Optimization of an age estimation model for semen samples based on DNA methylation using SNaPshot technique.

Final Master Project

Marc Català Blanco

Master in Genetics, Physics and Forensic Chemistry

Project performed in “Luís Concheiro” Forensic Sciences Institute (INCIFOR)

Tutored in USC by Ana Freire Aradas, Ph.D.

Tutored in URV by Raúl Beltrán Debón, Ph.D.

UNIVERSITAT ROVIRA i VIRGILI

Tarragona, June 2018

Acknowledgements

First of all, I would like to special thank Ana Freire Aradas, Ph.D. and all the team from "Luís Concheiro" Forensic Sciences Institute (INCIFOR) for giving me the chance to work in this research project in which I have received guidance whenever I have needed it. It has been a very fruitful experience and I have improved my knowledge about forensic genetics, coping with all strict requirements which entails laboratory work in this interesting field of science.

On the other hand, I would also like to thank Raúl Beltrán Debón, Ph.D. from the Faculty of Chemistry at Rovira i Virgili University. He has advised me whenever I had troubles or questions during the development of this project.

Finally, I want to express my gratitude to my parents and sister for providing me with complete support through all my years of study; without their encouragement, this project would not have been possible.

Index

Prevailing abbreviations

-bp: base pair(s).

-ddNTP: Dideoxynucleotide Triphosphate.

-DNA: Deoxyribonucleic Acid.

-DNMT: DNA methyltransferase.

-dNTP: Deoxynucleotide Triphosphate.

-dsDNA: double-stranded DNA.

-MAD: Median Absolute Deviation.

-mtDNA: mitochondrial DNA.

-NTP: Nucleoside Triphosphate.

-PCR: Polymerase Chain Reaction.

-RNA: Ribonucleic Acid.

-RFLP: Restriction Fragment Length Polymorphism.

-SBE: Single Base Extension.

-SNP: Single Nucleotide Polymorphism.

-ssDNA: single-stranded DNA.

-STR: Short Tandem Repeat.

Abstract

Epigenetic interest for age estimation in the forensic context has risen noticeably during last few years. DNA methylation analysis in CpG dinucleotides has turned into be the reference biomarker for age estimation of individuals. Many articles have provided valuable information for body fluids such as blood or saliva samples; nevertheless, few has been reported about semen body fluid. A recent report by Lee et al. was able to develop two age prediction regression models by using only 3 very informative CpG sites (cg06304190 in *TTC7B* gene, cg12837463, and cg06979108 in *NOX4* gene) according to DNA methylation patterns in semen samples. During the present project, 50 μL and 300 μL semen samples from 5 individuals have been bisulfite-converted and analysed by methylation SNaPshot technique for the CpGs selected in the reference paper. Thus, by following the regression models described in Lee et al., an age approach has been achieved successfully with an average absolute difference (MAD) of approximately 5 years between the predicted age and the chronological age in 4 out of 5 individuals. This way, the accuracy and the adaptability of these initial regression models has been verified meaning that, the MAD obtained for original samples in Lee et al. is maintained in other samples from individuals with distinct geographical precedence and subjected to different environmental conditions. However, the limited number of samples considered in this project has not enabled statistical treatment of data, therefore the development of a new regression model has not been possible. That is why further large-scale studies may be developed in order to implement new age estimation models which can aid to solve forensic relevant cases such as sexual assaults or rapes, in which semen is a biological fluid frequently found as an evidence in the crime scene.

1. Introduction

1.1 Historical Background

Forensic genetics is a subspecialty of genetics and Legal Medicine which can be defined as the application of genetics in order to study the variation of inherited traits in populations in matters to the resolution of legal issues (Carracedo 2013; Carracedo, Salas, and Lareu 2010).

The tasks carried out by a forensic geneticist can be clustered in three main groups: analysis of the evidences collected from a crime scene, paternity testing and the identification of human remains (Goodwin 2007).

The beginning of forensic genetics dates back to the year 1900, year in which the Nobel laureate Karl Landsteiner published an article under the name of *Anti-fermentative, lytic and agglutinating effects of blood serum and lymph* (Landsteiner 1900). In this article, the author described that the serum of healthy humans has an agglutinating effect in animal blood cells and among individuals as well: that is how ABO grouping system was created. The implementation of ABO typing marked a milestone in the history of forensic genetics and, by 1931, this assay became a standard method in most of forensic laboratories. These serological techniques were considered to be a great tool; however, many inconveniences aroused by the great amount of biological sample which is needed to perform them in order to obtain discerning results (Goodwin 2007).

During the 1960s and 1970s, Sanger and Coulson developed a simple and fast procedure for determining nucleotide sequences in ssDNA through *Escherichia Coli* DNA polymerase I and bacteriophage T4 DNA polymerase under conditions of variable limiting nucleoside triphosphate (NTPs) and further separation of the products by acrylamide gel electrophoresis (Sanger and Coulson 1975). Shortly after, a new method for sequencing DNA was described; in this case 2',3'-dideoxynucleoside triphosphates were added, therefore acting as chain-terminating inhibitors of DNA polymerase (Sanger, Nicklen, and Coulson 1977). Additionally, a new method, called Southern blotting, described a procedure to transfer DNA fragments from agarose gels to cellulose nitrate filters; this technique has been very useful for the detection of DNA polymorphisms such as RFLPs (Kan and Dozy 1978; Southern 1975).

In 1983, Sir Alec John Jeffreys discovered that, within the human myoglobin gene, there were short fragments of DNA (15-50 bp) repeated in tandem: that is how the term 'minisatelite' emerged (Blanchetot et al. 1983).

Minisatellites were originally detected by hybridization of probes to Southern blots of genomic DNA digested with restriction enzymes, giving rise to multiband patterns known as DNA fingerprints (Carracedo 2013). These hypervariable fragments are unique to individuals and very stable in the DNA comprised in sexual chromosomes. They can be easily isolated from blood and semen stains, hence demonstrating the DNA fingerprinting method potential for the identification of rape suspects (Gill, Jeffreys, and Werrett 1985).

In forensic terminology, DNA fingerprinting (also known as DNA typing or DNA profiling) can be defined as the comparison of the nuclear DNA from an individual cells with the biological evidence found in a crime scene, or with the DNA of another person in matters of identification or exclusion of potential suspects (Lutz Roewer 2013).

Despite the discovery of 'minisatelites' and their application to forensic genetics by DNA fingerprinting, the major issue was, once again, the limited amount of sample available in forensic samples. In this context, it is important to highlight the introduction of a technique which allows to amplify specific regions of interest of DNA: the polymerase chain reaction (PCR). This technique was introduced by Kary Mullis and applied for the first time for the detection of sickle cell mutation in a rapid and, at least, two orders more sensitive than conventional Southern blotting; this protocol comprised oligonucleotide primers and DNA polymerase so specific β-globin genes are amplified (Saiki et al. 1985).

The PCR, apart from raising the sensitivity of genomic analysis, reduced the time spent to obtain a DNA profile and could be used with degraded samples, a fact which is very common in the forensic field. The first application of the PCR in a forensic case dates back to 1991, when an innovative study implemented an SSO-typing system to study variations in human mtDNA control-region sequences. The SSO-typing system refers to analysis of DNA sequences already amplified by PCR and afterwards hybridized with sequence-specific oligonucleotide (SSO) probes. The authors chose the mtDNA control region because it is the main noncoding polymorphic region of human mtDNA genome and includes the origin of replication of one strand, D-loop region and both origins of transcription. This mtDNA SSO-typing system was overwhelmingly applied to a case involving individual identification of skeletal remains (Stoneking et al. 1991).

1.2 Polymorphisms of DNA

1.2.1 Short Tandem Repeats (STRs)

The application of PCR in forensic genetics was quickly followed by the analysis of short tandem repeats (STRs), also known as microsatellites or simple sequence repeats (SSRs). STRs are accordion-like short repeated DNA regions, usually between 2 and 6 bp, which have been widely used since early 1990s. These regions do not alter cellular functions and the number of copies are very variable among individuals (Butler 2004). STRs are widely found in prokaryotes and eukaryotes (including humans) but their distribution is not uniform: more than 90% are located in noncoding regions and they are less usual in subtelomeric regions (Fan and Chu 2007). Despite the fact that human genome contains abundant STRs, only a small portion of them have been chosen as markers for forensic purposes such as identity testing (Butler 2006).

The application of microsatellites to the forensic field led to the creation of databases such as CoDIS (Combined DNA Index System), an American database developed by FBI which contains many STRs profiles in matters to the investigation of missing people, criminals and also from forensic evidences found at a crime scene. The initial amount of STRs loci included into CoDIS were 13 (D3S1358, D5S818, D7S820, D8S1179, D13S317, D16S539, D18S51, D21S11, CSF1PO, FGA, TH01, TPOX and vWA) plus the sex determining locus Amelogenin (AMEL). At a later stage, the FBI Laboratory announced that 7 additional loci (D1S1656, D2S441, D2S1338, D10S1248, D12S391, D19S433 and D22S1045) would be added and could be analysed by specific multiplex amplification commercial kits such as the Global-Filer® PCR Amplification Kit (Moretti et al. 2016). Further investigations, highlighted the role of STRs located in chromosome Y (Y-STRs) in applications in linkage studies, paternity cases or ethnological issues among others (L Roewer et al. 1992). Taking into account their importance in forensic genetics, STRs profiling is still considered the first and foremost choice for genotyping in this field.

1.2.2 Single Nucleotide Polymorphisms (SNPs)

During last few years, Single Nucleotide Polymorphisms (SNPs) which are defined as single base pairs positions in genomic DNA which can vary in different individuals, have gained a lot of attention in forensic genetics. To be considered as SNP, the allelic frequency of the nucleotide variation has to be at least of 1% in the population considered (Brookes 1999). The appearance rate of SNPs is considerably high (approximately one every 1000-2000 bp), that is why they are suitable markers for haplotype analysis (especially in Y chromosome and mtDNA) and easy to automate by high-throughput genotyping techniques (Allah, Yang, and Li 2007; Sachidanandam et al. 2001).

The main advantages by which these kind of polymorphisms play a crucial role in forensic genetics are the following (Sobrino and Carracedo 2005):

- SNPs have lower mutation rates than STRs, thus being very suitable for paternity testing.
- They can be analysed from short amplicons, therefore being very useful for amplification of degraded samples.
- Finally, and as mentioned above, SNPs are appropriate for high-throughput technologies analysis, a tool very relevant for criminal DNA databases and large population studies.

Despite the fact that some SNPs can be the cause of many pathologies, they are located in intronic regions predominantly thus playing no role in the phenotype of the individual. These non-coding polymorphisms have been noted for their implication in the human genome in terms of molecular evolution and populations (Zhao et al. 2003) and also for forensic studies. For instance, small biallelic SNPs arrays are capable to differentiate closely related individuals such as brothers (Gill 2001). They have even been used in relevant forensic caseworks where STR profiles were not sufficient such as in the investigation of the 11-M Madrid terrorist attack (Carracedo 2013).

1.3 Initial techniques for Age Determination

1.3.1 Physical and skeletal analysis

In some forensic cases, such as the identification of human bodies, an approach to the age of individuals is a great tool when there are no further identity clues (Alkass et al. 2010). Initial techniques for age estimation in children and adolescents (young adults) were based on morphological methods which combine physical examinations, X-ray imaging of the hand and dental examination with a radiograph of the jaw region. In the case that the hand skeletal development was complete, additional X-ray or Computed Tomography from clavicles were also taken into consideration (Schmeling et al. 2016).

From all of these analysis mentioned above, it is important to underline the dental observation; in particular the degree of mineralization of the dentition, is the chosen method for most of the estimations of children and adolescent skeletal remains (Cavrić et al. 2016; Freire-Aradas, Phillips, and Lareu 2017; Lewis and Flavel 2006). The use of tooth mineralization has been successfully used in criminal and civil procedures; in detail, the study of the third molar has been widely accepted as a distinguishing feature between child and adult status (under and over 18 years old) (Lucas et al. 2016). Moreover, during last years, Magnetic Resonance Imaging has replaced X-Ray analysis due to its higher ratio assessment of third molars and its less invasiveness (De Tobel, Hillewig, and Verstraete 2017).

The major issue of the skeletal analysis is the lack of accuracy to elderly people (over 65 years of age; therefore, many molecular techniques have been described as it is shown in the section below.

1.3.2 Molecular methods

Up to now, five essential molecular alterations have been studied (Freire-Aradas, Phillips, and Lareu 2017):

1. Mitochondrial DNA deletions.
2. Telomeres shortening.
3. Advanced Glycation End-products (AGEs).
4. Aspartic Acid Racemization (AAR).
5. Signal-joint T-cell Receptor Excision Circles (sjTRECs).

The first molecular technique is based on the study of the relation between the accumulation of deletions in mtDNA and aging; in detail, the 4977 bp deletion has been widely studied in various tissues, specifically in postmitotic tissues with high energy demand (Cristoph Meissner et al. 1999). This nucleotide deletion has been linked to the lack of capacity of human mitochondria to accomplish the oxidative phosphorylation due to high amounts of free radicals (reactive oxygen species), therefore, leading to apoptotic signalling into cells (Chomyn and Attardi 2003; Christoph Meissner et al. 2008).

Secondly, telomere are structures located in the tips of chromosomes and their shortening occur in every round of mitosis due to DNA-polymerase is not able to completely replicate these last nucleotides during the S-phase of cell cycle. These nucleotide deletions, may be also influenced by reactive oxygen species. In order to solve this problem, cells induce the expression of telomerase but as aging occur this enzyme is supressed. Therefore, when telomeres reach a critical short length, are recognized as DNA damage and activate DNA damage checkpoints (Jiang, Ju, and Rudolph 2007).

Another molecular technique comprises the analysis of advanced glycation end-products (AGEs), which are a heterogeneous group of macromolecules in which lipids, nonenzymatic glycation of proteins and nucleic acids are included. AGEs can be exogenous (if they are absorbed by the diet such as in deep-fried, roasted or grilled food) or endogenous (when produced by the body; for instance diabetics show high levels of AGEs due to altered glucose metabolism). This products increase the oxidative damage and are accumulated in brain, eyes, erythrocytes or liver among others with increasing age. That is why, lower intakes of AGEs may increase healthy aging and greater longevity (Semba, Nicklett, and Ferrucci 2010).

On the other hand, the determination of the enantiomers L and D from acid aspartic (Asp) from tooth enamel and later from dentin and cementum has also been applied in age estimation for forensic purposes (Arany et al. 2004). This technique is considered to be very accurate but teeth need to be exposed to high temperatures, leading to bias in results. (Freire-Aradas, Phillips, and Lareu 2017).

Finally, the last molecular method is the study of sjTRECs, which are circular DNA molecules which are produced during the rearrangement of naïve T-cells. The amount of sjTRECs are reduced in an age-related way, that is why this method can be useful in cases in which morphological or skeletal information is lacking.

However, the standard error of this age predicting method is relatively high (approximately ± 10 years) and it has only been proved in peripheral blood samples (Ou et al. 2011, 2012).

Despite the fact that these molecular approaches have contributed to ameliorate the techniques for age estimation, they have many disadvantages (high standard errors, specificity for one or few sample types, and lack of reproducibility...), hence other methods have gained prominence in other scientific fields: that is the case of epigenetics.

1.4 Epigenetics and Aging: DNA methylation

Aging is defined as a multifactorial process characterized by a progressive decay of physiological functions, prompting the emergence of age-related diseases and increasing the susceptibility of environmental effects; this can lead to a decrease of the lifespan of individuals and, eventually into death (D'Aquila et al. 2013; Pal and Tyler 2016). The modifications resulted from the interaction between genetics and environment are named under the name of the term "epigenetic clock" or "epigenetic drift" (Freire-Aradas, Phillips, and Lareu 2017). In this context, it is crucial to take into account the role of epigenetics, which refer to any heritable biological processes in which DNA, RNA and proteins are chemically or structurally altered without changing their primary sequence. There are many epigenetic modifications linked to variations in gene expression and DNA replication and recombination (Sierra, Fernández, and Fraga 2015; Weinhold 2006); that is the case of histone modifications, the regulation of noncoding short and long RNAs and the methylation of DNA (Guillaumet-Adkins et al. 2017). During this project, DNA methylation has been crucial to provide an age estimation; that is why this epigenetic mechanism will be focused during this section.

The DNA methylation is a heritable epigenetic mark which consists on the covalent addition of a methyl group to nucleotide rings in specific cytosines. This process is catalysed by DNA methyltransferases (DNMTs), a group of enzymes which transfer methyl groups from S-adenosylmethionine (SAM) to the fifth carbon of a cytosine residue, therefore forming the 5-methylcytosine (Moore, Le, and Fan 2013) (Figure 1). DNA methylation can occur in any part of the genome; nevertheless, most of this epigenetic mark takes place in CpG dinucleotides (Jin, Li, and Robertson 2011).

The term CpG is an abbreviation for cytosine and guanine separated by a phosphate group, which acts as a linker between the two nucleotides. The great majority of CpG nucleotides (also named CpG positions or sites) are predominantly clustered in the so called 'CpG islands' (Derek K Lim et al. 2010; Li and Zhang 2014). Hence, CpG islands are defined as 1000 bp long DNA sequences in which high density of unmethylated CpG sites is found. It is important to highlight that about 70% of promoters for housekeeping genes are located in CpG islands. Moreover, and taken into account that DNA methylation is associated with gene silencing, the methylation in CpG islands is a crucial mechanisms for the regulation of the gene expression (Moore, Le, and Fan 2013; Schübeler 2015).

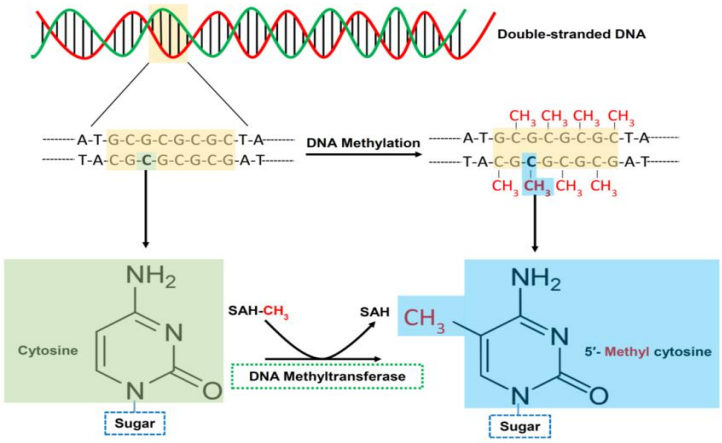

Figure 1. DNA methylation catalysed by DNA methyltransferases (Alam et al. 2016).

The relation between aging and DNA methylation is complex and can be defined, in general terms, as a ratio between the amount of hypermethylated and hypomethylated CpG positions (López-Otín et al. 2013). It has been widely reported that hypomethylation patterns proportionally increase while aging. Nevertheless, in detail, this affirmation cannot be applied to all CpG sites; for instance cancer cells and other age-related diseases may be the consequence of aberrant methylation (hypermethylation) of promoter CpG islands related with silencing of tumor-supressor genes (Maegawa et al. 2010).

These pathologies may be related with DNA methyltransferases which fail and become less active as aging occurs; for instance, it can be noted in the lower addition of methyl moieties after each cell division. Furthermore, changes in metabolism and diet can lead to a reduction of the intake of folate, from which methyl groups are synthetized (Heyn et

al. 2012). Other changes in DNA methylation patterns seem to be related with continuous and systemic exposures such as to cigarette smoke (Qiu et al. 2015).

In the forensic context, the study of the DNA methylation has been a turning point in recent years. This technique provide an accurate approach for age estimation in a wide type of samples (blood, saliva, semen) which are frequently found in a crime scene. Therefore, by calculating with statistical methods the predicted age, the number of suspects can be reliably delimited. Moreover, the DNA methylation analysis have less disadvantages than the molecular techniques described in the previous section in terms of the amount of sample required, the greater accuracy and the capacity to deal with degraded samples.

1.5 Analytical Methods for DNA methylation

Nowadays, there exists a great number of techniques for the study of DNA methylation; The first step before profiling consists on a pre-treatment phase of the DNA prior to analysis, herein enzyme digestion (applying restriction endonucleases), affinity enrichment (using immunoglobulins or methyl-binding proteins) and sodium bisulfite conversion are contemplated. On the other hand, the amount of techniques for DNA methylation profiling is also very wide (Whole-Genome Bisulfite Sequencing (WGBS), Pyrosequencing, EpiTYPER ® system coupled with MALDI-TOF spectrometry and Single Base Extension (SBE) among others) (Freire-Aradas, Phillips, and Lareu 2017).

This section will be focused on the bisulfite conversion pre-treatment and the SBE technique (commercially known under the name of SNaPshot) as these have been the chosen methods for the development of this project.

1.5.1 Bisulfite Conversion

The discovery of the bisulfite conversion dates back to 1970, when 4-thiouridine and sodium sulfite brought the formation of uridine-4-sulfonate in an oxygen mediated reaction (Hayatsu, Wataya, and Kai 1970). This led investigators to study the effect of sodium sulfite in other nucleosides and, in further studies, the nucleoside cytosine was shown to be deaminated into 5,6-dihydrouracil-6-sulfonate due to the effect of bisulfite in hot aqueous solution at pH 6 (Hayatsu et al. 1970).

Therefore, the application of this pre-treatment method to distinguish between unmethylated and methylated cytosines (Frommer et al. 1992) in CpG sites is relevant for forensic studies, specifically for age estimation. Further amplification of the bisulfite-converted products by PCR is performed using specific primer pairs specifically designed for that purpose.

On the one hand, this results in the amplification as thymines if the original cytosine was deaminated into uracil, meaning that the original cytosine was not methylated.

On the other hand, if the original cytosine was methylated, the sodium bisulfite is not able to deaminate this residue; therefore, during the amplification the nucleosides will remain as cytosines (Huang et al. 2010).

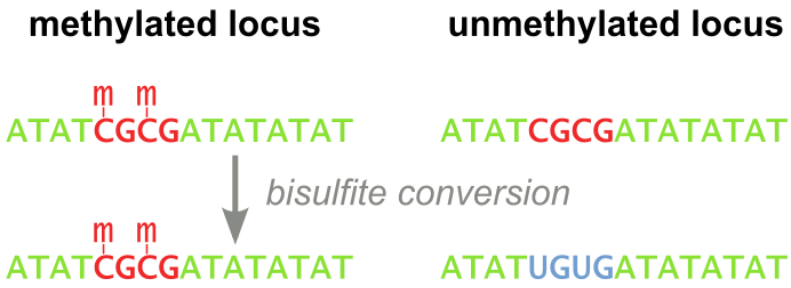

Figure 2. Molecular mechanism detail of bisulfite conversion in methylated and unmethylated locus.

1.5.2 Single Base Extension (SNaPshot technique)

The Single Base Extension (SBE) reaction consists on repeated annealing of a primer probe (forward and/or reverse) (figure 3) exactly one bp upstream of the target of the CpG site. During the reaction a unique fluorescent dideoxynucleotide (ddNTP) is added, therefore the proportion of the fluorescent signals can be measured by capillary electrophoresis (CE). The widest platform which supports the SBE technique is the ABI-SNaPshot (CE-SBE). This system allows the study of multiplexed reactions with primers of different size, so the study of many CpG dinucleotides is achieved in a single reaction. In CE, primers with the ddNTPs attached migrate according to their electrophoretic mobility; in the end the establishment of a fluorescent signal proportion is possible and the calculus of the methylation percentage is achieved (Kaminsky and Petronis 2009).

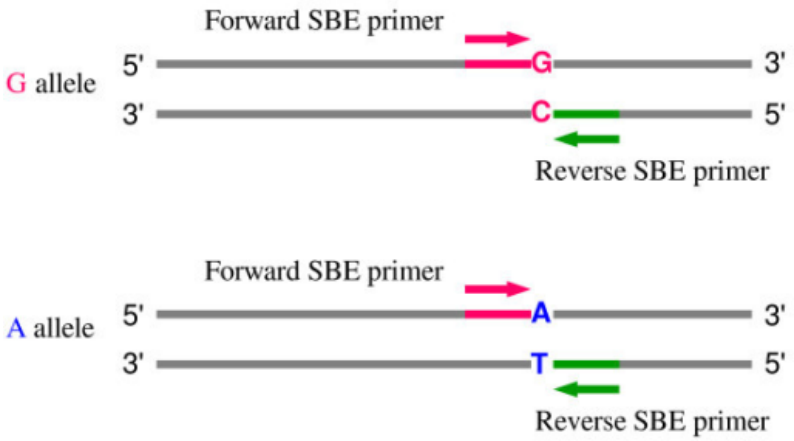

Figure 3. Annealing of Forward and Reverse probes during a Single Base Extension reaction (You et al. 2008).

1.6 Current Forensic Models for Age Estimation

At the present day, there are several methods applied for age estimation; however, the large number of CpGs considered in some of these models lead to the impossibility of their application in forensic genetics, a discipline in which samples are usually limited and even degraded. All of these age estimation models have been used to provide an age estimation according to the DNA methylation levels and the CpG sites studied in each case. During this section, the most determining methods will be reviewed, with special mention to the single method existing for semen samples.

In 2013, Horvath created the first forensic model for age prediction. This project consisted on the characterization of 353 CpG dinucleotides which have been linked to aging process. This model was able to provide an age estimation with a median absolute deviation (MAD) of 3.6 years regarding the chronological age (the real age of the suspect). Despite the fact that it was studied in 51 health tissues and cell types, the excessive number of CpGs comprised in this model has been the main triggering factor to avoid its application in forensic studies (Horvath 2013).

One year later, Weidner et al. developed initially a detailed analysis of 102 CpG sites in blood linked to aging obtaining a MAD of less than 5 years. Moreover, in the same study, the authors describe another approach based on only three 3 CpGs located in the genes *ASPA*, *ITGA2B* and *PDE4C*. This model has a MAD of just 5.4 years, a surprising fact taking into account the few CpGs considered in the development of this model (Weidner et al. 2014).

Further blood studies led to the emergence of another model by Zbieć-Piekarska et al., where authors were able to provide an age estimation model with a MAD of 3.9 years, by just considering 2 CpG sites from a single gene: the *ELOVL2*, one of the most important locus which provides information about chronological age in blood samples (Zbieć-Piekarska et al. 2015).

On the other hand, Lee et al. tested the accuracy of the previous models in 36 body fluids including blood, saliva and semen. Therefore, they selected age-related CpGs from semen profiles and through regression analysis, they were able to build a model with a MAD of approximately 5 years, and composed by only 3 CpGs: cg06304190 in the *TTC7B* gene, cg06979108 in the *NOX4* gene and cg12837463 (Lee et al. 2015). At the present time, this age estimation model is the only one existing for semen body fluid samples based on DNA methylation patterns.

Last but not least, it is relevant to highlight that the epigenetic marks by which all of these models are built may show dissimilarities bearing in mind the epigenetic variability according to the type of tissue (Armstrong et al. 2014). Individual characteristics could also be linked to raise the disparity of results due to differences in sex, ancestry and the occurrence of diseases during lifespan (Freire-Aradas, Phillips, and Lareu 2017).

2. Objectives

The age estimation by epigenetic signatures is a powerful tool which can help to solve forensic cases, especially in crimes in which a wide number of suspects is contemplated. In this particular case, the main aim of this project is to implement and ready the model created by Lee et al. in order to be able to provide an age estimation from semen samples obtained in crime scenes.

This task will be developed in the genetics department laboratories from the "Luís Concheiro" Forensic Sciences Institute (INCIFOR) located in Santiago de Compostela (Spain). Besides, another objective is to test if this model is applicable to semen samples from individuals with different ancestry and subjected to distinct environmental conditions than original samples tested in Lee et al.

3. Methodology

All techniques applied in this project have been tried to be as close as possible as the ones applied in the reference paper (Lee et al. 2015). However, some conditions have been modified according to indications provided by the experienced supervisor of this project.

Five freshly ejaculated samples were obtained from five individuals. Moreover, two different starting extraction volumes were considered for each of the samples independently: 50 µL and 300 µL. This way, it is possible to achieve two different concentrated end product for each sample. Finally, note that all thermocycler reactions described along the methodology have been performed in a GeneAmp® PCR System 9700. The general guideline followed through methodology can be found in figure 4.

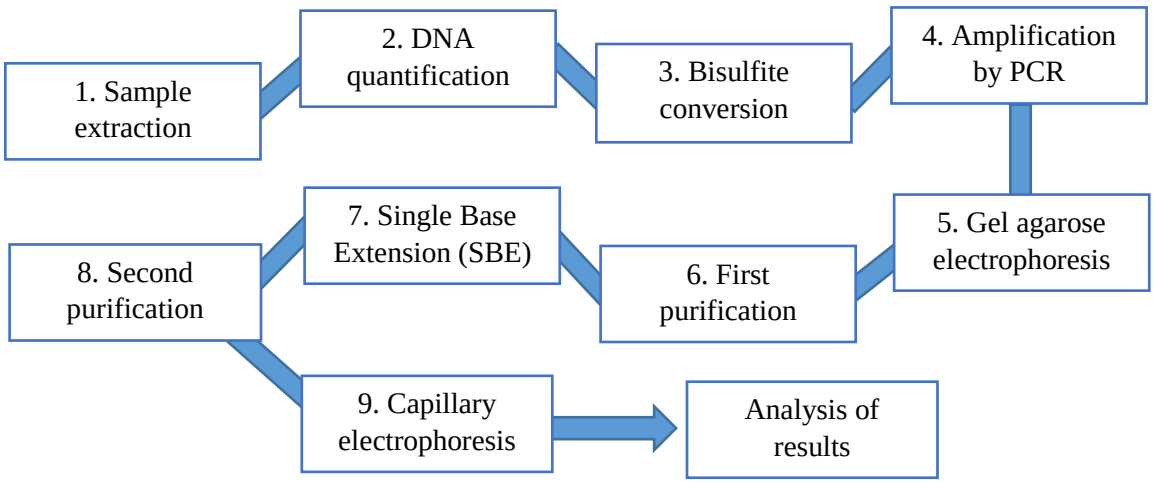

Figure 4. General scheme of the methodology applied during the development of this project.

3.1 Sample extraction: Phenol/Chloroform

The extraction method applied is based on a phenol/chloroform guideline reported for genomic DNA extraction from semen samples (Anvar et al. 2015).

The steps followed are briefly described hereunder:

1) Centrifuge the 50 µL and 300 µL samples at 13200 rpm (max. speed) 10 min to obtain the cellular pellet. Discard the supernatant.
2) Add to the pellet 500 µL of lysis buffer (10 mM TRIS HCl (pH 8.0), 25 mM EDTA, 1% SDS and 75 mM NaCl) + 21 µL DTT + 10 µL Proteinase K.

3) Vortex gently and incubate overnight with agitation at 50ºC.
4) Centrifuge at max. speed for 5 min. Transfer supernatant to a new tube and discard the pellet.
5) Add the same supernatant volume (~400 µL) of phenol/chloroform/isoamyl alcohol (25:24:1)
6) Mix by inversion and centrifuge at max. speed for 5 min to separate DNA from proteins.
7) Transfer supernatant to a new tube and add 1/10 of volume of 3 M Sodium acetate (NaAc) + 1 µL of glycogen + 2 volumes of EtOH 100%. Respect the mentioned sequence for adding the reagents.
8) Mix by inversion and incubate for 1 hour at -80ºC.
9) Centrifuge at max speed at 4ºC during 15 min to precipitate DNA. Discard supernatant.
10) Wash DNA pellet with 70% cold EtOH (add same volume as EtOH 100% in step 7).
11) Centrifuge at max. speed for 5 min. Aspirate EtOH completely without contacting the pellet.
12) Open tubes and let air dry for 2h. Resuspend the pellet in 50 µL of distilled water.
13) Incubate samples at 56ºC for 2h under agitation. Hence, samples can be stored long term periods at -20 ºC.

3.2 Quantification: Qubit 3 Fluorometer

The Invitrogen Qubit 3 allows a fast quantitation of DNA, RNA and protein samples in less than 5 seconds per specimen. It provides a high accuracy by using only 1-20 µL of sample, even with diluted samples, an interesting key point taking into account the often low quantity of DNA in forensic samples.

The Qubit 3 is fitted with 2 assay types for DNA: dsDNA High sensitivity (for low concentration or degraded samples) and dsDNA Broad Range. For this project, and bearing in mind the great concentration of DNA in sperm, the Broad Range will be the selected analysis. The procedure is shortly described below:

1) Set up two assay tubes for the Broad Range standards and one more for each sample.

2) Prepare the working solution by mixing 199 μL of Qubit ® buffer and 1 μL of Qubit ® reagent for each tube.
3) Prepare the tubes according table 1:

Table 1. Assay tubes preparation for Qubit 3 Fluorometer.

	Standard tubes	**Sample tubes**
Volume of working solution	190 μL	198 μL
Volume of Standard	10 μL	-
Volume of Sample	-	2 μL
Total volume	200 μL	200 μL

4) Vortex tubes during 2-3 sec using a glove to prevent direct contact of them with the device; this could lead to failure or mistakes in the quantification.
5) Incubate tubes for 2 min at room temperature.
6) Select Broad Range assay and insert tubes in the Qubit ® fluorometer as device indicates. First, standard tubes should be quantified and then, the samples.

3.3 Bisulfite Conversion: EpiTect® Fast DNA Bisulfite Kit

EpiTect Fast DNA Bisulfite Conversion Kits provide an accurate conversion and purification of DNA prepared from FFPE (Formalin-fixed paraffin embedded), blood, cell and tissue samples. The high concentration Bisulfite Solution from the kit, reduces the time needed to convert unmethylated cytosines into uracil from 5 hour to 30 minutes. Additional DNA protect Buffer, a pH indicator dye, prevents DNA fragmentation.

This kit allows the user to work with high (1μg-2μg) or low-concentrations (1-500 ng); for the present case, since quantified samples have shown values between 100-400 ng/μL, the low-concentration method has been chosen. The amount of DNA considered for the bisulfite conversion has been 300 ng.

The main steps of this kit are provided here below:

1) Add in PCR tubes the following reagents as listed in Table 2.

Table 2. Reagents and Samples for High and Low Concentration Bisulfite Conversions.

Bisulfite reaction components	**High-concentration (1-2µg)**	**Low-concentration (1-500 ng)**
DNA	Variable (max. 20 µL)	Variable (max. 40 µL)
RNase-free water	Variable (20 µL – Vol. DNA)	Variable (40 – Vol. DNA)
Bisulfite Solution	85 µL	85 µL
DNA Protect Buffer	35 µL	15 µL
Total Volume	140 µL	140 µL

2) Mix completely the tubes until DNA Protect Buffer turns into blue colour.
3) Incubate in thermocycler under conditions exposed in table 3.

Table 3. Thermocycler conditions for Bisulfite Conversion.

Step	**Time**	**Temperature**
Denaturation	5 min	95ºC
Incubation	20 min	60ºC
Denaturation	5 min	95ºC
Incubation	20 min	60ºC
Hold	Infinite (∞)	20ºC

Desulfonation and Washing

4) Transfer volume from PCR tubes to 1.5 mL Tubes.
5) Add 310 µL of Buffer BL. Vortex and Spin samples.
6) Add 250 µL of absolute EtOH. Vortex for 15 sec. Spin to remove bubbles.
7) Transfer volume to MinElute DNA columns.
8) Centrifuge at max. speed for 1 min. Empty the collector tube.
9) Add 500 µL of Buffer BW. Repeat step 8.
10) Add 500 µL of Buffer BD, close the tubes and incubate at room temperature for 15 min. Repeat step 8.
11) Repeat step 9 two times.
12) Add 250 µL of absolute EtOH. Repeat step 8.

13) Transfer the column to a new collector tube and centrifuge under the same conditions to remove the EtOH remains.
14) Place the column in a new elution tube. Add 15 µL of Buffer EB straight in the middle of the membrane.
15) Incubate at room temperature for 1 min. Centrifuge at max. speed for 1 min. The converted samples can be kept between 2-8ºC for 24h. For longer periods, store samples frozen at -20ºC.

3.4 Amplification by Polymerase Chain Reaction (PCR)

The initial chosen concentration of primers has been 25 µM, specific for the amplification of the three CpG dinucleotides located in *TTC7B*, *NoGene* and *NOX4* loci respectively (considering primers pair Fw/Rv for each CpG), see table 9.

Further dilutions need to be done, therefore PCR mixes are prepared in order to achieve the adequate proportion of primer for each CpG site.

The assays performed indicating the proportion variations within these three primers will be described in the results section. The final PCR conditions which have been suitable for multiplexed assays along this project are shown in tables 4 and 5.

Table 4. PCR content for 1 tube.

Reagent	Volume for 1 tube
Buffer 10x	1.5 µL
BSA	1.5 µL
$MgCl_2$	3.9 µL
Primer Mix*	4 µL
dNTPs	1 µL
TaqGold Polymerase	0.3 µL
DNA sample	3 µL

*Note that, despite introducing variations in primers proportions, the final volume of primer mix added is 4 µL in any case.

Table 5. Thermocycler conditions for PCR.

	34 cycles				
95°C	**94°C**	**56°C**	**72°C**	**72°C**	**8°C**
11m	20s	60s	30s	7m	Infinite (∞)

Note: Additional agarose gel electrophoresis step is recommended to ensure that the amplification has been successful.

3.5 First purification: ExoSAP-IT Reagent

The ExoSAP-IT™ PCR Product Cleanup Reagent is a one-step solution used for the cleanup of PCR amplified product. It removes the excess of primers and dNTPs enzymatically. For that, add in new tubes 2.5 μL of PCR product + 1 μL of ExoSAP-IT reagent. The thermocycler conditions are described in table 6.

Table 6. Thermocycler conditions for ExoSAP-IT purification step.

37 ºC	**80ºC**	**8ºC**
45 min	15 min	Infinite (∞)

3.6 Single Base Extension (SBE): SNaPshot Technique

Following the same guideline as in the primer mix preparation for PCR amplification, aliquots are prepared in order to avoid prolonged contact with comercial tubes; on this occasion the starting concentration has been considered of 50 μM and further SBE mixes are prepared according to the required proportion for each CpG site (see results section for more information). In this case, probes are only reverse sense for the three CpGs (located in *TTC7B*, *NoGene* and *NOX4* loci), see table 10.

To perform this reaction, it is necessary to add in new tubes for each sample, 2.5 μL SNaPshot reagent + 1.5 μL SBE mix (probes) + 2 μL of ExoSAP-IT purified product. Thermocycler conditions are further shown in table 7.

Note that, despite introducing variations in probes proportions, the final volume of SBE mix added is 1.5 µL in any case.

Table 7. Thermocycler conditions for SNaPshot technique.

30 cycles			
96 °C	**55 °C**	**60 °C**	**8 °C**
10 sec	5 sec	30 sec	Infinite (∞)

3.7 Second purification: SAP Reagent

Shrimp Alkaline Phosphatase (SAP) is a high activity, heat-labile alkaline phosphatase isolated from *Pandalus borealis* (arctic shrimp) which removes 5'-phosphates from DNA, RNA, dNTPs, and proteins. It is highly indicated for the purification of unincorporated ddNTPs prior to DNA sequencing or SNP analysis. Hence, it is needed to add 1 µL of SAP to the whole product of SNaPshot (6 µL) for each sample. The thermocycler conditions are described in table 8.

Table 8. Thermocycler conditions for SAP purification step.

37 ºC	**85ºC**	**8ºC**
80 min	15 min	Infinite (∞)

3.8 Capillary Electrophoresis: 3130xl Genetic Analyzer

The ABI PRISM® 3130xl Genetic Analyzer from Applied Biosystems™ is a 16-capillary electrophoresis instrument for medium throughput research laboratories.

To perform the sequencing reaction, 96-well plates have been used. First of all a loading mix needs to be prepared; it contains (for each sample) 10 µL of formamide (used to unfold DNA molecules and allow them to migrate and separate according only to their size and not because of the structure) and 0.1 µL of Liz120 (the internal reference standard which contains well known length DNA fragments). Besides, for each well, 9.5 µL of loading mix + 1.5 µL of purified sample have to be added. Then, the plate is ready to be analysed. The results are visualized using the GeneMapper® ID v3.2.1 software.

Figure 5 provides a general view of the main molecular mechanisms related to the bisulfite conversion, PCR amplification and Single Base Extension steps of the methodology. Further information about complete nucleotide sequences of PCR primers and reverse probes can be found in tables 9 and 10.

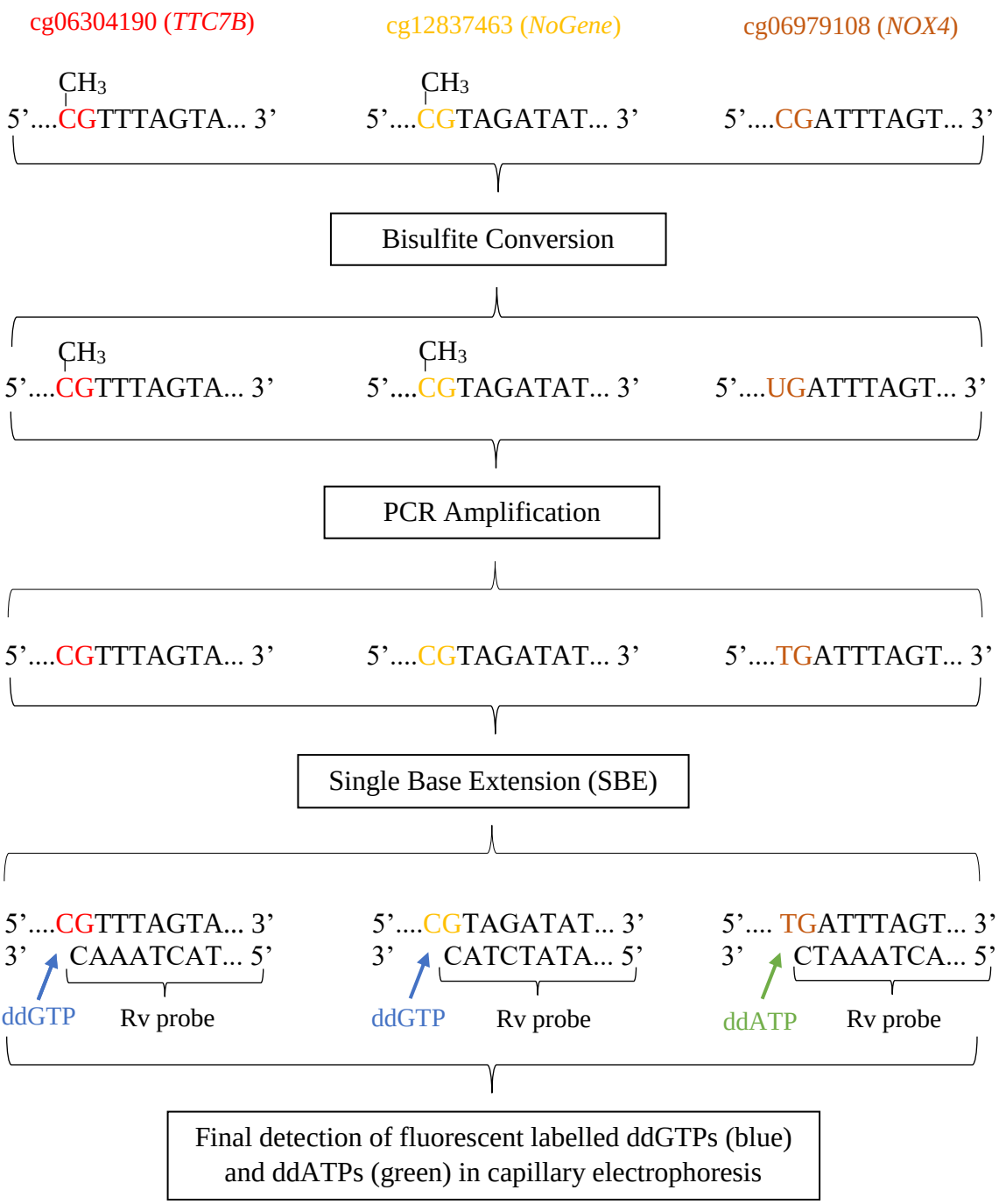

Figure 5. Molecular insight for Bisulfite Conversion, PCR amplification and SBE assays.

Note that, in figure 5 as well, to ease the comprehension, only plus strands (5'→3') are displayed due to the fact that probes used have been only reverse (3'→5') sense. However, during PCR amplification, copies of both strands have been performed due to the addition of primer pairs (Fw and Rv) for each CpG site. Also, it must be taken into account that, just to give an example, *TTC7B* and *NoGene* have been considered as methylated loci while *NOX4* has been chosen as an unmethylated locus. In order to avoid confusions, remark also that reactions for the three CpGs occur simultaneously since it is a multiplex assay.

Table 9. Primers sequences used in PCR amplification for the three CpG sites (Lee et al. 2015).

PCR amplification			
Target ID	**Gene Locus**	**Primer Fw (5'→3')** **Primer Rv (5'→3')**	**Amplicon size (bp)**
cg06304190	*TTC7B*	F: AATTTTATTTTTGGTATTTAAAGTAG R: AAACAAAAACTACCACTCTCACAC	195
cg12837463	*NoGene*	F: AGTTGGTATTAGGGTTTGAAATGTA R: TCTCAAAAACTCTACAATAAAAAAAA	209
cg06979108	*NOX4*	F: TAGTTATTTGAGTGAAGTGTGTTGG R: ACCTCCCAAAATACTAAATTACTC	194

Table 10. Probes sequences used in Single Base Extension (SBE) for the three CpG sites (Lee et al. 2015).

Single Base Extension (SBE)			
Target ID	**Gene Locus**	**Probe Rv (5'→3')**	**Length (nt)**
cg06304190	*TTC7B*	R: AATAATCACCTACTATATACTAAAC	25
cg12837463	*NoGene*	R: CCTTCTTTAACTCATATACTTTAAAAATATCTAC	34
cg06979108	*NOX4*	R: TCAATTAAATCCTCAACTAAATC	23

4. Results and Discussion

4.1 Protocol optimization

One of the main objectives of this project is to implement and ready the protocol applied in the reference paper (Lee et al. 2015). That is why multiple assays have been performed until best conditions have been achieved. All experiments with key variations in the conditions can be briefly visualized in table 11.

Table 11. Changing conditions performed along different assays.

Assay Number	**Assay Type**	**PCR Primer Mix (from 25 µM alicuots)**	**SBE Probe Mix (from 50 µM alicuots)**	**Annealing Temperature in SBE method**	**Capillary Electrophoresis Polymer Type**
1	Singleplex	No Primer Mix. 1/30 dilutions for each primer performed in separate tubes.	No SBE Mix. 1/30 dilutions for each probe performed in separate tubes.	52 °C	POP-7™
2	Multiplex	2 µL *TTC7B* + 4 µL *NoGene* + 2 µL *NOX4*. Final volume of 80 µL.	2 µL *TTC7B* + 4 µL *NoGene* + 2 µL *NOX4*. Final volume of 80 µL.	52 °C	POP-7™
3	Multiplex	2 µL *TTC7B* + 2 µL *NoGene* + 2 µL *NOX4*. Final volume of 80 µL.	2 µL *TTC7B* + 2 µL *NoGene* + 2 µL *NOX4*. Final volume of 80 µL.	52 °C	POP-4®
4	Multiplex	2 µL *TTC7B* + 0.75 µL *NoGene* + 2 µL *NOX4*. Final volume of 100 µL.	2 µL *TTC7B* + 1 µL *NoGene* + 2 µL *NOX4*. Final volume of 100 µL.	52 °C & 55 °C	POP-4®

5	Multiplex	2 µL *TTC7B* + 1.25 µL *NoGene* + 2 µL *NOX4*. Final volume of 150 µL.	2 µL *TTC7B* + 1 µL *NoGene* + 2 µL *NOX4*. Final volume of 150 µL.	55 °C	POP-4®
6	Multiplex	2 µL *TTC7B* + 1.25 µL *NoGene* + 2 µL *NOX4*. Final volume of 150 µL.	2 µL *TTC7B* + 1 µL *NoGene* + 2 µL *NOX4*. Final volume of 150 µL.	55 °C	POP-4®

According to Table 11, the first singleplex assay was performed in order to test if commercial primers and probes were working properly. This way, miscarriages can be identified easily due to the independence of the reactions performed in samples.

Once the singleplex assay was proven in a successfully way, multiplex assays were executed. Consecutively, the first multiplex (assay 2) lead to two crucial statements which can be visualized in figure 6:

- The signal obtained in capillary electrophoresis is too high (saturated) for the CpG site associated with the *NoGene* locus. This feature can be seen as a vertical pink bar in the figure mentioned; in detail, saturation leads to poor-defined peaks and also to the emergence of new unexpected ones such as the small peak in the T lane. This, in fact, is due to that the wavelength (λ) at which the saturated nucleotide is analysed intercepts with λ of other nucleotides, giving rise to false positive peaks or intensifying their signal.

- The polymer POP-7™ is not suitable to differentiate between the CpGs located in the *TTC7B* and *NOX4* loci; both peaks overlap, therefore being impossible to obtain a reliable peak height measurement later used for the calculations of the %GC.

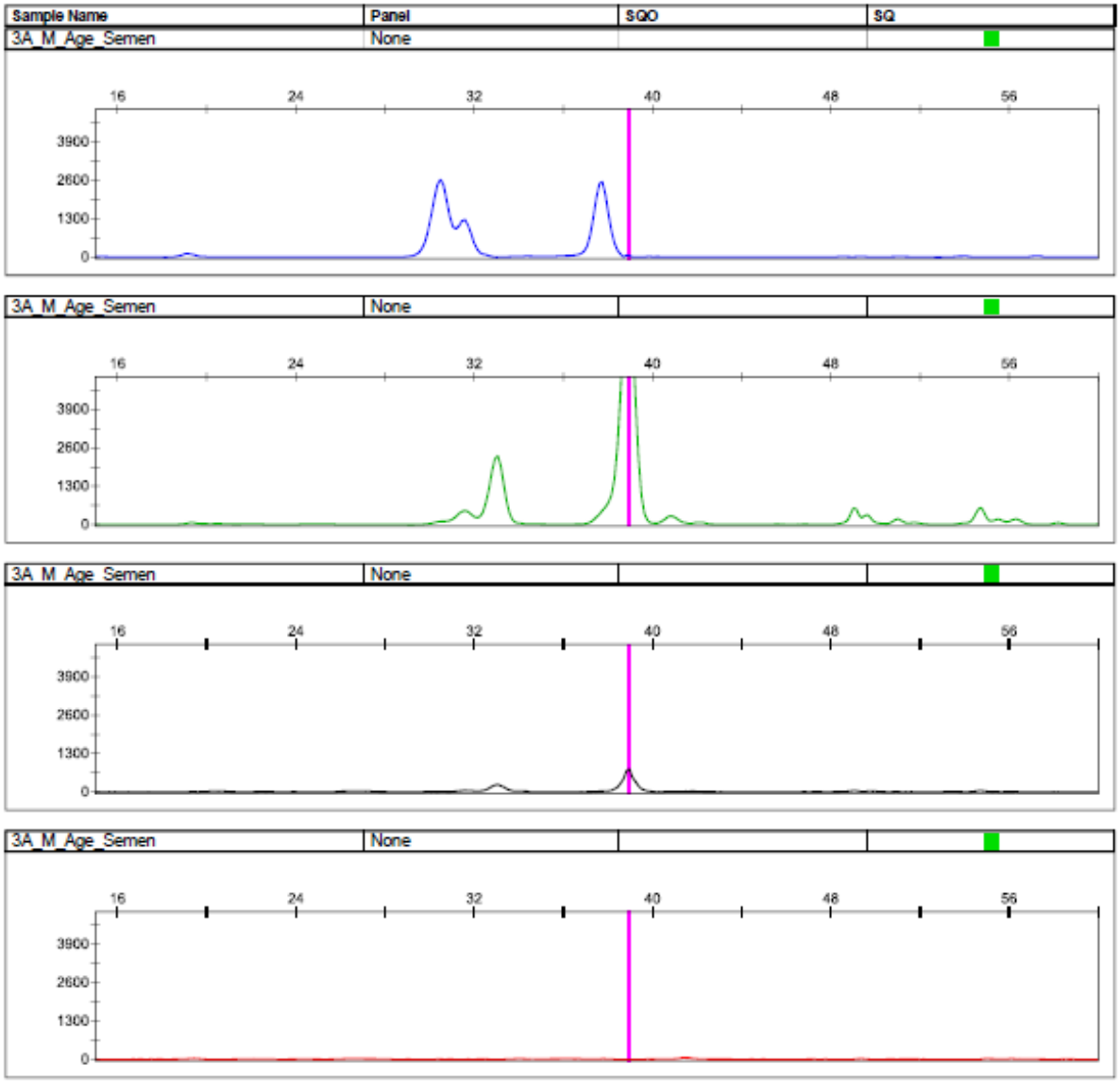

Figure 6. Electropherogram of a semen sample analysed in assay 2 (first multiplex) according to table 1. The lanes match to G (blue), A (green), T (black) and C (red) nucleotides respectively. The X-axis represent the base pairs (bp) while the Y-axis means the Relative Fluorescent Units (RFUs).

Note that the first peak in figure 6 correspond to the *NOX4* CpG (at 24 bp), the second to the *TTC7B* CpG (at 26 bp) and the third to the *NoGene* CpG (at 35 bp). However, even real length of fragments (probes + ddNTP added) are 24, 26 and 35 bp respectively, the observed peaks are a bit shifted to higher bp positions. Moreover, it is relevant to highlight that only G and A signals are expected due to the fact that probes added in the multiplex SNaPshot reaction are reverse sense (if probes were oriented forwardly peaks should be observed in T and C lanes). For that, the right false-positive peak observed in T lane (black) is due to the high intensity of the signal existing in the A lane (green).

Coming back to Table 11, and bearing in mind the results obtained in the first multiplex assay, primers and probes dilutions were adjusted in the following multiplexes until optimal conditions were achieved (2 µL *TTC7B* + 1.25 µL *NoGene* + 2 µL *NOX4* for PCR primers and 2 µL *TTC7B* + 1 µL *NoGene* + 2 µL *NOX4* for SBE probes mix. In both cases, a final volume of 150 µL was used).

Moreover, annealing temperature for probes in SNaPshot reaction were finally optimized to 55 °C instead of the 52 °C as considered initially. This temperature modification may have been one of the most important triggering factors to attain valuable results.

Last but not least, polymer POP-7™ is considered as one of the most versatile polymers for long and short fragments length sequencing. Nonetheless, the analysis with POP-4® polymer, designed specifically for forensic applications, has been able to solve the overlapping issue between *NOX4* and *TTC7B* CpG dinucleotides. Also, note that observed length of fragments are closer to real sizes under POP-4® analysis (24, 26 and 35 pb for *NOX4*, *TTC7B*, and *NoGene* associated respectively) (figure 7).

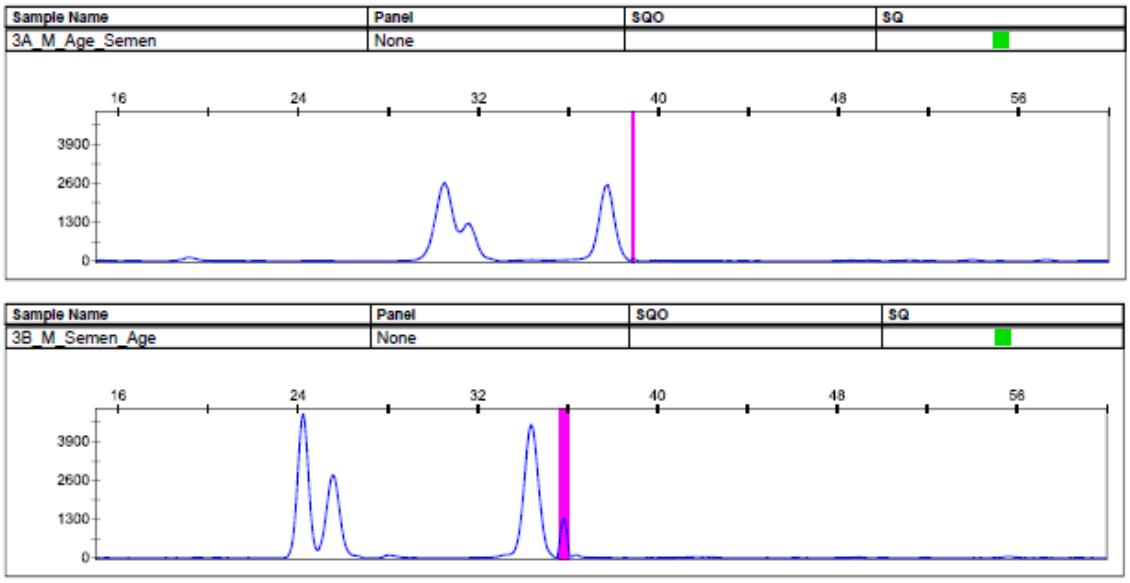

Figure 7. Electropherogram G peaks (in blue) of the three studied CpGs in two semen samples from the same individual analysed with POP-7™ (top) and POP-4® (bottom). Again, and as stated before, the pink vertical bar refers to a saturation which comes from an adenine signal (herein not displayed); that is why this small peak which appears at almost 35 bp in the lower electropherogram is not a real signal.

4.2 Age estimation through DNA methylation

In the manuscript of Lee et al., two regression models were developed: one built from a training set samples (31 semen samples) and the second from the total analysed samples (68 semen samples= 31 semen samples from training set + 37 semen samples from testing set). However, estimated age is calculated following the same equation structure in both models:

$$\text{Predicted age} = \beta_0 + \beta_1 \text{ x } \text{CpG}_1 + \beta_2 \text{ x } \text{CpG}_2 + \beta_3 \text{ x } \text{CpG}_3$$

In which:

- β_0 represents the intercept.
- $\beta_{1,2,3}$ represent estimate values (coefficients) in table 10.
- $\text{CpG}_{1,2,3}$ represent methylation values in % at each CpG site as obtained from the peak height measurement in electropherograms.

Table 12. Estimate values (coefficients) for each CpG site according to two different samples set (Lee et al. 2015).

Samples Set	CpG Name	Locus Gene Name	Estimate Value
Training (n=31)	(Intercept)	-	74.153
Training (n=31)	cg06304190	*TTC7B*	-0.460
Training (n=31)	cg12837463	*NoGene*	-0.353
Training (n=31)	cg06979108	*NOX4*	0.304
Total (n=68)	(Intercept)	-	55.357
Total (n=68)	cg06304190	*TTC7B*	-0.471
Total (n=68)	cg12837463	*NoGene*	-0.269
Total (n=68)	cg06979108	*NOX4*	0.491

Along the present project, an age estimation approach for 4 individuals has been done. It has not been possible to analyse the remaining one due to a bad quality of the sperm and/or technical issues during the extraction of the sample.

For this purpose, figures 8, 9, 10, 11 and table 13 provide significant information about electropherograms and age estimation calculus for the four individuals.

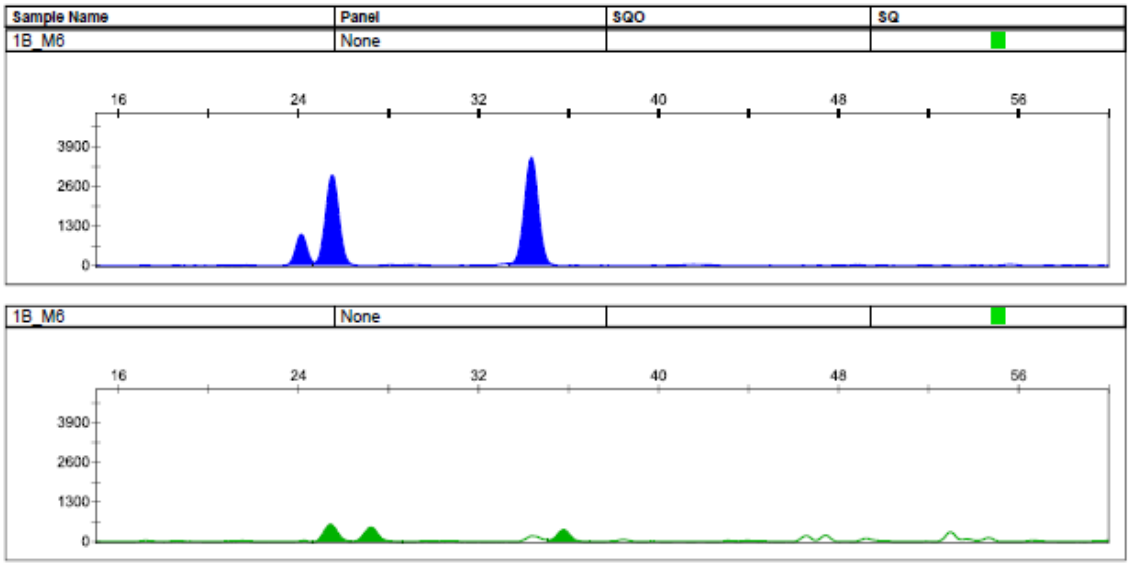

Figure 8. Guanine (blue) and Adenine (green) electropherogram lanes applicable to the semen sample from individual 1.

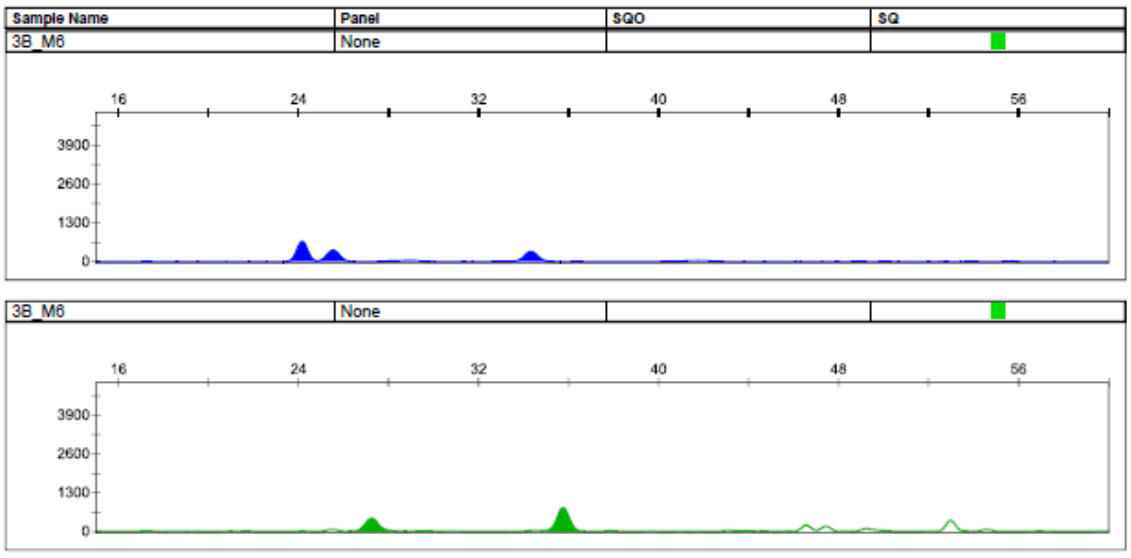

Figure 9. Guanine (blue) and Adenine (green) electropherogram lanes applicable to the semen sample from individual 2.

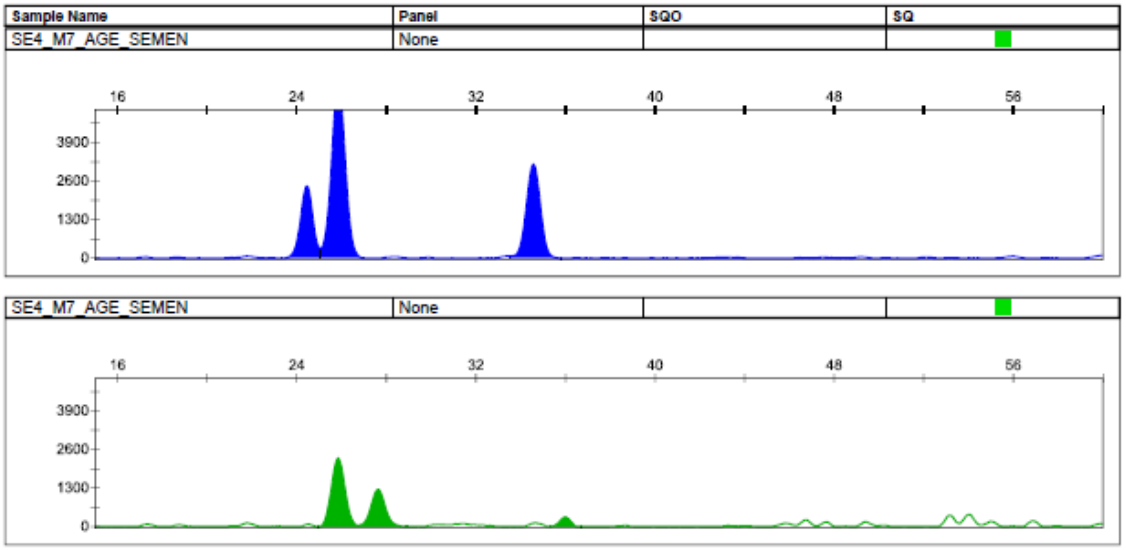

Figure 10. Guanine (blue) and Adenine (green) electropherogram lanes applicable to the semen sample from individual 3.

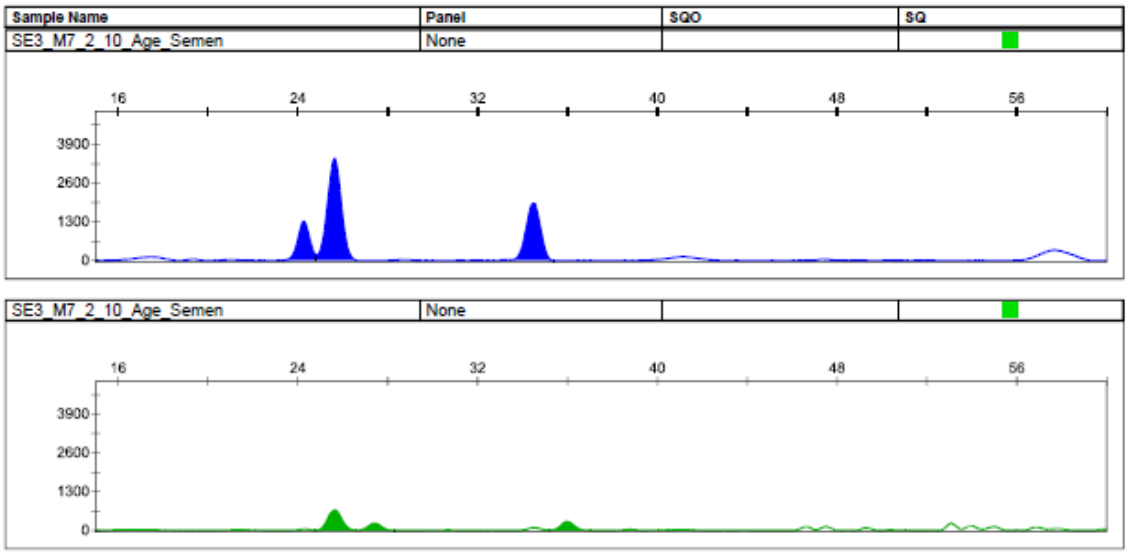

Figure 11. Guanine (blue) and Adenine (green) electropherogram lanes applicable to the semen sample from individual 4.

Table 13. Age estimation of semen samples from each individual according the n=31 (training set of samples) and n=68 (total set of samples: training + testing) regression models by Lee et al.

Individual	CpG site locus	Peak Height (URFs)		Methylation Percentage (%)	Predicted Age (by n=31 model) (in years)	Predicted Age (by n=68 model) (in years)	Chronological (Real) Age (in years)
		G	A				
1	*TCC7B*	2977	468	86.42	22.17	21.99	22
	NoGene	3561	398	89.95			
	NOX4	1028	573	64.21			
2	*TCC7B*	391	438	47.17	69.56	69.75	64
	NoGene	340	799	29.85			
	NOX4	671	67	90.92			
3	*TCC7B*	5922	1258	82.48	19.72	17.18	22
	NoGene	3184	327	90.69			
	NOX4	2425	2325	51.05			
4	*TCC7B*	3436	237	93.55	20.50	20.24	24
	NoGene	1925	296	86.67			
	NOX4	1325	692	65.69			

As shown above (table 13), age estimation of the four individuals has been considerably successful. In detail, for individual 1, there has been a close match between the predicted age in both regression models (22.17 and 21.99 years) compared to the real age (22 years).

Regarding individuals 3 and 4, closely related to individual 1 in age, the age estimation has also been accurate enough taking into account the MAD (median absolute deviation) assumed by Lee et al. in the development of the two regression models (for the training set (n=31), the MAD from chronological age was defined in 4.2 years while for the total set of samples (n=68), the MAD was 4.7 years). However, it is relevant to note that the predicted age for individual 3 is slightly out of the established MAD (according to the

total set of samples model) by the reference paper authors. Actually, for individual 3, the age estimation difference between the two regression models is considerably wider than in the rest of individuals.

Finally, focusing on the results obtained in the age prediction for individual 2, the MAD has been exceeded a little bit for both regression models. This, actually, can be due to the low signal obtained in the electropherogram (Figure 9), a fact which may be linked to the decrease of the accuracy in both of the models; as well as the reduction of DNA methylation in aging could also be a contributing factor related to the lack of precision.

Additionally, Lee et al. observed frequently high deviations in individuals aged more than 60 years; they attribute these variations for aspects of biological age. As exposed in table 13, chronological age is calculated; this term refers to the total amount of years which have elapsed since individuals are born. On the other hand, biological age relates to the systems, tissues and cells age in relation to common patterns; therefore, the biological age is a physiological concept and does not depend on timeline: it is strongly affected by environmental or genetic conditions and circumstances. In young individuals, the biological age often matches closely to the chronological age, but in elder individuals these differences may play a key role in the age determination techniques and models.

That is why, independent models should be performed in order to discern among individuals with large age differences. Nevertheless, Lee et al. were not able to test separately these two groups of individuals because of the small size of samples considered in their project.

5. Conclusions

The raising interest in epigenetics last few years has led to the emergence of interesting techniques and age estimation methods in the forensic genetics context. Being able to provide an age approach can be a firm step forward in terms of solving a crime in which many suspects are considered or also in humanitarian civil cases. The two regression models proposed by Lee et al. for semen samples are the unique insights for the study of DNA methylation in semen samples. In this pioneer article, authors reviewed other regression models and were able to provide this two specific models by using only three CpG sites (cg06304190 in the *TTC7B* gene, cg06979108 in the *NOX4* gene and cg12837463) which have been linked to age, specifically in sperm.

During the present study, 4 semen samples have been successfully analysed; it has been able to provide an age estimation within the deviation values obtained by Lee et al. in the development of their regression models (4.2 and 4.7 years MAD respectively). However, in one of these samples, coming from a 64 year old individual, the deviation was considerably high in relation with other samples from young individuals. That is why, further studies and new models according to the age group should be developed to provide a closer estimation between the chronological age ('timeline age') and the biological ('cellular age'), being the second one highly influenced by external and genetic conditions specially for advanced aged individuals. Note that the lack of a higher amount of samples for the development of the current project has made impossible the application of statistical analysis and the creation of mathematical regression models for age estimation.

On the other hand, it is important to emphasize that the establishment of the conditions applied in techniques is a key event so compelling results are achieved. Therefore, previous knowledge in forensic cases and protocol optimization by experienced people in the field is a great support to obtain this objective.

In summary, the validation of Lee et al. regression models have been satisfactorily accomplished in samples of different origin than the original published work. Therefore, the high versatility of this age predicting models have been proved. Nevertheless, precise and deeper studies are needed in order to acquire wider and trusty validation. For instance, large scale studies could help to develop complementary regression models which may bear in mind the geographical region or even ancestry origin of individuals.

Last but not least, recent investigations are focusing in the study of other CpG sites such as two positions located in *ACCN4* and *PLEC* loci respectively. This two dinucleotides, in spite of not being related to aging, show fully methylated patterns in sperm genomic DNA while in other fluids like blood the methylation percentage decreases remarkably. Therefore, they can be useful to discern among semen samples, mixtures and other sample types. This means that, by combining the analysis of just 5 CpGs (*TTC7B*, *NoGene*, *NOX4* as reported during this project, in addition to *ACCN4* and *PLEC4*) in a single multiplex assay, it may be possible to provide an age estimation of the individual and also, check with a high reliability the sample type. Hence, a great step forward to solve difficult forensic cases, especially those related to rapes or sexual assaults, is being undertaken.

6. References

Alam, Fahmida et al. 2016. 'DNA Methylation: An Epigenetic Insight into Type 2 Diabetes Mellitus.' *Current pharmaceutical design* 22(28): 4398–4419.

Alkass, Kanar et al. 2010. 'Age Estimation in Forensic Sciences: Application of Combined Aspartic Acid Racemization and Radiocarbon Analysis.' *Molecular & cellular proteomics : MCP* 9(5): 1022–30.

Allah, Rakha, Li Yang, and Sheng-bin Li. 2007. 'SNPs and Forensic DNA Typing.' *Fa yi xue za zhi* 23(5): 373–79.

Anvar, Zahra, Bahia Namavar-Jahromi, Samaneh Ebrahimi, and Behrouz Gharesi-Fard. 2015. 'Genomic DNA Extraction from Sperm'. *Journal of Advanced Medical Sciences and Applied Technologies (JAMSAT)* 1(1).

Arany, Szilvia, Susumu Ohtani, Naofumi Yoshioka, and Kunio Gonmori. 2004. 'Age Estimation from Aspartic Acid Racemization of Root Dentin by Internal Standard Method'. *Forensic Science International* 141(2–3): 127–30.

Armstrong, David A. et al. 2014. 'Global and Gene-Specific DNA Methylation across Multiple Tissues in Early Infancy: Implications for Children's Health Research'. *The FASEB Journal* 28(5): 2088–97.

Blanchetot, A, V Wilson, D Wood, and A J Jeffreys. 1983. 'The Seal Myoglobin Gene: An Unusually Long Globin Gene.' *Nature* 301(5902): 732–34.

Brookes, A J. 1999. 'The Essence of SNPs.' *Gene* 234(2): 177–86.

Butler, John M. 2004. 'Short Tandem Repeat Analysis for Human Identity Testing'. In *Current Protocols in Human Genetics*, Hoboken, NJ, USA: John Wiley & Sons, Inc.

———. 2006. 'Genetics and Genomics of Core Short Tandem Repeat Loci Used in Human Identity Testing'. *Journal of Forensic Sciences* 51(2): 253–65.

Carracedo, Angel. 2013. Encyclopedia of Forensic Sciences *Forensic Genetics: History*. 2nd ed. Elsevier Ltd.

Carracedo, Angel, Antonio Salas, and M V Lareu. 2010. 'Problemas y Retos de Futuro de La Genética Forense En El Siglo XXI'. *Cuadernos de Medicina Forense* 16: 31–35.

Cavrić, Jelena, Marin Vodanović, Ana Marušić, and Ivan Galić. 2016. 'Time of Mineralization of Permanent Teeth in Children and Adolescents in Gaborone, Botswana'. *Annals of Anatomy - Anatomischer Anzeiger* 203: 24–32.

Chomyn, Anne, and Giuseppe Attardi. 2003. 'MtDNA Mutations in Aging and Apoptosis.' *Biochemical and biophysical research communications* 304(3): 519–29.

D'Aquila, Patrizia, Giuseppina Rose, Dina Bellizzi, and Giuseppe Passarino. 2013. 'Epigenetics and Aging'. *Maturitas* 74(2): 130–36.

Derek K Lim, Authors H, Eamonn R Maher, Derek H K Lim DCH MRCPCH, and Eamonn R Maher FRCP FMedSci Professor. 2010. 'SAC Review DNA Methylation: A Form of Epigenetic Control of Gene Expression'. *SAC review* 1212: 37–4237.

Fan, Hao, and Jia-You Chu. 2007. 'A Brief Review of Short Tandem Repeat Mutation.' *Genomics, proteomics & bioinformatics* 5(1): 7–14.

Freire-Aradas, A, C Phillips, and M V Lareu. 2017. 'Forensic Individual Age Estimation with DNA: From Initial Approaches to Methylation Tests.' *Forensic science review* 29(2): 121–44.

Frommer, M et al. 1992. 'A Genomic Sequencing Protocol That Yields a Positive Display of 5-Methylcytosine Residues in Individual DNA Strands.' *Proceedings of the National Academy of Sciences of the United States of America* 89(5): 1827–31.

Gill, P. 2001. 'An Assessment of the Utility of Single Nucleotide Polymorphisms (SNPs) for Forensic Purposes.' *International journal of legal medicine* 114(4–5): 204–10.

Gill, P, A J Jeffreys, and D J Werrett. 1985. 'Forensic Application of DNA "Fingerprints".' *Nature* 318(6046): 577–79.

Goodwin, William. 2007. 'Introduction to Forensic Genetics'. In *An Introduction to Forensic Genetics*, , 1–5.

Guillaumet-Adkins, Amy et al. 2017. 'Epigenetics and Oxidative Stress in Aging'. *Oxidative Medicine and Cellular Longevity* 2017: 1–8.

Hayatsu, Hikoya., Yusuke. Wataya, and Kazushige. Kai. 1970. 'Addition of Sodium Bisulfite to Uracil and to Cytosine'. *Journal of the American Chemical Society* 92(3): 724–26.

Hayatsu, Hikoya, Yusuke Wataya, Kazushige Kai, and Shigeru Iida. 1970. 'Reaction of Sodium Bisulfite with Uracil, Cytosine, and Their Derivatives'. *Biochemistry* 9(14): 2858–65.

Heyn, H. et al. 2012. 'Distinct DNA Methylomes of Newborns and Centenarians'. *Proceedings of the National Academy of Sciences* 109(26): 10522–27.

Horvath, Steve. 2013. 'DNA Methylation Age of Human Tissues and Cell Types'. *Genome Biology* 14(10): R115.

Huang, Yun et al. 2010. 'The Behaviour of 5-Hydroxymethylcytosine in Bisulfite Sequencing.' *PloS one* 5(1): e8888.

Jiang, H., Z. Ju, and K. L. Rudolph. 2007. 'Telomere Shortening and Ageing'. *Zeitschrift für Gerontologie und Geriatrie* 40(5): 314–24.

Jin, Bilian, Yajun Li, and Keith D Robertson. 2011. 'DNA Methylation: Superior or Subordinate in the Epigenetic Hierarchy?' *Genes & cancer* 2(6): 607–17.

Kaminsky, Zachary, and Arturas Petronis. 2009. 'Methylation SNaPshot: A Method for the Quantification of Site-Specific DNA Methylation Levels'. In , 241–55.

Kan, Y W, and A M Dozy. 1978. 'Polymorphism of DNA Sequence Adjacent to Human Beta-Globin Structural Gene: Relationship to Sickle Mutation.' *Proceedings of the National Academy of Sciences of the United States of America* 75(11): 5631–35.

Landsteiner, Karl. 1900. 'Non-Fermentative, Lytic and Agglutinating Effects of Blood Serum and Lymph'. *Primary Journal of Bacteriology* 27: 357–62.

Lee, Hwan Young et al. 2015. 'Epigenetic Age Signatures in the Forensically Relevant Body Fluid of Semen: A Preliminary Study'. *Forensic Science International: Genetics* 19: 28–34.

Lewis, Mary E., and Ambika Flavel. 2006. 'Age Assessment of Child Skeletal Remains in Forensic Contexts'. In *Forensic Anthropology and Medicine*, Totowa, NJ: Humana Press, 243–57.

Li, En, and Yi Zhang. 2014. 'DNA Methylation in Mammals.' *Cold Spring Harbor perspectives in biology* 6(5): a019133.

López-Otín, Carlos et al. 2013. 'The Hallmarks of Aging.' *Cell* 153(6): 1194–1217.

Lucas, Victoria S., Manoharan Andiappan, Fraser McDonald, and Graham Roberts. 2016. 'Dental Age Estimation: A Test of the Reliability of Correctly Identifying a Subject Over 18 Years of Age Using the Gold Standard of Chronological Age as the Comparator'. *Journal of Forensic Sciences* 61(5): 1238–43.

Maegawa, S. et al. 2010. 'Widespread and Tissue Specific Age-Related DNA Methylation Changes in Mice'. *Genome Research* 20(3): 332–40.

Meissner, Christoph et al. 2008. 'The 4977bp Deletion of Mitochondrial DNA in Human Skeletal Muscle, Heart and Different Areas of the Brain: A Useful Biomarker or More?' *Experimental Gerontology* 43(7): 645–52.

Meissner, Cristoph, N von Wurmb, B Schimansky, and M Oehmichen. 1999. 'Estimation of Age at Death Based on Quantitation of the 4977-Bp Deletion of Human Mitochondrial DNA in Skeletal Muscle.' *Forensic science international* 105(2): 115–24.

Moore, Lisa D, Thuc Le, and Guoping Fan. 2013. 'DNA Methylation and Its Basic Function.' *Neuropsychopharmacology : official publication of the American College of Neuropsychopharmacology* 38(1): 23–38.

Moretti, Tamyra R et al. 2016. 'Population Data on the Expanded CODIS Core STR Loci for Eleven Populations of Significance for Forensic DNA Analyses in the United States.' *Forensic science international. Genetics* 25: 175–81.

Ou, Xueling et al. 2011. 'Detection and Quantification of the Age-Related SjTREC Decline in Human Peripheral Blood'. *International Journal of Legal Medicine* 125(4): 603–8.

———. 2012. 'Predicting Human Age with Bloodstains by SjTREC Quantification.' *PloS one* 7(8): e42412.

Pal, Sangita, and Jessica K Tyler. 2016. 'Epigenetics and Aging.' *Science advances* 2(7): e1600584.

Qiu, Weiliang et al. 2015. 'The Impact of Genetic Variation and Cigarette Smoke on DNA Methylation in Current and Former Smokers from the COPDGene Study'. *Epigenetics* 10(11): 1064–73.

Roewer, L et al. 1992. 'Simple Repeat Sequences on the Human Y Chromosome Are Equally Polymorphic as Their Autosomal Counterparts.' *Human genetics* 89(4): 389–94.
Roewer, Lutz. 2013. 'DNA Fingerprinting in Forensics: Past, Present, Future.' *Investigative genetics* 4(1): 22.
Sachidanandam, Ravi et al. 2001. 'A Map of Human Genome Sequence Variation Containing 1.42 Million Single Nucleotide Polymorphisms'. *Nature* 409(6822): 928–33.
Saiki, R K et al. 1985. 'Enzymatic Amplification of Beta-Globin Genomic Sequences and Restriction Site Analysis for Diagnosis of Sickle Cell Anemia.' *Science (New York, N.Y.)* 230(4732): 1350–54.
Sanger, F., and A.R. Coulson. 1975. 'A Rapid Method for Determining Sequences in DNA by Primed Synthesis with DNA Polymerase'. *Journal of Molecular Biology* 94(3): 441–48.
Sanger, F, S Nicklen, and A R Coulson. 1977. 'DNA Sequencing with Chain-Terminating Inhibitors.' *Proceedings of the National Academy of Sciences of the United States of America* 74(12): 5463–67.
Schmeling, Andreas et al. 2016. 'Forensic Age Estimation.' *Deutsches Arzteblatt international* 113(4): 44–50.
Schübeler, Dirk. 2015. 'Function and Information Content of DNA Methylation'. *Nature* 517(7534): 321–26.
Semba, Richard D, Emily J Nicklett, and Luigi Ferrucci. 2010. 'Does Accumulation of Advanced Glycation End Products Contribute to the Aging Phenotype?' *The journals of gerontology. Series A, Biological sciences and medical sciences* 65(9): 963–75.
Sierra, Marta I, Agustín F Fernández, and Mario F Fraga. 2015. 'Epigenetics of Aging.' *Current genomics* 16(6): 435–40.
Sobrino, Beatriz, and Angel Carracedo. 2005. 'SNP Typing in Forensic Genetics: A Review.' *Methods in molecular biology (Clifton, N.J.)* 297: 107–26.
Southern, E.M. 1975. 'Detection of Specific Sequences among DNA Fragments Separated by Gel Electrophoresis'. *Journal of Molecular Biology* 98(3): 503–17.
Stoneking, M et al. 1991. 'Population Variation of Human MtDNA Control Region Sequences Detected by Enzymatic Amplification and Sequence-Specific Oligonucleotide Probes.' *American journal of human genetics* 48(2): 370–82.
De Tobel, Jannick, Elke Hillewig, and Koenraad Verstraete. 2017. 'Forensic Age Estimation Based on Magnetic Resonance Imaging of Third Molars: Converting 2D Staging into 3D Staging'. *Annals of Human Biology* 44(2): 121–29.
Weidner, Carola et al. 2014. 'Aging of Blood Can Be Tracked by DNA Methylation Changes at Just Three CpG Sites'. *Genome Biology* 15(2): R24.
Weinhold, Bob. 2006. 'Epigenetics: The Science of Change.' *Environmental health perspectives* 114(3): A160-7.
You, Frank M et al. 2008. 'BatchPrimer3: A High Throughput Web Application for PCR and Sequencing Primer Design.' *BMC bioinformatics* 9(1): 253.
Zbieć-Piekarska, Renata et al. 2015. 'Examination of DNA Methylation Status of the ELOVL2 Marker May Be Useful for Human Age Prediction in Forensic Science'. *Forensic Science International: Genetics* 14: 161–67.
Zhao, Zhongming, Yun-Xin Fu, David Hewett-Emmett, and Eric Boerwinkle. 2003. 'Investigating Single Nucleotide Polymorphism (SNP) Density in the Human Genome and Its Implications for Molecular Evolution.'*Gene* 312: 207–13.