sampleNames(rgSet) <- substr(sampleNames(rgSet),1,10)

rgSet <- rgSet[,sampleNames(mSetSq3_1)]

cell_counts <- estimateCellCounts(rgSet, compositeCellType = "Blood",processMethod = "aut
o")
head(cell_counts)
colData(mSetSq3_1) <- DataFrame(Sample=colData(mSetSq3_1)[1],Age=colData(mSetSq3_
1)[2],Sex=colData(mSetSq3_1)[3],cell_counts)
colnames(mSetSq3_1) <- sampleNames(rgSet)
rm(rgSet)


rgSetNormSnpFree3_1 <-dropLociWithSnps(mSetSq3_1)

# the number of probes that has been dropped is:
 nrow(mSetSq3_1) - nrow(rgSetNormSnpFree3_1)
 rm(rgSet3_1)
 rm(mSetSq3_1)
 rm(pD_1)

# preparing the data for the regression analysis
GSE116339_regression <- data.frame(PBB=as.numeric(pD.all$`ln(totalpbb):ch1`[1:673]),Age=
as.numeric(pD.all$`age:ch1`[1:673]))

# simple linear regression
```

```
summary(lm(PBB~Age, data = GSE116339_regression))

library( broom )
write.csv( tidy( lm(PBB~Age, data = GSE116339_regression) ) , "coefss.csv" )
write.csv( glance( lm(PBB~Age, data = GSE116339_regression)) , "ann.csv" )

rm(pD.all)
rm(GSE116339_regression)

# preparing the data
all_cell_counts <-rbind(colData(rgSet_GSE103189)[,4:9],colData(rgSet_GSE123914)[,4:9],col
Data(rgSetNormSnpFree3_1)[,4:9],colData(rgSetNormSnpFree3_2)[,4:9],colData(rgSetNormS
npFree3_3)[,4:9],colData(rgSetNormSnpFree3_4)[,4:9],colData(rgSetNormSnpFree3_5)[,4:9],c
olData(rgSetNormSnpFree3_6)[,4:9],colData(rgSetNormSnpFree3_7)[,4:9])

# storing the chronological ages in a vector which will be used for the correlation test

age <- c(pD_GSE103189_age, pD_GSE123914_age, pD_GSE116339_age)

counts_age <- cbind(age,all_cell_counts)

hist(age,
     main="",
     xlab="Chronological Age", ylab="Sample Size",
     freq=TRUE, border="black",cex.lab=1.2,cex.main=1, font=2, font.lab=2,col="cornsilk", xlim
=c(0,100),ylim=c(0,300))
text(x=20,y=180,label=paste("Age range",min(age), "-",as.integer(max(age))),offset = 0.1, fon
t=2)
text(x=20,y=160,label=paste("Median",median(age)),offset = 0.1, font=2)
text(x=20,y=140,label=paste("Female Prop. 0.68"),offset = 0.1, font=2)

plot(counts_age$age,counts_age$CD8T, xlab="Age", ylab="CD8+ Composition", pch=20, ce
x.lab=1.3, font.lab=2, col="orange")
lines(lowess(counts_age$CD8T~counts_age$age, f=2/3),col="red", lwd=3)
text(x=10,y=.3,"rho = - 0.35" , font=2)

plot(counts_age$age,counts_age$CD4T, xlab="Age", ylab="CD4+ Composition", pch=20, ce
x.lab=1.3, font.lab=2, col="green")
lines(lowess(counts_age$CD4T~counts_age$age, f=2/3),col="red", lwd=3)
text(x=10,y=.3,"rho = - 0.19" , font=2)

plot(counts_age$age,counts_age$NK,xlab="", ylab="NK Composition", pch=20, cex.lab=1.3, f
ont.lab=2, col="blue")
lines(lowess(counts_age$NK~counts_age$age, f=2/3),col="red", lwd=3)
text(x=10,y=.2,"rho = 0.32" , font=2)

plot(counts_age$age,counts_age$Bcell,xlab="", ylab="B-cell Composition", pch=20, cex.lab=
1.3, font.lab=2)
lines(lowess(counts_age$Bcell~counts_age$age, f=2/3),col="red", lwd=3)
text(x=10,y=.2,"rho = -0.12" , font=2)
```

```
plot(counts_age$age,counts_age$Mono,xlab="", ylab="Mono Composition", pch=20, cex.lab=
1.3, font.lab=2, col=22)
lines(lowess(counts_age$Mono~counts_age$age, f=2/3),col="red", lwd=3)
text(x=10,y=.25,"rho = 0.19" , font=2)

plot(counts_age$age,counts_age$Gran,xlab="", ylab="Gran Composition", pch=20, cex.lab=1
.3, font.lab=2, col="gray45")
lines(lowess(counts_age$Gran~counts_age$age, f=2/3),col="red", lwd=3)
text(x=10,y=.8,"rho = 0.09" , font=2)

# using simple linear regression to test whether the cell type proportions changes with age
a<-lm(age~., data = counts_age)

library( broom )
write.csv( tidy( a ) , "coefs.csv" )
write.csv( glance( a ) , "an.csv" )

# extracting DNAm M-values from the data sets
GSE103189M <- getM(rgSet_GSE103189)
GSE123914M <- getM(rgSet_GSE123914)
rgSet3_1M <- getM(rgSetNormSnpFree3_1)
rgSet3_2M <- getM(rgSetNormSnpFree3_2)
rgSet3_3M <- getM(rgSetNormSnpFree3_3)
rgSet3_4M <- getM(rgSetNormSnpFree3_4)
rgSet3_5M <- getM(rgSetNormSnpFree3_5)
rgSet3_6M <- getM(rgSetNormSnpFree3_6)
rgSet3_7M <- getM(rgSetNormSnpFree3_7)

# combining the data sets
comb_data <- cbind(GSE103189M,GSE123914M,rgSet3_1M,rgSet3_2M,rgSet3_3M,rgSet3_4
M,rgSet3_5M,rgSet3_6M,rgSet3_7M)

# Get annotation EPIC
library("IlluminaHumanMethylationEPICanno.ilm10b4.hg19")
data(Locations)
data(Other)
data(Islands.UCSC)

# removing sex linked CpG probes
xy_probes <- rownames(Locations)[which(Locations$chr == "chrY" | Locations$chr == "chrX")
]
comb_data <- comb_data[-which(rownames(comb_data) %in% xy_probes == "TRUE"),]

Train_data=rbind(comb_data,age)
Train_data_transp = t(Train_data)
rm(Train_data)
rm(comb_data)

Cor.pvalues <-  vector()
for (i in 1:816126) {
 Cor.pvalues[i]<-cor.test(Train_data_transp[,816127], Train_data_transp[,i], method = "spearm
an", exact = FALSE)$p.value
```

```
}
Cor.p<-unlist(Cor.pvalues)

library(brainwaver)
pvalue.thresh<-compute.FDR(Cor.p,0.05)

# preparing train data with only probes passed FDR0.05 and then calculate the R-square
corTrain_data_FDR=Train_data_transp[,which(Cor.p <= pvalue.thresh)]
corTrain_data_FDR_age <-cbind(corTrain_data_FDR, Train_data_transp["age"] )
colnames(corTrain_data_FDR_age)[ncol(corTrain_data_FDR_age)] <-"age"
rm(corTrain_data_FDR)

# script to calculate rho
Cor.rvaluesFDR <-  vector()
for (i in 1:(ncol(corTrain_data_FDR_age)-1)) {
 Cor.rvaluesFDR[i]<-cor(corTrain_data_FDR_age[,"age"], corTrain_data_FDR_age[,i], method
= "spearman")
}
Cor.rvaluesFDR<-unlist(Cor.rvaluesFDR)

length(which(Cor.rvaluesFDR >= 0.5))
length(which(Cor.rvaluesFDR <= - 0.5))
length(which(Cor.rvaluesFDR > 0.6)) + length(which(Cor.rvaluesFDR < - 0.6))

# Manhattan plot
library(qqman)
manh_pval <-data.frame(CpG=colnames(Train_data_transp)[-816127],CHR=Locations[colna
mes(Train_data_transp)[-816127],1],BP=Locations[colnames(Train_data_transp)[-816127],2],
P=Cor.p)
manh_pval[,2] <- substr(manh_pval[,2],4,5)
manh_pval[,2] <- as.integer(manh_pval[,2])

manhattan(manh_pval, snp = "CpG", suggestiveline = F, genomewideline = -log10(4.55016e-7
5),main = "Manhattan Plot", cex = 0.5, cex.axis = 1,col = c("blue4", "orange3"))

probes_27k <- rownames(Other)[which(Other$Methyl27_Loci == TRUE )]
probes_450k <- rownames(Other)[which(Other$Methyl450_Loci == TRUE )]
probes_EPIC <- rownames(Other)[which(Other$Methyl27_Loci == "" & Other$Methyl450_Loci
== "")]

sig_probes_27k <- colnames(corTrain_data_FDR_age)[which(colnames(corTrain_data_FDR_
age) %in% probes_27k == TRUE)]
cor_probes_27k <- Cor.rvaluesFDR[which(colnames(corTrain_data_FDR_age) %in% probes_
27k == TRUE)]

sig_probes_450k <- colnames(corTrain_data_FDR_age)[which(colnames(corTrain_data_FDR
_age) %in% probes_450k == TRUE)]
cor_probes_450k <- Cor.rvaluesFDR[which(colnames(corTrain_data_FDR_age) %in% probes
_450k == TRUE)]

sig_probes_epic <- colnames(corTrain_data_FDR_age)[which(colnames(corTrain_data_FDR
_age) %in% probes_EPIC == TRUE)]
cor_probes_epic <- Cor.rvaluesFDR[which(colnames(corTrain_data_FDR_age) %in% probes_
```

```
EPIC == TRUE)]

ooo <- c(sig_probes_27k,sig_probes_450k,sig_probes_epic)
oioi <- c(rep("27k",length(sig_probes_27k)), rep("450k", length(sig_probes_450k)), rep("EPI
C",length(sig_probes_epic)))

sig_arrayType <- data.frame(Cor=c(cor_probes_27k,cor_probes_450k,cor_probes_epic), Array
=oioi)
rm(ooo)
rm(oioi)

ggplot(sig_arrayType) + geom_density(aes(x = abs(Cor), color = Array), size=1) + xlim(0.6,0.
8)+ylim(0,65) + xlab("abs(rho)") +ylab("Density")+ ggtitle(paste("AR CpG probes"))+theme
(plot.title = element_text(size=20,face="bold")) +theme(axis.title.x = element_text(size=15,fac
e="bold"))+theme(axis.title.y=element_text(size=15,face="bold"))+ theme(plot.title = element
_text(hjust = 0.5))

# the ID of CpG probes with correlation coefficient above 0.6
colnames(corTrain_data_FDR_age)[which(Cor.rvaluesFDR > 0.6 | Cor.rvaluesFDR < - 0.6)]

# creating data frame contianing detailed annotation about these age related CpG sites
sig_probes_anno <- DataFrame(Other[colnames(corTrain_data_FDR_age)[which(Cor.rvalues
FDR > 0.6 | Cor.rvaluesFDR < - 0.6)] ,c(3,5,7,25,26)], Locations[colnames(corTrain_data_FDR
_age)[which(Cor.rvaluesFDR > 0.6 | Cor.rvaluesFDR < - 0.6)],],Cor.eff=Cor.rvaluesFDR[which(
colnames(corTrain_data_FDR_age) %in% colnames(corTrain_data_FDR_age)[which(Cor.rval
uesFDR > 0.6 | Cor.rvaluesFDR < - 0.6)]  == TRUE)])

# creating data frame contianing detailed annotation about the novel age related CpG sites foun
d only on EPIC array
novel_probes_anno <- sig_probes_anno[rownames(sig_probes_anno)[which(sig_probes_anno
$Methyl27_Loci == "" & sig_probes_anno$Methyl450_Loci == "")],]

novel_probes_anno

# Heatmap

library(ComplexHeatmap)

trainHeat<-Train_data_transp[,c(rownames(novel_probes_anno),"age")]
rownames(trainHeat)<-c(1:754)
trainHeat <- M2Beta(trainHeat)
trainHeat[,"age"] <- Train_data_transp[,"age"]

zd<-trainHeat[order(trainHeat[,22] ),]
aa<-zd[,22]
zdd<-zd[,-22]
zzd<-scale(zdd) # to convert to z-score scale fucntion uses (value) - (its mean) / (its sd)


ha = HeatmapAnnotation(df = data.frame(age = aa),annotation_legend_param = list(age = lis
t(title = "Age",color_bar="continuous", legend_height = unit(4, "cm"),title_gp = gpar(fontsize =
10,fontface = "bold"))))
```

```
hh=Heatmap(t(zzd),km=2,row_dend_width = unit(2,"cm"),heatmap_legend_param = list(color_
bar = "continuous"),name = "Z-score",column_title = "Samples",column_title_side = "bottom"
,column_title_gp = gpar(fontsize = 15, fontface = "bold"),row_title_gp = gpar(fontsize = 13, fontf
ace = "bold"), cluster_columns = FALSE,top_annotation = ha)
draw(hh,row_title = "CpG markers",row_title_gp = gpar(fontsize = 22, fontface = "bold"))

# function for plotting ggplots side by side

multiplot <- function(..., plotlist=NULL, file, cols=1, layout=NULL) {
  library(grid)

  # Make a list from the ... arguments and plotlist
  plots <- c(list(...), plotlist)

  numPlots = length(plots)

  # If layout is NULL, then use 'cols' to determine layout
  if (is.null(layout)) {
    # Make the panel
    # ncol: Number of columns of plots
    # nrow: Number of rows needed, calculated from # of cols
    layout <- matrix(seq(1, cols * ceiling(numPlots/cols)),
                     ncol = cols, nrow = ceiling(numPlots/cols))
  }

  if (numPlots==1) {
    print(plots[[1]])

  } else {
    # Set up the page
    grid.newpage()
    pushViewport(viewport(layout = grid.layout(nrow(layout), ncol(layout))))

    # Make each plot, in the correct location
    for (i in 1:numPlots) {
      # Get the i,j matrix positions of the regions that contain this subplot
      matchidx <- as.data.frame(which(layout == i, arr.ind = TRUE))

      print(plots[[i]], vp = viewport(layout.pos.row = matchidx$row,
                                      layout.pos.col = matchidx$col))
    }
  }
}


sig1 = as.matrix(Train_data_transp[,c("cg16867657")])
sig1 <- M2Beta(sig1)
sig1<- data.frame(sig1,Train_data_transp[,"age"])
colnames(sig1)[c(1,2)] <-c("cg16867657","age")

ggplot(data = as.data.frame(sig1), aes(x = sig1[,2], y = sig1[,1])) + geom_point(size=1.5) + ge
om_smooth(method=lm, fill="pink",se=TRUE) + xlab("Chronological Age") +ylab("Methylati
```

```
on level") +ggtitle(paste(colnames(sig1)[1],"450k probe \n (ELOVL2 gene)"))+theme(plot.ti
tle = element_text(size=20,face="bold")) +theme(axis.title.x = element_text(size=15,face="bol
d"))+theme(axis.title.y=element_text(size=15,face="bold")) + annotate("text", x = 20, y=0.86, l
abel ="Spearman's rho = 0.82",fontface =2,size=5) +theme(plot.title = element_text(hjust = 0.
5))

sig2 = as.matrix(Train_data_transp[,c("cg21572722")])
sig2 <- M2Beta(sig2)
sig2<- data.frame(sig2,Train_data_transp["age"])
colnames(sig2)[c(1,2)] <-c("cg21572722","age")

ggplot(data = as.data.frame(sig2), aes(x = sig2[,2], y = sig2[,1])) + geom_point(size=1.5) + ge
om_smooth(method=lm, fill="pink",se=TRUE) + xlab("Chronological Age") +ylab("Methylati
on level") +ggtitle(paste(colnames(sig2)[1],"450k probe \n (ELOVL2 gene)"))+theme(plot.ti
tle = element_text(size=20,face="bold")) +theme(axis.title.x = element_text(size=15,face="bol
d"))+theme(axis.title.y=element_text(size=15,face="bold")) + annotate("text", x = 20, y=0.6, la
bel ="Spearman's rho = 0.76",fontface =2,size=5)+ theme(plot.title = element_text(hjust = 0.5
))

sig3 = as.matrix(Train_data_transp[,c("cg17268658")])
sig3 <- M2Beta(sig3)
sig3<- data.frame(sig3,Train_data_transp["age"])
colnames(sig3)[c(1,2)] <-c("cg17268658","age")

ggplot(data = as.data.frame(sig3), aes(x = sig3[,2], y = sig3[,1])) + geom_point(size=1.5) + ge
om_smooth(method=lm, fill="pink",se=TRUE) + xlab("Chronological Age") +ylab("Methylati
on level") +ggtitle(paste(colnames(sig3)[1]," EPIC probe \n (FHL2 gene)"))+theme(plot.title
= element_text(size=20,face="bold")) +theme(axis.title.x = element_text(size=15,face="bold")
)+theme(axis.title.y=element_text(size=15,face="bold")) + annotate("text", x = 20, y=0.7, label
="Spearman's rho =  0.76",fontface =2,size=5)+ theme(plot.title = element_text(hjust = 0.5))

# plotting DNAm level at the top 4 (in terms of correlation coefficient) novel are related CpG sites
found exclusively on the EPIC array
sig1 = as.matrix(Train_data_transp[,c("cg17268658")])
sig1 <- M2Beta(sig1)
sig1<- data.frame(sig1,Train_data_transp["age"])
colnames(sig1)[c(1,2)] <-c("cg17268658","age")

p1= ggplot(data = as.data.frame(sig1), aes(x = sig1[,2], y = sig1[,1])) + geom_point(size=1.5)
+ geom_smooth(method=lm, fill="pink",se=TRUE) + xlab("") +ylab("Methylation level") +ggti
tle(paste(colnames(sig1)[1],"(FHL2 gene)"))+theme(plot.title = element_text(size=20,face="b
old")) +theme(axis.title.x = element_text(size=15,face="bold"))+theme(axis.title.y=element_te
xt(size=15,face="bold")) + annotate("text", x = 20, y=0.7, label ="Spearman's rho = 0.76",fon
tface =2,size=5) +theme(plot.title = element_text(hjust = 0.5))

sig2 = as.matrix(Train_data_transp[,c("cg24866418")])
sig2 <- M2Beta(sig2)
sig2<- data.frame(sig2,Train_data_transp["age"])
colnames(sig2)[c(1,2)] <-c("cg24866418","age")
```

```
p2= ggplot(data = as.data.frame(sig2), aes(x = sig2[,2], y = sig2[,1])) + geom_point(size=1.5)
+ geom_smooth(method=lm, fill="pink",se=TRUE) + xlab("") +ylab("Methylation level") +ggti
tle(paste(colnames(sig2)[1],"(LHFPL4 gene)"))+theme(plot.title = element_text(size=20,face
="bold")) +theme(axis.title.x = element_text(size=15,face="bold"))+theme(axis.title.y=element
_text(size=15,face="bold")) + annotate("text", x = 20, y=0.5, label ="Spearman's rho = 0.66",
fontface =2,size=5)+ theme(plot.title = element_text(hjust = 0.5))


 multiplot(p1,p2, cols = 2)

sig3 = as.matrix(Train_data_transp[,c("cg07323488")])
sig3 <- M2Beta(sig3)
sig3<- data.frame(sig3,Train_data_transp[,"age"])
colnames(sig3)[c(1,2)] <-c("cg07323488","age")

p1= ggplot(data = as.data.frame(sig3), aes(x = sig3[,2], y = sig3[,1])) + geom_point(size=1.5)
+ geom_smooth(method=lm, fill="pink",se=TRUE) + xlab("Chronological Age") +ylab("Meth
ylation level") +ggtitle(paste(colnames(sig3)[1],"(EGFEM1P gene)"))+theme(plot.title = ele
ment_text(size=20,face="bold")) +theme(axis.title.x = element_text(size=15,face="bold"))+the
me(axis.title.y=element_text(size=15,face="bold")) + annotate("text", x = 60, y=0.7, label ="S
pearman's rho = - 0.69",fontface =2,size=5)+ theme(plot.title = element_text(hjust = 0.5))


sig4 = as.matrix(Train_data_transp[,c("cg13552692")])
sig4 <- M2Beta(sig4)
sig4<- data.frame(sig4,Train_data_transp[,"age"])
colnames(sig4)[c(1,2)] <-c("cg13552692","age")

p2= ggplot(data = as.data.frame(sig4), aes(x = sig4[,2], y = sig4[,1])) + geom_point(size=1.5)
+ geom_smooth(method=lm, fill="pink",se=TRUE) + xlab("Chronological Age") +ylab("Meth
ylation level") +ggtitle(paste(colnames(sig4)[1],"(CCDC102B gene)"))+theme(plot.title = ele
ment_text(size=20,face="bold")) +theme(axis.title.x = element_text(size=15,face="bold"))+the
me(axis.title.y=element_text(size=15,face="bold")) + annotate("text", x = 60, y=0.7, label ="S
pearman's rho = - 0.67",fontface =2,size=5)+ theme(plot.title = element_text(hjust = 0.5))

 multiplot(p1,p2, cols = 2)

 rm(sig1,sig2,sig3,sig4)

cont_train<-data.frame(counts_age,Train_data_transp[,rownames(novel_probes_anno)])

summary(lm(cg17268658~age, data = as.data.frame(cont_train)))
summary(lm(cg17268658~age+Gran, data = as.data.frame(cont_train)))

summary(lm(cg24866418~age, data = as.data.frame(cont_train)))
summary(lm(cg24866418~age+Gran+Bcell+Mono+CD8T+CD4T+NK, data = as.data.frame(co
nt_train)))

summary(lm(cg13552692~age, data = as.data.frame(cont_train)))
summary(lm(cg13552692~age+CD8T, data = as.data.frame(cont_train)))
```

```
# top 10 AR CpG markers from Garganani 2012

GargnaniMarkers <- c("cg06639320","cg16867657","cg22454769","cg24079702","cg1641
9235","cg21572722","cg24724428","cg16219603","cg12877723")
length(which(colnames(Train_data_transp) %in% GargnaniMarkers == TRUE)) # Only 1 drop
ped after snp filteration
data.frame(ID=GargnaniMarkers[which(GargnaniMarkers %in% colnames(corTrain_data_FD
R_age) == TRUE )],Cor.coeffic=Cor.rvaluesFDR[which(colnames(corTrain_data_FDR_age) %i
n% GargnaniMarkers == TRUE)])
# how many markers were abs > 0.5
length(which(data.frame(ID=GargnaniMarkers[which(GargnaniMarkers %in% colnames(corT
rain_data_FDR_age) == TRUE )],Cor.coeffic=abs(Cor.rvaluesFDR[which(colnames(corTrain_d
ata_FDR_age) %in% GargnaniMarkers == TRUE)]))$Cor.coeffic > 0.5))

# 71 AR CpG markers which were selected by elastic net regression (Hannum et al. 2013)
HannumMarkers <- read.table("HanumMarkers.txt", header = TRUE)
length(which(colnames(Train_data_transp) %in% HannumMarkers$Marker == TRUE)) # 1 ma
rkers from Hannum were dropped from EPIC by SNP filteration and 11 markers were totally drop
ped from EPIC chip.

# the correlation coefficient of Hannum markers resulted in our analysis
data.frame(ID=HannumMarkers$Marker[which(HannumMarkers$Marker %in% colnames(corT
rain_data_FDR_age) == TRUE )],Cor.coeffic=Cor.rvaluesFDR[which(colnames(corTrain_data_
FDR_age) %in% HannumMarkers$Marker == TRUE)])

# how many markers were abs > 0.5
length(which(data.frame(ID=HannumMarkers$Marker[which(HannumMarkers$Marker %in%
colnames(corTrain_data_FDR_age) == TRUE )],Cor.coeffic=abs(Cor.rvaluesFDR[which(colna
mes(corTrain_data_FDR_age) %in% HannumMarkers$Marker == TRUE)]))$Cor.coeffic > 0.5))

# 11 markers were found by Xu et al. (2015) (450K and linear regression test were used to identi
fication)
XuMarkers <- read.table("XuMarkers.txt", header = TRUE)

length(which(colnames(corTrain_data_FDR_age) %in% XuMarkers$ID == TRUE)) # 1 marker
has been dropped from EPIC chip, and 1 marker was dropped by SNP filteration.

# the correlation coefficients resulted in our analysis
data.frame(ID=XuMarkers$ID[which(XuMarkers$ID %in% colnames(corTrain_data_FDR_age)
== TRUE )],Cor.coeffic=Cor.rvaluesFDR[which(colnames(corTrain_data_FDR_age) %in% Xu
Markers$ID == TRUE)])

# 162 markers found by Florath et al. (2014) (the coefficients value were not included in the supp
l materials)
FlorathMarkers <- read.table("FlorathMarkers.txt", header = TRUE)

length(which(colnames(corTrain_data_FDR_age) %in% FlorathMarkers$ID == TRUE)) # 10
markers were dropped from EPIC chip and 2 markers from SNP filteration

# the correlation coefficient from our analysis
data.frame(ID=FlorathMarkers$ID[which(FlorathMarkers$ID %in% colnames(corTrain_data_F
DR_age) == TRUE )],Cor.coeffic=Cor.rvaluesFDR[which(colnames(corTrain_data_FDR_age)
```

```
%in% FlorathMarkers$ID == TRUE)])

# 102 markers were found by Weidner et al. (2014) (from 27K Pearson's test)
WeidnerMarkers <- read.table("WeidnerMarkers.txt", header = TRUE)

length(which(colnames(corTrain_data_FDR_age) %in% WeidnerMarkers$ID == TRUE)) # 1
marker dropped from EPIC and 1 from SNP filteration

# the correlation coefficient from our analysis
ourWeidner<-data.frame(ID=WeidnerMarkers$ID[which(WeidnerMarkers$ID %in% colnames(
corTrain_data_FDR_age) == TRUE )],Cor.coeffic=Cor.rvaluesFDR[which(colnames(corTrain_d
ata_FDR_age) %in% WeidnerMarkers$ID == TRUE)])

## 80% will be training and 20% testing data
smp_size <- floor(0.7 * nrow(Train_data_transp))

## set the seed to make your partition reproducible
set.seed(123)
train_ind <- sample(seq_len(nrow(Train_data_transp)), size = smp_size)

train <- Train_data_transp[train_ind, ]
test <- Train_data_transp[-train_ind, ]

hist(train[,"age"],
     main="Training set \n (n=527)",
     xlab="Age (years)", ylab="Frequency",
     freq=TRUE, border="black",cex.lab=1.2,cex.main=1.3, font=2, font.lab=2,col="gray", ylim=c
(0,200), xlim=c(0,100))

hist(test[,"age"],
     main="Testing set \n (n=227)",
     xlab="Age (years)", ylab="Frequency",
     freq=TRUE, border="black",cex.lab=1.2,cex.main=1.3, font=2, font.lab=2,col="gray", ylim=c
(0,100), xlim=c(0,100))

library(glmnet)

# use 10 fold cross validation to estimate the lambda parameter in the
# training data

glmnet.Training.CV <- cv.glmnet(as.matrix(train[, 1:(ncol(train) -1)]), train[,"age"], nfolds = 10,
                        alpha = 0.5, family = "gaussian")

# The definition of the lambda parameter:
lambda.glmnet.Training <- glmnet.Training.CV$lambda.min

# Fit the elastic net predictor to the training data
glmnet.Training <- glmnet(as.matrix(train[, 1:(ncol(train))-1]), train[, "age"], alpha = 0.5, family
= "gaussian", nlambda = 100)


# predicting age of the samples in the training set
```

```
DNAmAge_Training <- predict(glmnet.Training, as.matrix(train[, 1:(ncol(train)) - 1]), type="res
ponse", s=lambda.glmnet.Training)

# Calculating MAD value
glm_Predicted_Age <- DNAmAge_Training[, 1]
Actual_Age <- train[, "age"]

# The prediction accuracy is
mean(abs(Actual_Age - glm_Predicted_Age))

plot(glmnet.Training.CV)


library(broom)
tidied.cv<-as.data.frame(tidy(glmnet.Training.CV))
tidied.cv

glmnet.Training.CV$lambda.min

# Ploting predicted vs actual
t<-data.frame(Actual_Age,glm_Predicted_Age)
ggplot(as.data.frame(t), aes(x = Actual_Age, y = glm_Predicted_Age, color="Blood")) +
geom_point(size=4.5,alpha=0.9, color="orange")+ xlab("Actual Age")+ylab("Predicted Age"
)+labs(color="Tissue Type",face="bold",size=3)+ geom_abline(slope=1,
intercept=0, size=1, color="black")+theme(plot.title = element_text(size=21,face="bold"))+
theme(axis.title.x = element_text(size=18,face="bold"))+theme(axis.title.y=element_text(size=
18,face="bold"))+coord_fixed(ratio=3/4) +xlim(15, 70) +ylim(15, 70)

cor.test(Actual_Age , glm_Predicted_Age)

# extracting the IDs of the CpG probes that have been selected by elastic net regression and cre
ating data frame containing annotation details:

trainWithIntercept=cbind(rep(0,836), Train_data_transp)
colnames(trainWithIntercept)[1]<-"Intercept"

elastic_probes_anno <- data.frame(Other[colnames(trainWithIntercept)[which(coef(glmnet.Tra
ining.CV, s = 0.2260869) != 0)][-1],c(3,5,7,25,26)], Locations[colnames(trainWithIntercept)[whic
h(coef(glmnet.Training.CV, s = 0.2260869) != 0)][-1],])

# the number of probes that are coming from each BeadChip array
# The number of probes from 27K
length(which(elastic_probes_anno$Methyl27_Loci == "TRUE"))
# The number of probes from 450K
length(which(elastic_probes_anno$Methyl450_Loci == "TRUE"))
# The number of probes from EPIC
length(which(elastic_probes_anno$Methyl450_Loci == "" & elastic_probes_anno$Methyl27_Lo
ci == ""))

rm(trainWithIntercept)
rm(Other)
```

```
rm(Locations)
rm(Islands.UCSC)

# Validating the model using independent data set

# Validating the model with the randomly selected samples
DNAmAge_Testing <- predict(glmnet.Training, as.matrix(test[, 1:(ncol(test)) - 1]), type="respo
nse", s=lambda.glmnet.Training)

# Calculating MAD value
test_Predicted_Age <- DNAmAge_Testing[, 1]
test_Actual_Age <- test[, "age"]

# The prediction accuracy is
mean(abs(test_Actual_Age - test_Predicted_Age))

s <- data.frame(test_Actual_Age,test_Predicted_Age )

ggplot(as.data.frame(s), aes(x = test_Actual_Age, y =test_Predicted_Age))+geom_point(size=
2.8,alpha=0.9)+
                    ggtitle("") + xlab("Actual age") +
                    ylab("Predicted age")+geom_abline(slope=1,intercept=0,size=1,color="red")
+
                    theme(plot.title=element_text(size=21,face="bold"))+
                    theme(axis.title.x=element_text(size=18,face="bold"))+
                    theme(axis.title.y=element_text(size=18,face="bold"))+coord_fixed(ratio=3/4)
+
                    theme(plot.title = element_text(hjust = 0.5))

no_markers <- tidied.cv$nzero[-1]
lamd <- tidied.cv$lambda[-1]
mad_models <- vector()
for (i in 61:98) { DNAge <- predict(glmnet.Training, as.matrix(test[, 1:(ncol(test)) - 1]), type="re
sponse", s=lamd[i])
tes_P_Age <- DNAge[, 1]
mad_models[i] <-mean(abs(test_Actual_Age - tes_P_Age))
rm(DNAge)
}

no_markers1 <- no_markers[-which(duplicated(no_markers) == TRUE)]
mad_models <- mad_models[-which(duplicated(no_markers) == TRUE)]

y <- data.frame(Model=no_markers1, MAD=mad_models)

ggplot(as.data.frame(y), aes(x = Model, y = MAD)) +geom_line(size=1) + xlab("Number of C
pG sites")+
                     ylab("MAD (years)")+labs(face="bold",size=3) +
                    theme(plot.title=element_text(size=21,face="bold"))+
                    theme(axis.title.x=element_text(size=18,face="bold"))+
                    theme(axis.title.y=element_text(size=18,face="bold"))+
                    coord_fixed(ratio=4/0.1)+scale_y_discrete(limits=c(seq(1,12,1)))
```

```
# Simple linear regression code
CoefEsts.forM = matrix(NA, nrow = (ncol(train)-1), ncol = 4) # this is to design matrix for inputs
rownames(CoefEsts.forM) = colnames(train)[1:(ncol(train)-1)]
colnames(CoefEsts.forM) = c("Intercept", "Coef","Pvalue","Rsquare")
for (i in 1:(ncol(train)-1)) { fit = lm(train[,ncol(train)]~train[,i]) ;CoefEsts.forM[i,c(1,2,3)] = summ
ary(fit)[[4]][c(1,2,8)]; CoefEsts.forM[i,4] = summary(fit)[[8]] }

Reg.p <- CoefEsts.forM[,3]
library(brainwaver)

Rpvalue.thresh<-compute.FDR(Reg.p,0.05)

CoefEsts.forM[which(CoefEsts.forM[,4] > 0.6),]

# step-wise regression
library(leaps)
train <- as.data.frame(train)
train <- train[,rownames(CoefEsts.forM[which(CoefEsts.forM[,4] > 0.5),])]
train <- cbind( train, age[train_ind])
colnames(train)[ncol(train)] <- "age"

regfitfull=regsubsets(age~., data = train, nvmax = 10,nbest=1, method = "exhaustive")
reg.summary=summary(regfitfull)
plot(reg.summary$bic, xlab="Number of variables", cex.lab=1.3,ylab="BIC stats",cex=2, col=
"red",pch=20,main=" Bayesian information criterion (BIC) ",font.lab=2,cex.main=2)

abline(v=which.min(reg.summary$bic),  lwd=2, col="blue")

# Based on Bayesian Information Criterion (BIC) algorithm, the best CpG markers combination i
s at
which.min(reg.summary$bic)

# The names of the CpG sites that are included in the best model are
names(coef(regfitfull,(which.min(reg.summary$bic))))[-1]

#Multiple linear regression - CpG markers from the model that has the lowest BIC value
Model_M_Sig <- lm(age~cg18933331+cg10501210+cg06639320+cg24866418+cg16867657+c
g17110586, data = as.data.frame(train))

Predicted_M_Sig <- fitted(Model_M_Sig) # predicted values

k <- cbind(Actual_age=train["age"], sig_predicted_train=Predicted_M_Sig)

ggplot(as.data.frame(k), aes(x = Actual_age, y =sig_predicted_train))+geom_point(size=2.8,al
pha=0.9)+
                    ggtitle("") + xlab("Actual age") +
                    ylab("Predicted age")+geom_abline(slope=1,intercept=0,size=1,color="red")
+
                    theme(plot.title=element_text(size=21,face="bold"))+
                    theme(axis.title.x=element_text(size=18,face="bold"))+
                    theme(axis.title.y=element_text(size=18,face="bold"))+coord_fixed(ratio=3/4)
+
```

```
                    theme(plot.title = element_text(hjust = 0.5)) +
          annotate("text", x = 20, y=80, label =" Training data \n MAD = 4.5 years \n r = 0.9",
fontface=2,size=5)


write.csv( tidy( lm(age~cg18933331+cg10501210+cg06639320+cg24866418+cg16867657+cg
17110586, data = as.data.frame(train)) ) , "train6.csv" )
write.csv( glance( lm(age~cg18933331+cg10501210+cg06639320+cg24866418+cg16867657
+cg17110586, data = as.data.frame(train))) , "ann.csv" )

test <- test[,names(coef(regfitfull,(which.min(reg.summary$bic))))[-1]]
test <- cbind(test, age=age[-train_ind])

sig_predicted_test <- predict(Model_M_Sig, as.data.frame(test))

mean(abs(test["age"] - sig_predicted_test))

l <- data.frame(Actual_age=test["age"], Predicted_age=sig_predicted_test)

ggplot(as.data.frame(l), aes(x = Actual_age, y =sig_predicted_test))+geom_point(size=2.8,alp
ha=0.9)+
                    ggtitle("") + xlab("Actual age") +
                    ylab("Predicted age")+geom_abline(slope=1,intercept=0,size=1,color="red")
+
                    theme(plot.title=element_text(size=21,face="bold"))+
                    theme(axis.title.x=element_text(size=18,face="bold"))+
                    theme(axis.title.y=element_text(size=18,face="bold"))+coord_fixed(ratio=3/4)
+
           annotate("text", x = 20, y=65, label =" Testing data \n MAD = 4.6 years \n r = 0.9",
fontface=2,size=5)

# Bootstrap function
library(mgsub)
ss <-vector()
for (i in 1:10000 ) {
  nx <- test[sample(rownames(test),227, replace = TRUE),]
  px<-predict(Model_M_Sig, as.data.frame(nx))
  trt<-as.numeric(px)
  names(trt)<-names(px)
  pree <- test_Actual_Age[mgsub( string=names(trt), pattern=c(".1",".2",".3",".4",".5",".6"), repl
acement=c("","","","","",""), fixed = TRUE)]
  ss[i] <-mean(abs(pree - trt))
}

t.test(ss)

# Plot hist of MAD estimates

par(mar=c(5, 6, 4, 2))

hist(ss,
    main="MAD estimation by bootstrap analysis",
```

```
    cex.main=2.5,
    xlab="MAD values",
    border="blue",
    col="pink",
    las=1,
    cex.axis=1,
    cex.lab=2,
    breaks=50,
    prob = TRUE)
lines(density(na.omit(ss)), lwd=4)

gender_GSE116339<-gender_GSE116339[which(gender_GSE116339[,1] %in% rownames(Tr
ain_data_transp) == TRUE),]
gender_GSE116339[,2] <- as.character(gender_GSE116339[,2])

gender_data <- c("Male","Male", "Female", "Female", "Male" ,"Female", "Male", "Male", re
p("Female",69))
gender_data <- c(gender_data, gender_GSE116339[,2])
names(gender_data) <- rownames(Train_data_transp)

test_gender <- gender_data[which(names(gender_data) %in% rownames(test) == TRUE)]

# female
mean(abs( test[which(test_gender == "Female"),"age"] - predict(Model_M_Sig, as.data.fram
e(test[which(test_gender == "Female"),])))

# male

mean(abs( test[which(test_gender == "Male"),"age"] - predict(Model_M_Sig, as.data.frame(t
est[which(test_gender == "Male"),])))

# t.test between male and female

absPredicted_male<-abs( test[which(test_gender == "Male"),"age"] - predict(Model_M_Sig, a
s.data.frame(test[which(test_gender == "Male"),]))

absPredicted_female<-abs( test[which(test_gender == "Female"),"age"] - predict(Model_M_S
ig, as.data.frame(test[which(test_gender == "Female"),]))

t.test(absPredicted_female, absPredicted_male)
```

Appendix C3: R codes used for Chapter 6

This section provides the R codes used in Chapter 6, to normalise the DNA methylation profiles assayed on Illumina HumanMethylation27 and HumanMethylation450 based on Horvath's algorithm.

```
library(GEOquery)

# Downloading the data set
exprss <- getGEO("GSE-data",getGPL = FALSE,AnnotGPL = FALSE,GSEMatrix = TRUE)

# Extracting the DNAm data, samples' meta-data, and features annotation
expression_data <- as.data.frame(exprs(exprss[[1]]))
meta_data<-pData(phenoData(exprss[[1]]))
feature_annotation <- pData(featureData(exprss[[1]]))

dat0 <- cbind(rownames(expression_data),expression_data)
colnames(dat0)[1] <-"ProbeID"

probeAnnotation27k=read.csv("datMiniAnnotation27k.csv")
probeAnnotation21kdatMethUsed=read.csv("probeAnnotation21kdatMethUsed.csv")

nSamples=dim(dat0)[[2]]-1
nProbes= dim(dat0)[[1]]
# the following command may not be needed. But it is sometimes useful when you use read.csv.
sql
dat0[,1]= gsub(x=dat0 [,1],pattern="\"",replacement="")
#Create a log file which will be output into your directory
# The code looks a bit complicated because it serves to create a log file (for error checks etc).
# It will automatically create a log file.
file.remove("LogFile.txt")
file.create("LogFile.txt")
DoNotProceed=FALSE
cat(paste( "The methylation data set contains", nSamples, "samples (e.g. arrays) and ", n
Probes, " probes."),file="LogFile.txt")
if (nSamples==0) {DoNotProceed=TRUE; cat(paste( "\n ERROR: There must be a data inp
ut error since there seem to be no samples.\n Make sure that you input a comma delimi
ted file (.csv file)\n that can be read using the R command read.csv.sql . Samples corre
spond to columns in that file  ."), file="LogFile.txt",append=TRUE) }
if (nProbes==0) {DoNotProceed=TRUE; cat(paste( "\n ERROR: There must be a data input
error since there seem to be zero probes.\n Make sure that you input a comma delimite
d file (.csv file)\n that can be read using the R command read.csv.sql  CpGs correspond
to rows.")  , file="LogFile.txt",append=TRUE) }
if (  nSamples > nProbes  ) { cat(paste( "\n MAJOR WARNING: It worries me a lot that ther
e are more samples than CpG probes.\n Make sure that probes correspond to rows and
samples to columns.\n I wonder whether you want to first transpose the data and then r
esubmit them? In any event, I will proceed with the analysis."),file="LogFile.txt",append=T
RUE) }
```

```
if (  is.numeric(dat0[,1]) ) { DoNotProceed=TRUE; cat(paste( "\n Error: The first column doe
s not seem to contain probe identifiers (cg numbers from Illumina) since these entries ar
e numeric values. Make sure that the first column of the file contains probe identifiers su
ch as cg00000292. Instead it contains ", dat0[1:3,1]  ),file="LogFile.txt",append=TRUE)  }
if (  !is.character(dat0[,1]) ) {  cat(paste( "\n Major Warning: The first column does not see
m to contain probe identifiers (cg numbers from Illumina) since these entries are numeri
c values. Make sure that the first column of the file contains CpG probe identifiers such
as cg00000292. Instead it contains ", dat0[1:3,1]  ),file="LogFile.txt",append=TRUE)  }
datout=data.frame(Error=c("Input error. Please check the log file for details","Please read
the instructions carefully."), Comment=c("", "email Steve Horvath."))

if ( ! DoNotProceed ) {
nonNumericColumn=rep(FALSE, dim(dat0)[[2]]-1)
for (i in 2:dim(dat0)[[2]] ){ nonNumericColumn[i-1]=! is.numeric(dat0[,i]) }
if (  sum(nonNumericColumn) >0 ) { cat(paste( "\n MAJOR WARNING: Possible input error.
The following samples contain non-numeric beta values: ", colnames(dat0)[-1][ nonNumeri
cColumn], "\n Hint: Maybe you use the wrong symbols for missing data. Make sure to co
de missing values as NA in the Excel file. To proceed, I will force the entries into numeri
c values but make sure this makes sense.\n" ),file="LogFile.txt",append=TRUE)  }
XchromosomalCpGs=as.character(probeAnnotation27k$Name[probeAnnotation27k$Chr=="X"]
)
selectXchromosome=is.element(dat0[,1], XchromosomalCpGs )
selectXchromosome[is.na(selectXchromosome)]=FALSE
meanXchromosome=rep(NA, dim(dat0)[[2]]-1)
if (   sum(selectXchromosome) >=500 )  {
meanXchromosome= as.numeric(apply( as.matrix(dat0[selectXchromosome,-1]),2,mean,na.r
m=TRUE)) }
if (  sum(is.na(meanXchromosome)) >0 ) { cat(paste( "\n \n Comment: There are lots of mis
sing values for X chromosomal probes for some of the samples. This is not a problem w
hen it comes to estimating age but I cannot predict the gender of these samples.\n " ),fil
e="LogFile.txt",append=TRUE)  }

match1=match(probeAnnotation21kdatMethUsed$Name , dat0[,1])
if  ( sum( is.na(match1))>0 ) {
missingProbes= probeAnnotation21kdatMethUsed$Name[!is.element( probeAnnotation21kdat
MethUsed$Name , dat0[,1])]
DoNotProceed=TRUE; cat(paste( "\n \n Input error: You forgot to include the following ", le
ngth(missingProbes), " CpG probes (or probe names):\n ", paste( missingProbes, sep="",col
lapse=", ")),file="LogFile.txt",append=TRUE)  }}

#STEP 2: Restrict the data to 21k probes and ensure they are numeric
match1=match(probeAnnotation21kdatMethUsed$Name , dat0[,1])
if  ( sum( is.na(match1))>0 ) stop(paste(sum( is.na(match1)), "CpG probes cannot be matc
hed"))
dat1= dat0[match1,]
asnumeric1=function(x) {as.numeric(as.character(x))}
dat1[,-1]=apply(as.matrix(dat1[,-1]),2,asnumeric1)

plot(density(dat1[,2],na.rm=TRUE,from=0,to=1), col="red" ,xlab = "DNA Methylation Level -
Beta values", main="Density plot")
```

```
for (i in 3:ncol(dat1)) {lines(density(dat1[,i], na.rm=TRUE,from=0,to=1),col="red") }


require(RPMM);
library("dynamicTreeCut")
library(WGCNA)

betaEst2=function (y, w, weights)
{
  yobs = !is.na(y)
  if (sum(yobs) <= 1)
    return(c(1, 1))
  y = y[yobs]
  w = w[yobs]
  weights = weights[yobs]
  N = sum(weights * w)
  p = sum(weights * w * y)/N
  v = sum(weights * w * y * y)/N - p * p
  logab = log(c(p, 1 - p)) + log(pmax(1e-06, p * (1 - p)/v -
                                        1))
  if (sum(yobs) == 2)
    return(exp(logab))
  opt = try(optim(logab, betaObjf, ydata = y, wdata = w, weights = weights,
                  method = "Nelder-Mead",control=list(maxit=50) ), silent = TRUE)
  if (inherits(opt, "try-error"))
    return(c(1, 1))
  exp(opt$par)
} # end of function betaEst

blc2=function (Y, w, maxiter = 25, tol = 1e-06, weights = NULL, verbose = TRUE)
{
  Ymn = min(Y[Y > 0], na.rm = TRUE)
  Ymx = max(Y[Y < 1], na.rm = TRUE)
  Y = pmax(Y, Ymn/2)
  Y = pmin(Y, 1 - (1 - Ymx)/2)
  Yobs = !is.na(Y)
  J = dim(Y)[2]
  K = dim(w)[2]
  n = dim(w)[1]
  if (n != dim(Y)[1])
    stop("Dimensions of w and Y do not agree")
  if (is.null(weights))
    weights = rep(1, n)
  mu = a = b = matrix(Inf, K, J)
  crit = Inf
  for (i in 1:maxiter) {
    warn0 = options()$warn
    options(warn = -1)
    eta = apply(weights * w, 2, sum)/sum(weights)
    mu0 = mu
    for (k in 1:K) {
      for (j in 1:J) {
        ab = betaEst2(Y[, j], w[, k], weights)
```

```
      a[k, j] = ab[1]
      b[k, j] = ab[2]
      mu[k, j] = ab[1]/sum(ab)
    }
  }
  ww = array(0, dim = c(n, J, K))
  for (k in 1:K) {
    for (j in 1:J) {
      ww[Yobs[, j], j, k] = dbeta(Y[Yobs[, j], j],
                                  a[k, j], b[k, j], log = TRUE)
    }
  }
  options(warn = warn0)
  w = apply(ww, c(1, 3), sum, na.rm = TRUE)
  wmax = apply(w, 1, max)
  for (k in 1:K) w[, k] = w[, k] - wmax
  w = t(eta * t(exp(w)))
  like = apply(w, 1, sum)
  w = (1/like) * w
  llike = weights * (log(like) + wmax)
  crit = max(abs(mu - mu0))
  if (verbose)
    print(crit)
  if (crit < tol)
    break
 }
 return(list(a = a, b = b, eta = eta, mu = mu, w = w, llike = sum(llike)))
}


# The function BMIQcalibration was created by Steve Horvath by heavily recycling code
# from A. Teschendorff's BMIQ function.
# BMIQ stands for beta mixture quantile normalization.
# Explanation: datM is a data frame with Illumina beta values (rows are samples, colums are Cp
Gs.
# goldstandard is a numeric vector with beta values that is used as gold standard for calibrating t
he columns of datM.
# The length of goldstandard has to equal the number of columns of datM.
# Example code: First we impute missing values.
# library(WGCNA); dimnames1=dimnames(datMeth)
# datMeth= data.frame(t(impute.knn(as.matrix(t(datMeth)))$data))
# dimnames(datMeth)=dimnames1
# gold.mean=as.numeric(apply(datMeth,2,mean,na.rm=TRUE))
#datMethCalibrated=BMIQcalibration(datM=datMeth,goldstandard.beta=gold.mean)

BMIQcalibration=function(datM,goldstandard.beta,nL=3,doH=TRUE,nfit=20000,th1.v=c(0.2,0.7
5),th2.v=NULL,niter=5,tol=0.001,plots=FALSE,calibrateUnitInterval=TRUE){
 if (length(goldstandard.beta) !=dim(datM)[[2]] ) {stop("Error in function arguments length(g
oldstandard.beta) !=dim(datM)[[2]]. Consider transposing datM.")}
 if (plots ) {par(mfrow=c(2,2))}
 beta1.v = goldstandard.beta

 if (calibrateUnitInterval ) {datM=CalibrateUnitInterval(datM)}
```

```
 ### estimate initial weight matrix from type1 distribution
 w0.m = matrix(0,nrow=length(beta1.v),ncol=nL);
 w0.m[which(beta1.v <= th1.v[1]),1] = 1;
 w0.m[intersect(which(beta1.v > th1.v[1]),which(beta1.v <= th1.v[2])),2] = 1;
 w0.m[which(beta1.v > th1.v[2]),3] = 1;
 ### fit type1
 print("Fitting EM beta mixture to goldstandard probes");
 set.seed(1)
 rand.idx = sample(1:length(beta1.v),min(c(nfit, length(beta1.v))  )   ,replace=FALSE)
 em1.o = blc(matrix(beta1.v[rand.idx],ncol=1),w=w0.m[rand.idx,],maxiter=niter,tol=tol);
 subsetclass1.v = apply(em1.o$w,1,which.max);
 subsetth1.v = c(mean(max(beta1.v[rand.idx[subsetclass1.v==1]]),min(beta1.v[rand.idx[subset
class1.v==2]])),mean(max(beta1.v[rand.idx[subsetclass1.v==2]]),min(beta1.v[rand.idx[subsetcla
ss1.v==3]])));
 class1.v = rep(2,length(beta1.v));
 class1.v[which(beta1.v < subsetth1.v[1])] = 1;
 class1.v[which(beta1.v > subsetth1.v[2])] = 3;
 nth1.v = subsetth1.v;
 print("Done");

 ### generate plot from estimated mixture
 if(plots){
   print("Check");
   tmpL.v = as.vector(rmultinom(1:nL,length(beta1.v),prob=em1.o$eta));
   tmpB.v = vector();
   for(l in 1:nL){
     tmpB.v = c(tmpB.v,rbeta(tmpL.v[l],em1.o$a[l,1],em1.o$b[l,1]));
   }
   plot(density(beta1.v),main= paste("Type1fit-", sep=""));
   d.o = density(tmpB.v);
   points(d.o$x,d.o$y,col="green",type="l")
   legend(x=0.5,y=3,legend=c("obs","fit"),fill=c("black","green"),bty="n");
 }

 ### Estimate Modes
 if (  sum(class1.v==1)==1 ){ mod1U= beta1.v[class1.v==1]}
 if (  sum(class1.v==3)==1 ){ mod1M= beta1.v[class1.v==3]}
 if (  sum(class1.v==1) >1){
   d1U.o = density(beta1.v[class1.v==1])
   mod1U = d1U.o$x[which.max(d1U.o$y)]
 }
 if (  sum(class1.v==3)>1 ){
   d1M.o = density(beta1.v[class1.v==3])
   mod1M = d1M.o$x[which.max(d1M.o$y)]
 }

 ### BETA 2
 for (ii in 1:dim(datM)[[1]] ){
   printFlush(paste("ii=",ii))
   sampleID=ii
   beta2.v = as.numeric(datM[ii,])
```

```
  d2U.o = density(beta2.v[which(beta2.v<0.4)]);
  d2M.o = density(beta2.v[which(beta2.v>0.6)]);
  mod2U = d2U.o$x[which.max(d2U.o$y)]
  mod2M = d2M.o$x[which.max(d2M.o$y)]

  ### now deal with type2 fit
  th2.v = vector();
  th2.v[1] = nth1.v[1] + (mod2U-mod1U);
  th2.v[2] = nth1.v[2] + (mod2M-mod1M);

  ### estimate initial weight matrix
  w0.m = matrix(0,nrow=length(beta2.v),ncol=nL);
  w0.m[which(beta2.v <= th2.v[1]),1] = 1;
  w0.m[intersect(which(beta2.v > th2.v[1]),which(beta2.v <= th2.v[2])),2] = 1;
  w0.m[which(beta2.v > th2.v[2]),3] = 1;

  print("Fitting EM beta mixture to input probes");
  # I fixed an error in the following line (replaced beta1 by beta2)
  set.seed(1)
  rand.idx = sample(1:length(beta2.v),min(c(nfit, length(beta2.v)),na.rm=TRUE)  ,replace=FA
LSE)
  em2.o = blc2(Y=matrix(beta2.v[rand.idx],ncol=1),w=w0.m[rand.idx,],maxiter=niter,tol=tol,verb
ose=TRUE);
  print("Done");

  ### for type II probes assign to state (unmethylated, hemi or full methylation)
  subsetclass2.v = apply(em2.o$w,1,which.max);


  if (sum(subsetclass2.v==2)>0 ){
    subsetth2.v = c(mean(max(beta2.v[rand.idx[subsetclass2.v==1]]),min(beta2.v[rand.idx[subs
etclass2.v==2]])),
                mean(max(beta2.v[rand.idx[subsetclass2.v==2]]),min(beta2.v[rand.idx[subsetclas
s2.v==3]])));
  }
  if (sum(subsetclass2.v==2)==0 ){
    subsetth2.v = c(1/2*max(beta2.v[rand.idx[subsetclass2.v==1]])+ 1/2*mean(beta2.v[rand.idx[
subsetclass2.v==3]]), 1/3*max(beta2.v[rand.idx[subsetclass2.v==1]])+ 2/3*mean(beta2.v[rand.id
x[subsetclass2.v==3]]));
  }

  class2.v = rep(2,length(beta2.v));
  class2.v[which(beta2.v <= subsetth2.v[1])] = 1;
  class2.v[which(beta2.v >= subsetth2.v[2])] = 3;

  ### generate plot
  if(plots){
    tmpL.v = as.vector(rmultinom(1:nL,length(beta2.v),prob=em2.o$eta));
    tmpB.v = vector();
    for(lt in 1:nL){
      tmpB.v = c(tmpB.v,rbeta(tmpL.v[lt],em2.o$a[lt,1],em2.o$b[lt,1]));
    }
```

```
    plot(density(beta2.v),  main= paste("Type2fit-",sampleID,sep="")  );
    d.o = density(tmpB.v);
    points(d.o$x,d.o$y,col="green",type="l")
    legend(x=0.5,y=3,legend=c("obs","fit"),fill=c("black","green"),bty="n");
  }

  classAV1.v = vector();classAV2.v = vector();
  for(l in 1:nL){
    classAV1.v[l] =  em1.o$mu[l,1];
    classAV2.v[l] =  em2.o$mu[l,1];
  }

  ### start normalising input probes
  print("Start normalising input probes");
  nbeta2.v = beta2.v;
  ### select U probes
  lt = 1;
  selU.idx = which(class2.v==lt);
  selUR.idx = selU.idx[which(beta2.v[selU.idx] > classAV2.v[lt])];
  selUL.idx = selU.idx[which(beta2.v[selU.idx] < classAV2.v[lt])];
  ### find prob according to typeII distribution
  p.v = pbeta(beta2.v[selUR.idx],em2.o$a[lt,1],em2.o$b[lt,1],lower.tail=FALSE);
  ### find corresponding quantile in type I distribution
  q.v = qbeta(p.v,em1.o$a[lt,1],em1.o$b[lt,1],lower.tail=FALSE);
  nbeta2.v[selUR.idx] = q.v;
  p.v = pbeta(beta2.v[selUL.idx],em2.o$a[lt,1],em2.o$b[lt,1],lower.tail=TRUE);
  ### find corresponding quantile in type I distribution
  q.v = qbeta(p.v,em1.o$a[lt,1],em1.o$b[lt,1],lower.tail=TRUE);
  nbeta2.v[selUL.idx] = q.v;

  ### select M probes
  lt = 3;
  selM.idx = which(class2.v==lt);
  selMR.idx = selM.idx[which(beta2.v[selM.idx] > classAV2.v[lt])];
  selML.idx = selM.idx[which(beta2.v[selM.idx] < classAV2.v[lt])];
  ### find prob according to typeII distribution
  p.v = pbeta(beta2.v[selMR.idx],em2.o$a[lt,1],em2.o$b[lt,1],lower.tail=FALSE);
  ### find corresponding quantile in type I distribution
  q.v = qbeta(p.v,em1.o$a[lt,1],em1.o$b[lt,1],lower.tail=FALSE);
  nbeta2.v[selMR.idx] = q.v;


  if(doH){ ### if TRUE also correct type2 hemimethylated probes
    ### select H probes and include ML probes (left ML tail is not well described by a beta-distri
bution).
    lt = 2;
    selH.idx = c(which(class2.v==lt),selML.idx);
    minH = min(beta2.v[selH.idx],na.rm=TRUE)
    maxH = max(beta2.v[selH.idx],na.rm=TRUE)
    deltaH = maxH - minH;
    #### need to do some patching
    deltaUH = -max(beta2.v[selU.idx],na.rm=TRUE) + min(beta2.v[selH.idx],na.rm=TRUE)
    deltaHM = -max(beta2.v[selH.idx],na.rm=TRUE) + min(beta2.v[selMR.idx],na.rm=TRUE)
```

```
    ## new maximum of H probes should be
    nmaxH = min(nbeta2.v[selMR.idx],na.rm=TRUE) - deltaHM;
    ## new minimum of H probes should be
    nminH = max(nbeta2.v[selU.idx],na.rm=TRUE) + deltaUH;
    ndeltaH = nmaxH - nminH;

    ### perform conformal transformation (shift+dilation)
    ## new_beta_H(i) = a + hf*(beta_H(i)-minH);
    hf = ndeltaH/deltaH ;
    ### fix lower point first
    nbeta2.v[selH.idx] = nminH + hf*(beta2.v[selH.idx]-minH);

   }

   ### generate final plot to check normalisation
   if(plots){
    print("Generating final plot");
    d1.o = density(beta1.v);
    d2.o = density(beta2.v);
    d2n.o = density(nbeta2.v);
    ymax = max(d2.o$y,d1.o$y,d2n.o$y);
    plot(density(beta2.v),type="l",ylim=c(0,ymax),xlim=c(0,1), main=paste("CheckBMIQ-",sa
mpleID,sep="") );
    points(d1.o$x,d1.o$y,col="red",type="l");
    points(d2n.o$x,d2n.o$y,col="blue",type="l");
    legend(x=0.5,y=ymax,legend=c("type1","type2","type2-BMIQ"),bty="n",fill=c("red","black
","blue"));
   }

   datM[ii,]= nbeta2.v ;
  } # end of for (ii=1 loop
  datM
} # end of function BMIQcalibration

BMIQ = function(beta.v,design.v,nL=3,doH=TRUE,nfit=50000,th1.v=c(0.2,0.75),th2.v=NULL,nit
er=5,tol=0.001,plots=TRUE,sampleID=1,calibrateUnitInterval=TRUE){

  if (calibrateUnitInterval) {
   rangeBySample=range(beta.v,na.rm=TRUE)
   minBySample=rangeBySample[1]
   maxBySample=rangeBySample[2]
   if ( (minBySample<0 | maxBySample>1) & !is.na(minBySample) & !is.na(maxBySample) ) {
    y1=c(0.001,.999)
    x1=c(minBySample,maxBySample)
    lm1=lm( y1 ~ x1 )
    intercept1=coef(lm1)[[1]]
    slope1=coef(lm1)[[2]]
    beta.v=intercept1+slope1*beta.v
   } # end of if
  } # end of if (calibrateUnitInterval
```

```
  type1.idx = which(design.v==1);
  type2.idx = which(design.v==2);

  beta1.v = beta.v[type1.idx];
  beta2.v = beta.v[type2.idx];

  ### estimate initial weight matrix from type1 distribution
  w0.m = matrix(0,nrow=length(beta1.v),ncol=nL);
  w0.m[which(beta1.v <= th1.v[1]),1] = 1;
  w0.m[intersect(which(beta1.v > th1.v[1]),which(beta1.v <= th1.v[2])),2] = 1;
  w0.m[which(beta1.v > th1.v[2]),3] = 1;

  ### fit type1
  print("Fitting EM beta mixture to goldstandard probes");
  set.seed(1)
  rand.idx = sample(1:length(beta1.v),min(c(nfit, length(beta1.v))  )  ,replace=FALSE)
  em1.o = blc2(Y=matrix(beta1.v[rand.idx],ncol=1),w=w0.m[rand.idx,],maxiter=niter,tol=tol);
  subsetclass1.v = apply(em1.o$w,1,which.max);
  subsetth1.v = c(mean(max(beta1.v[rand.idx[subsetclass1.v==1]]),min(beta1.v[rand.idx[subset
class1.v==2]])),mean(max(beta1.v[rand.idx[subsetclass1.v==2]]),min(beta1.v[rand.idx[subsetcla
ss1.v==3]],na.rm=TRUE)));
  class1.v = rep(2,length(beta1.v));
  class1.v[which(beta1.v < subsetth1.v[1])] = 1;
  class1.v[which(beta1.v > subsetth1.v[2])] = 3;
  nth1.v = subsetth1.v;
  print("Done");

  ### generate plot from estimated mixture
  if(plots){
    print("Check");
    tmpL.v = as.vector(rmultinom(1:nL,length(beta1.v),prob=em1.o$eta));
    tmpB.v = vector();
    for(l in 1:nL){
      tmpB.v = c(tmpB.v,rbeta(tmpL.v[l],em1.o$a[l,1],em1.o$b[l,1]));
    }

    pdf(paste("Type1fit-",sampleID,".pdf",sep=""),width=6,height=4);
    plot(density(beta1.v));
    d.o = density(tmpB.v);
    points(d.o$x,d.o$y,col="green",type="l")
    legend(x=0.5,y=3,legend=c("obs","fit"),fill=c("black","green"),bty="n");
    dev.off();
  }

  ### Estimate Modes
  if (  sum(class1.v==1)==1 ){ mod1U= beta1.v[class1.v==1]}
  if (  sum(class1.v==3)==1 ){ mod1M= beta1.v[class1.v==3]}
  if (  sum(class1.v==1) >1){
    d1U.o = density(beta1.v[class1.v==1])
    mod1U = d1U.o$x[which.max(d1U.o$y)]
  }
  if (  sum(class1.v==3)>1 ){
```

```
  d1M.o = density(beta1.v[class1.v==3])
  mod1M = d1M.o$x[which.max(d1M.o$y)]
 }


 d2U.o = density(beta2.v[which(beta2.v<0.4)]);
 d2M.o = density(beta2.v[which(beta2.v>0.6)]);
 mod2U = d2U.o$x[which.max(d2U.o$y)]
 mod2M = d2M.o$x[which.max(d2M.o$y)]

 ### now deal with type2 fit
 th2.v = vector();
 th2.v[1] = nth1.v[1] + (mod2U-mod1U);
 th2.v[2] = nth1.v[2] + (mod2M-mod1M);

 ### estimate initial weight matrix
 w0.m = matrix(0,nrow=length(beta2.v),ncol=nL);
 w0.m[which(beta2.v <= th2.v[1]),1] = 1;
 w0.m[intersect(which(beta2.v > th2.v[1]),which(beta2.v <= th2.v[2])),2] = 1;
 w0.m[which(beta2.v > th2.v[2]),3] = 1;

 print("Fitting EM beta mixture to input probes");
 set.seed(1)
 rand.idx = sample(1:length(beta2.v),min(c(nfit, length(beta2.v)),na.rm=TRUE)   ,replace=FAL
SE)
 em2.o = blc2(Y=matrix(beta2.v[rand.idx],ncol=1),w=w0.m[rand.idx,],maxiter=niter,tol=tol);
 print("Done");

 ### for type II probes assign to state (unmethylated, hemi or full methylation)
 subsetclass2.v = apply(em2.o$w,1,which.max);

 if (sum(subsetclass2.v==2)>0 ){
   subsetth2.v = c(mean(max(beta2.v[rand.idx[subsetclass2.v==1]]),min(beta2.v[rand.idx[subse
tclass2.v==2]])),
                mean(max(beta2.v[rand.idx[subsetclass2.v==2]]),min(beta2.v[rand.idx[subsetclass
2.v==3]])));
 }
 if (sum(subsetclass2.v==2)==0 ){
   subsetth2.v = c(1/2*max(beta2.v[rand.idx[subsetclass2.v==1]])+ 1/2*mean(beta2.v[rand.idx[s
ubsetclass2.v==3]]), 1/3*max(beta2.v[rand.idx[subsetclass2.v==1]])+ 2/3*mean(beta2.v[rand.idx
[subsetclass2.v==3]]));
 }

 class2.v = rep(2,length(beta2.v));
 class2.v[which(beta2.v <= subsetth2.v[1])] = 1;
 class2.v[which(beta2.v >= subsetth2.v[2])] = 3;

 ### generate plot
 if(plots){
   tmpL.v = as.vector(rmultinom(1:nL,length(beta2.v),prob=em2.o$eta));
   tmpB.v = vector();
   for(lt in 1:nL){
     tmpB.v = c(tmpB.v,rbeta(tmpL.v[lt],em2.o$a[lt,1],em2.o$b[lt,1]));
```

```
  }
  pdf(paste("Type2fit-",sampleID,".pdf",sep=""),width=6,height=4);
  plot(density(beta2.v));
  d.o = density(tmpB.v);
  points(d.o$x,d.o$y,col="green",type="l")
  legend(x=0.5,y=3,legend=c("obs","fit"),fill=c("black","green"),bty="n");
  dev.off();
}

classAV1.v = vector();classAV2.v = vector();
for(l in 1:nL){
  classAV1.v[l] =  em1.o$mu[l,1];
  classAV2.v[l] =  em2.o$mu[l,1];
}

### start normalising input probes
print("Start normalising input probes");
nbeta2.v = beta2.v;
### select U probes
lt = 1;
selU.idx = which(class2.v==lt);
selUR.idx = selU.idx[which(beta2.v[selU.idx] > classAV2.v[lt])];
selUL.idx = selU.idx[which(beta2.v[selU.idx] < classAV2.v[lt])];
### find prob according to typeII distribution
p.v = pbeta(beta2.v[selUR.idx],em2.o$a[lt,1],em2.o$b[lt,1],lower.tail=FALSE);
### find corresponding quantile in type I distribution
q.v = qbeta(p.v,em1.o$a[lt,1],em1.o$b[lt,1],lower.tail=FALSE);
nbeta2.v[selUR.idx] = q.v;
p.v = pbeta(beta2.v[selUL.idx],em2.o$a[lt,1],em2.o$b[lt,1],lower.tail=TRUE);
### find corresponding quantile in type I distribution
q.v = qbeta(p.v,em1.o$a[lt,1],em1.o$b[lt,1],lower.tail=TRUE);
nbeta2.v[selUL.idx] = q.v;

### select M probes
lt = 3;
selM.idx = which(class2.v==lt);
selMR.idx = selM.idx[which(beta2.v[selM.idx] > classAV2.v[lt])];
selML.idx = selM.idx[which(beta2.v[selM.idx] < classAV2.v[lt])];
### find prob according to typeII distribution
p.v = pbeta(beta2.v[selMR.idx],em2.o$a[lt,1],em2.o$b[lt,1],lower.tail=FALSE);
### find corresponding quantile in type I distribution
q.v = qbeta(p.v,em1.o$a[lt,1],em1.o$b[lt,1],lower.tail=FALSE);
nbeta2.v[selMR.idx] = q.v;


if(doH){ ### if TRUE also correct type2 hemimethylated probes
  ### select H probes and include ML probes (left ML tail is not well described by a beta-distrib
ution).
  lt = 2;
  selH.idx = c(which(class2.v==lt),selML.idx);
  minH = min(beta2.v[selH.idx],na.rm=TRUE)
  maxH = max(beta2.v[selH.idx],na.rm=TRUE)
  deltaH = maxH - minH;
```

```
    #### need to do some patching
    deltaUH = -max(beta2.v[selU.idx],na.rm=TRUE) + min(beta2.v[selH.idx],na.rm=TRUE)
    deltaHM = -max(beta2.v[selH.idx],na.rm=TRUE) + min(beta2.v[selMR.idx],na.rm=TRUE)

    ## new maximum of H probes should be
    nmaxH = min(nbeta2.v[selMR.idx],na.rm=TRUE) - deltaHM;
    ## new minimum of H probes should be
    nminH = max(nbeta2.v[selU.idx],na.rm=TRUE) + deltaUH;
    ndeltaH = nmaxH - nminH;

    ### perform conformal transformation (shift+dilation)
    ## new_beta_H(i) = a + hf*(beta_H(i)-minH);
    hf = ndeltaH/deltaH ;
    ### fix lower point first
    nbeta2.v[selH.idx] = nminH + hf*(beta2.v[selH.idx]-minH);

  }

  pnbeta.v = beta.v;
  pnbeta.v[type1.idx] = beta1.v;
  pnbeta.v[type2.idx] = nbeta2.v;

  ### generate final plot to check normalisation
  if(plots){
    print("Generating final plot");
    d1.o = density(beta1.v);
    d2.o = density(beta2.v);
    d2n.o = density(nbeta2.v);
    ymax = max(d2.o$y,d1.o$y,d2n.o$y);
    pdf(paste("CheckBMIQ-",sampleID,".pdf",sep=""),width=6,height=4)
    plot(density(beta2.v),type="l",ylim=c(0,ymax),xlim=c(0,1));
    points(d1.o$x,d1.o$y,col="red",type="l");
    points(d2n.o$x,d2n.o$y,col="blue",type="l");
    legend(x=0.5,y=ymax,legend=c("type1","type2","type2-BMIQ"),bty="n",fill=c("red","black",
"blue"));
    dev.off();
  }

  print(paste("Finished for sample ",sampleID,sep=""));

  return(list(nbeta=pnbeta.v,class1=class1.v,class2=class2.v,av1=classAV1.v,av2=classAV2.v,
hf=hf,th1=nth1.v,th2=th2.v));

}

CheckBMIQ = function(beta.v,design.v,pnbeta.v){### pnbeta is BMIQ normalised profile

  type1.idx = which(design.v==1);
  type2.idx = which(design.v==2);
   beta1.v = beta.v[type1.idx];
  beta2.v = beta.v[type2.idx];
  pnbeta2.v = pnbeta.v[type2.idx];
```

```
} # end of function CheckBMIQ

CalibrateUnitInterval=function(datM,onlyIfOutside=TRUE){

  rangeBySample=data.frame(lapply(data.frame(t(datM)),range,na.rm=TRUE))
  minBySample=as.numeric(rangeBySample[1,])
  maxBySample=as.numeric(rangeBySample[2,])
  if (onlyIfOutside) { indexSamples=which((minBySample<0 | maxBySample>1) & !is.na(minBy
Sample) & !is.na(maxBySample))
  }
  if (!onlyIfOutside) { indexSamples=1:length(minBySample)}
  if ( length(indexSamples)>=1 ){
    for ( i in indexSamples) {
      y1=c(0.001,0.999)
      x1=c(minBySample[i],maxBySample[i])
      lm1=lm( y1 ~ x1 )
      intercept1=coef(lm1)[[1]]
      slope1=coef(lm1)[[2]]
      datM[i,]=intercept1+slope1*datM[i,]
    } # end of for loop
  }
  datM
} #end of function for calibrating to [0,1]



fastImputation= FALSE
nSamples=dim(dat1)[[2]]-1
nProbes= dim(dat1)[[1]]

meanMethBySample =as.numeric(apply(as.matrix(dat1[,-1]),2,mean,na.rm=TRUE))
minMethBySample   =as.numeric(apply(as.matrix(dat1[,-1]),2,min,na.rm=TRUE))
maxMethBySample  =as.numeric(apply(as.matrix(dat1[,-1]),2,max,na.rm=TRUE))

datMethUsed= t(dat1[,-1])
colnames(datMethUsed)=as.character(dat1[,1])

noMissingPerSample=apply(as.matrix(is.na(datMethUsed)),1,sum)
table(noMissingPerSample)

#STEP 2: Imputing
if (! fastImputation & nSamples>1 & max(noMissingPerSample,na.rm=TRUE)<3000 ){

  # run the following code if there is at least one missing
  if ( max(noMissingPerSample,na.rm=TRUE)>0 ){
    dimnames1=dimnames(datMethUsed)
    datMethUsed= data.frame(t(impute.knn(t(datMethUsed))$data))
    dimnames(datMethUsed)=dimnames1
  } # end of if
} # end of if (! fastImputation )

if ( max(noMissingPerSample,na.rm=TRUE)>=3000 ) fastImputation=TRUE
```

```
if ( fastImputation | nSamples==1 ){
  noMissingPerSample=apply(as.matrix(is.na(datMethUsed)),1,sum)
  table(noMissingPerSample)
  if ( max(noMissingPerSample,na.rm=TRUE)>0 & max(noMissingPerSample,na.rm=TRUE) >=
3000 ) {normalizeData=FALSE}

  # run the following code if there is at least one missing
  if ( max(noMissingPerSample,na.rm=TRUE)>0 & max(noMissingPerSample,na.rm=TRUE) < 3
000 ){
    dimnames1=dimnames(datMethUsed)
    for (i in which(noMissingPerSample>0) ){
      selectMissing1=is.na(datMethUsed[i,])
      datMethUsed[i,selectMissing1] = as.numeric(probeAnnotation21kdatMethUsed$goldstandar
d2[selectMissing1])
    } # end of for loop
    dimnames(datMethUsed)=dimnames1
  } # end of if
} # end of if (! fastImputation )

normalizeData <- as.logical(TRUE)
#gold.mean=as.numeric(apply(datMethUsed,2,mean,na.rm=TRUE))
gold.mean=probeAnnotation21kdatMethUsed$goldstandard2

if (normalizeData ){
  datMethUsedNormalized=BMIQcalibration(datM=datMethUsed,goldstandard.beta= gold.mean
,plots=FALSE)
}
if (!normalizeData ){ datMethUsedNormalized=datMethUsed }
```