7-19-2019

Development of DNA Methylation Markers to Infer Age, Smoking Status and Body Fluid Types for Forensic Application

Hussain J.H. Alghanim
halgh002@fiu.edu

FLORIDA INTERNATIONAL UNIVERSITY

Miami, Florida

DEVELOPMENT OF DNA METHYLATION MARKERS TO INFER AGE, SMOKING STATUS AND BODY FLUID TYPES FOR FORENSIC APPLICATION

A dissertation submitted in partial fulfillment of

the requirements for the degree of

DOCTOR OF PHILOSOPHY

in

CHEMISTRY

by

Hussain Jaffar Hussain Alghanim

2019

To: Dean Michael R. Heithaus
College of Arts, Sciences and Education

This dissertation, written by Hussain Jaffar Hussain Aghanim, and entitled Development of DNA Methylation Markers to Infer Age, Smoking Status and Body Fluid Types Forensic Application, having been approved in respect to style and intellectual content, is referred to you for judgment.

We have read this thesis and recommend that it be approved.

_____
Watson Lees

_____
Yuan Liu

_____
Yukching Tse Dinh

_____
Wensong Wu

_____
Bruce McCord, Major Professor

Date of Defense: July 19, 2019

The dissertation of Hussain Jaffar Hussain Alghanim is approved.

_____
Dean Michael R. Heithaus
College of Arts, Sciences and Education

_____
Andrés G. Gil
Vice President for Research and Economic Development
and Dean of the University Graduate School

Florida International University, 2019

DEDICATION

To my beloved Prophet Mohammad and his Household (blessings of God be upon them) who always encourage learning and teaching.

“One who follows a path in search of knowledge, Allah (God) will make easy for him a path to paradise” &

“Acquire knowledge and impart it to the people”_ Prophet Mohammad (SAWS)

To my parents

To everyone who love science

ACKNOWLEDGMENTS

In the name of Allah the most beneficent the most merciful

First and foremost, I can't thank enough my lord who created me without a need and never stop His shower of mercy over me. I also would like thanks my beloved prophet Mohammad and his family peace be upon them for guiding me to the best manner ever. I would like to thanks everyone who have been involved in this research throughout my years of studies.

The most heartfelt thanks and gratitude go to my parents for their continuous support and faith and for always believing in my success. Without their patience, understanding, and love, the completion of this work would not have been possible. I know you have both been through a lot of hard times to make sure I accomplish my dreams and I will always be grateful. I thank my brothers and sister for their encouragement and inspiration. To my grandmothers and grandfathers who always used to pray for me and for all the great people who also have me in their prayers, thanks you so much. Allah, bless you all.

I would like to express my great appreciation to my major advisor Dr. Bruce McCord. I am so grateful to his guidance and support during my graduate study and for the opportunity to work on this interesting epigenetic project. His optimism in science and life made it easy to keep finding new ways to overcome obstacles and succeed and learn from whatever I try. Thank you so much.

I also would like to thanks my committee members. Dr. Wensong Wu for her great help in the statistical part. Thanks for your patience and always finding time for

me to discuss statistical issues that I faced during my studies. You help was so much appreciated. Dr. Watson Lees for his always being available whenever I stop by to answers any questions and always trying to find solution for me. Dr. Yukching Tse Dinh and Dr. Yuan Liu for their support and being great teachers of biochemistry and always bringing up the more fundamental aspect of my research. Thanks for teaching me a great deal of science that I will carry out forever.

Special thanks to my sponsor, the General Headquarter of Dubai Police for the fellowship, for funding this research and for the generous support I received on all my undergraduate and graduate studies. I would like to appreciate the Lieutenant General Khamis Mattar AlMazeina for his encouragement and support to peruse this PhD degree. Allah bless him.

ABSTRACT OF THE DISSERTATION

DEVELOPMENT OF DNA METHYLATION MARKERS TO INFER AGE, SMOKING STATUS AND BODY FLUID TYPES FOR FORENSIC APPLICATION

by

Hussain Jaffar Hussain Alghanim

Florida International University, 2019

Miami, Florida

Professor Bruce McCord, Major Professor

In forensic investigations, biological evidences have great potential to place a suspect at the scene of a crime. However, in many cases suspect(s) usually can't be identified with any local or national databases. In such challenging cases, police need a tool that can provide investigative leads. Thus, different genetic biomarkers including DNA methylation markers have been developed for different phenotypic traits to serve as source of intelligence information to investigators in cases of unknown DNA profiles. In this project, novel sets of DNA methylation markers were developed for the forensic estimation of human age, determining the smoking status and body fluid identification.

First, single- and dual-locus age predictors for saliva and blood were developed from CpG sites in *KLF14* and *SCGN* based on pyrosequencing data and using a multivariate linear regression analysis. In saliva, single-locus model in particular was efficient for younger subjects (≥40 years) correctly predicting age of 78.9% of the samples with a MAD of 5.1 years in the validation set. Second, a quick and cost effective 4-CpG pyrosequencing assay was built to infer smoking habit using

multinomial logistic regression model (MLR). In blood, the model correctly predicted 90.0% of current smokers, 66.7% of former smokers, and 84.9% of never smokers. In addition, the MLR model correctly predicted 86.9% of current smokers, 54.5% of former smokers, and 77.8% of never smokers in saliva. Finally, new set of tissue specific differentially methylated region (tDMRs) were identified for forensic discrimination of body fluids. NMUR2 and UBE2U markers were found to be specific for sperm and the two assays developed can be employed using both pyrosequencing or high resolution melt (HRM) analysis. In addition, the developed AHRR marker can discriminate blood stains from other body fluids. These three piece of information may be very valuable source of investigative leads aiming to trace unknown perpetrators who could not be identified using the conventional autosomal DNA typing.

TABLE OF CONTENTS

LIST OF TABLES

TABLE PAGE

LIST OF FIGURES

FIGURE PAGE

ABBREVIATIONS AND ACCRONYMS

A	Adenine
AUC	Area under the curve
AHRR	Aryl hydrocarbon receptor repressor
ANOVA	Analysis of variance
bp	Base pairs
C	Cytosine
CpG	Cytosine phosphate guanine
DLX5	Distal-less homeobox 5
DNA	Deoxyribonucleic acid
DNMT	DNA methyltransferases
dNTP	Deoxynucleoside triphosphate
FDP	Forensic DNA phenotyping
G	Guanine
HRM	High resolution melt
KLF14	Kruppel-Like Factor 14
m5C	5-methylcystosine
m4C	N4-methylcystosine
MLR	Multinomial logistic regression
MSRE	Methylation-Sensitive Restriction Enzyme
Ms-SNuPE	Methylation-sensitive single nucleotide primer extension
MVPs	Methylation variable positions
m/z	Mass per charge

NMUR2	Neuromedin u receptor 2
PCR	Polymerase chain reaction
PPi	inorganic pyrophosphate
RNA	Ribonucleic acid
ROC	Receiver operating characteristic
SAM	S-adenyl methionine
SCGN	Secretagogin
SNP	Single nucleotide polymorphism
ssDNA	Single-stranded DNA
STR	Short tandem repeats
T	Thymine
Taq	Thermus acquaticus
tDMRs	Tissue-specific DNA methylation regions
TEs	transposable elements
Tm	Melting temperature
UBE2U	Ubiquitin conjugating enzyme E2 u
VMRs	Variably methylated regions

CHAPTER I- INTRODCUTION

1. Epigenetic

Historically, the word “epigenetics” was used to explain events that could not be descried though genetic concepts. The Greek prefix epi means “on top” or “in addition” and thus, epigenetics refers to the extra information in addition to the genetic information found in the DNA sequence (Brait et al., 2011). In the 1940s, Conrad Waddington was the first to introduced the term “epigenetics” in modern biology as “the branch of biology which studies the causal interactions between genes and their products, which bring the phenotype into being” (Waddington et al., 1942, Lim and Brunet et al., 2013). Currently, the widely accepted definition for epigenetics is the study of any stable processes that produce heritable changes in gene expression or cellular phenotype that does not strictly depend on the DNA sequence (Goldberg et al., 2007). Thus, the gene function and phenotypic outcome in different cells are caused by various mechanisms other than heritable changes found in the DNA sequence. Epigenetic modifications can be transmitted in multiple types, including DNA methylation, chromatin remodeling, post- translational modification of histone proteins, and non-coding RNA. These modifications play an important role in regulating gene expression without causing any changes in DNA sequence (Vidaki el al., 2013). Through the whole lifetime of an organism, epigenetics mechanisms provide a link between the various environmental factors and the different phenotypic changes (Tammen el al., 2013). In mammals, one vital function that is controlled by

epigenetic mechanisms is cell differentiation in which stem cells become fully differentiated cells during embryogenesis (Rando and Verstrepen et al., 2007). Among all the different epigenetic mechanisms, the field of DNA methylation is perhaps the best characterized and studied, making it the epicenter of numerous biomolecular investigations. The present study focuses merely on DNA methylation and its potential applications in the forensic science field.

2. DNA methylation

2.1 Definition of DNA Methylation

In general terms, the methylation of DNA refers to the addition of a methyl group at the 5' position of the cytosine residue in CpG dinucleotides forming 5-methylcystosine (m5C) (Rakyan et al., 2011) (Figure 1.1). It has been shown that cytosines in the form of CpC, CpA, and CpT dinucleotides may also be methylated but this occurs less frequently than that seen with the CpG dinucleotide form (Voet et al., 2011). In addition, methylated cytosines can also oxidize forming 5-hydroxymethylcytosines although this is less common (Ficz et al., 2011, Dahl et al., 2011). The conversion of methylcytosine to hydroxymethylcytosine could be the first step in the pathway that leads towards DNA demethylation. Some cytosines of DNA can also be methylated in a species-specific pattern to form N4-methylcystosine (m4C).

When looking at a single CpG site within a DNA strand, DNA methylation is said to be a binary trait in which the site can be either methylated or not. However, experimental samples such as extracted DNA from peripheral blood contain large number of DNA strands. Typically, a certain percentage of these DNA strands will

be methylated and the rest will not. Thus, the relative percentage of DNA methylation for a given sample at a single CpG site is said to be a quantitative trait indicating the proportion of the DNA strand that is methylated (Bibikova et al., 2011, Lee et al., 2013).

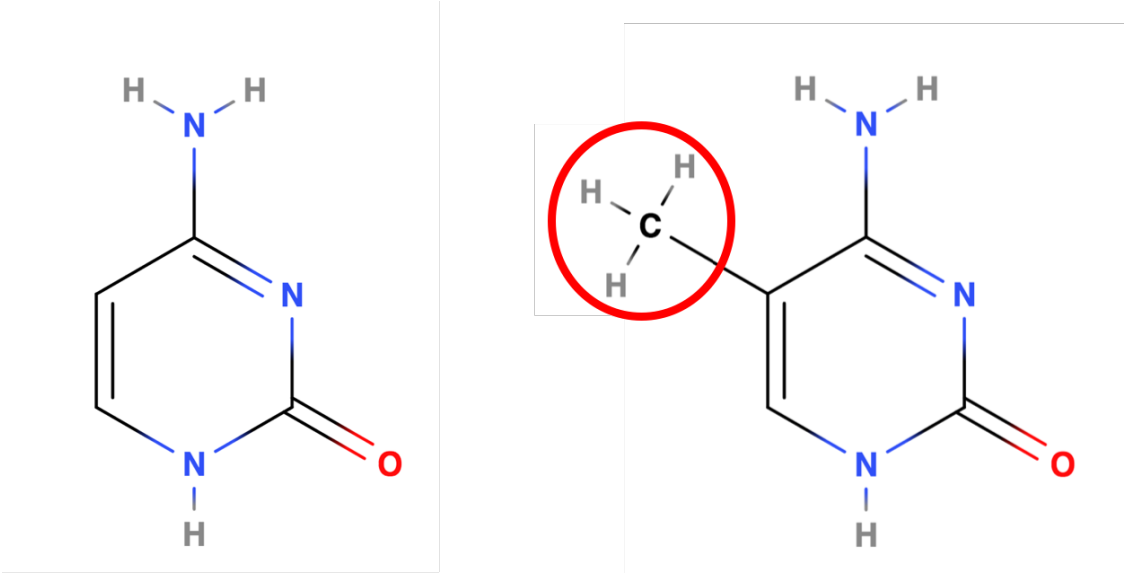


Figure 1.1: DNA methylation is the addition of methyl group (shown in red cycle) at the 5' position of the cytosine nucleotide

2.2 Genome Distribution of DNA Methylation

DNA methylation is the most widely studied epigenetic marker (Rakyan et al., 2011). In mitotically stable DNA, the predominant form is methylation of cytosines in the context of cytosine followed by guanine dinucleotides (CpGs). Depending on the cell type, the 5-methylcytosine, also known as the "fifth base", is estimated to present in about 4-6% of the cytosine bases in the human genome (Tammen el al., 2013). The CpG dinucleotides comprise of only around 1% of the human genome (Han et al., 2008). The upstream regions of many genes contain a high density of CpG dinucleotides in clusters known as CpG islands which have a role

in gene regulation. Those islands are between 300-3000 bp in length and have >55% GC content (Bird et al., 2002, Han et al., 2008). There are around 50200 CpG islands in the human genome which are found predominately in gene promoters and repetitive DNA elements. In addition, CpG islands are less commonly found in gene bodies (especially in exons) (Medvedeva et al., 2010).

It has been shown that not all CpGs are methylated. Analysis of DNA sequences in the human genome indicates that approximately 30% of CpGs are hypomethylated (i.e., less than 20% methylation level) and approximately 40% of CpGs are said to be hypermethylated (i.e. over 80% of DNA molecules are methylated). The remaining 30% of CpGs are shown to be methylated at an intermediary level meaning that between 20-80% of their DNA molecules are methylated (Eckhardt et al., 2006). Unmethylated CpGs are frequently found in CpG islands located at the 5' ends (gene-promoter regions) of many genes. CpGs in gene bodies and repetitive DNA elements are mostly methylated (Maunakea et al., 2010). In addition, methylation patterns are usually correlated across multiple neighboring CpGs and this co-methylation is stronger for CpGs located within CpG islands than for those found outside these regions (Eckhardt et al., 2006, Bell et al., 2011).

2.3 Function of DNA Methylation

Three families of DNA methyltransferases ($DNMT_1$, $DNMT_2$ and $DNMT_3$) are enzymes used to catalyze the de novo methylation and help to maintain DNA methylation (Kader et al., 2015). These DNMT enzymes assist in the transfer of a methyl group from S-adenyl methionine (SAM) to the fifth carbon of a cytosine

residue to form m5C (Aguilera et al., 2010). The main role of DNA methylation is to regulate gene expression and protect the genome integrity.

In general, the use of DNA methylation has been considered to inhibit DNA transcription (Figure 1.2) (Vidaki et al., 2013), although in some instances it is associated with gene activation (Straussman et al., 2009). This impact on gene promotors can be best understood in two ways. First, directly, methylation of DNA can hinder the physical binding of transcription enhancers. Second, indirectly, DNA methylation can cause chromatin remodeling that alters DNA accessibility during transcription (Portela and Esteller el al., 2010). This inverse correlation of DNA methylation and gene expression can also depend on the presence and locations of CpG islands.

In addition to gene regulation, methylation of DNA is used to protect the integrity of the human genome by inhibiting the mobility of transposable elements (TEs). These TEs are repetitive DNA sequences that are known to be able to be transcribed into new locations in the genome. TEs are hypermethylated and thus cause transcriptional silencing. Since the transposition mechanism needs TE-encoded enzymes to proceed, inhibition of TE transcription through DNA methylation can prevent translocations and gene disruptions (Levin and Moran et al., 2011). Therefore, loss of DNA methylation allows for reactivation of TE and transposition (Tsukahara et al., 2009).

The importance of DNA methylation in the function of normal cells is manifested by its role in cellular differentiation, X chromosome inactivation, genomic imprinting, maintenance of chromatin structure and suppression of “parasitic”

DNA. DNA methylation is important in cellular development. For example, the disruption of $DNMT_1$ alleles in embryonal stem cells results in embryonic death (Li et al.,1992) and deficiency of $DNMT_3A$ and $DNMT_3B$ enzymes is also lethal to these types of cells (Okano et al., 1999). In addition, mutations in the DNA methylation mechanism can lead to human disease. For example, mutations in $DNMT_3B$ results in ICF syndrome which is characterized by a B cell immunodeficiency, centromeric instability, and facial abnormalities.

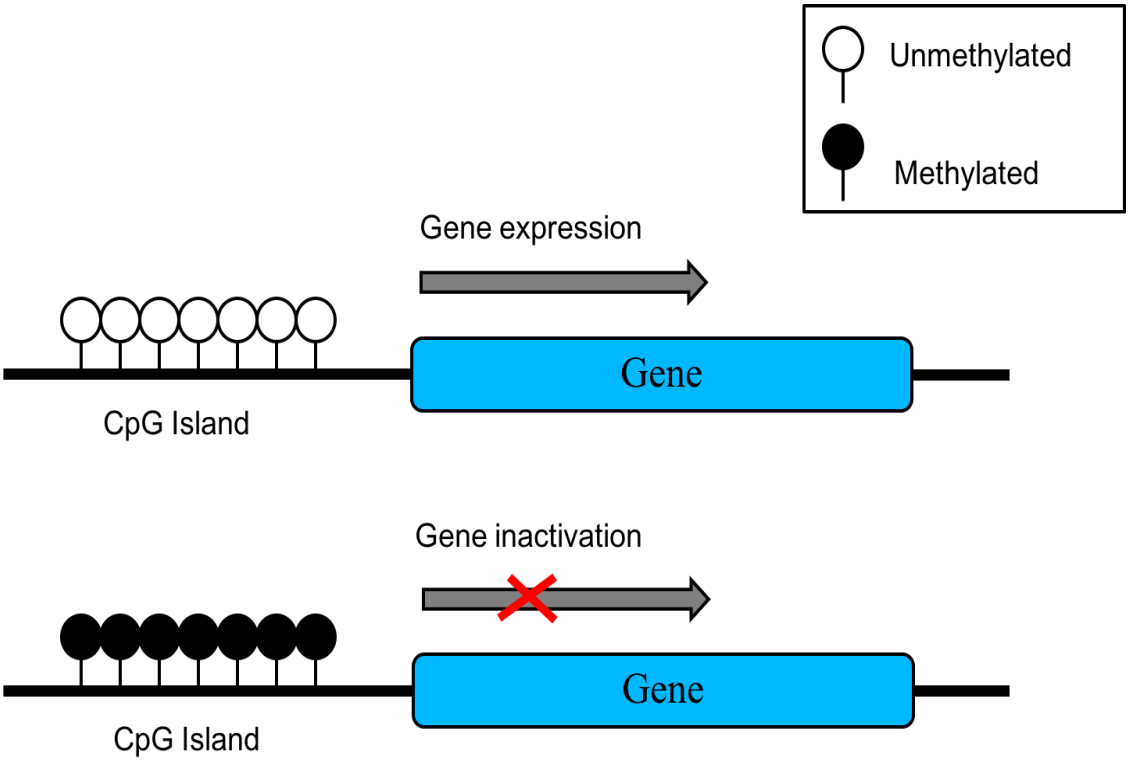


Figure 1.2: Representation of a gene containing a CpG island. Generally, CpG island that is unmethylated enable gene expression, whereas the gene expression is inhibited when CpG island is methylated

2.4 Variations and Factors Affecting DNA Methylation

The full range variations in DNA methylation is huge considering that the diploid human genome contains $\geq 10^7$ CpGs which may all potentially vary (Weidner et al., 2014). DNA methylation comprises of three common formats: methylation variable positions (MVPs), variably methylated regions (VMRs) or differentially methylated regions (DMRs) (Rakyan et al., 2004, Frigola et al., 2006). Methylation variable positions (MVPs) are defined as regions of the genome in which DNA methylation occurs at single CpG sites whereas VMRs are the regions of the genome which show increased variability instead of gain or loss of DNA methylation. Finally, DMRs are defined as regions of the genome in which multiple of adjacent CpG sites show patterns of differential methylation. These regions can encompass various states such as tissue-specific DMR (tDMR), aging-specific DMR (aDMR), imprinting-specific DMR (iDMR), cancer-specific DMR (cDMR), and reprogramming-specific DMR (rDMR).

Some studies suggested that the genetic sequences of the promoters are the main factors of variation that present at methylation sites. For example, mutations that occurs at the transcription factor-binding sites can prevent the maintenance of DNA methylation in the nearby locations (Bell et al., 2011, Lienert et al., 2011). DNA methylation can be influenced by complex exogenous and endogenous factors including early life experiences, nutrition and diet, aging, exposure to pollutants, smoking, and social environment (Kader et al., 2015). The three major factors which can alter controlling DNA methylation variability are aging, cell development and environmental exposures.

2.4.1 Aging Effects on DNA Methylation

The first study which associated the DNA methylation with aging came from investigating the organ and life stages in humpback salmon. This study found that global DNA methylation of 5-methylcytosine decreased significantly with age (Berdyshev et al., 1967). Other studies reported similar decreases in DNA methylation with age for rats, mice, and humans (Bjornsson et al., 2008, Vanyushin et al., 1973, Wilson et al., 1987). The highest level of 5-methylcytosine was detected in embryos and decreased gradually with age. To verify the link between global hypomethylation with aging, a comparison was made between the DNA methylomes of CD4+ T cells of newborns and centenarians using whole genome bisulfite sequencing. CpG methylation of newborns was shown to be more homogeneous compared to those found in centenarians, suggesting a scattered pattern of demethylation over the lifetime (Jung et al., 2015). This decease of global DNA methylation could be attributed mainly to a failure in DNMT1 functions with time (Casillas et al., 2003). The mechanism(s) by which aging can modify DNA methylation in cells is still poorly understood (Lillycrop et. al., 2014). However, it has been proposed that changes in DNA methylation associated with aging may be induced through both stochastic effects and environmental factors (Figure 1.3).

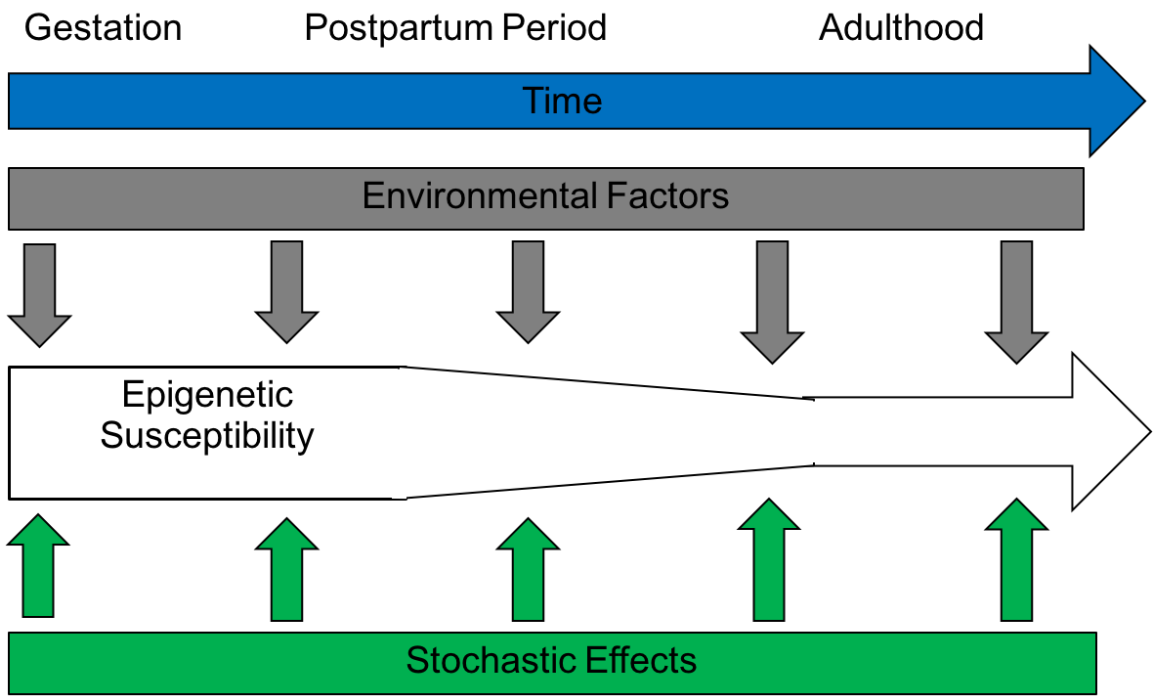


Figure 1.3: Age associated changes in DNA methylation patterns is shown to be based on two proposed mechanisms: 1) Stochastic factors, 2) Environmental factors (regenerated from Lillycrop et al., 2014).

Stochastic effects are considered one mechanism that leads to changes in DNA methylation patterns. These stochastic effects can occur due to different processes that take place inside human cells. After embryonic re-establishment, DNA methylation levels are maintained by DNMTs during successive cell divisions, and thus remain relatively stable across life. Maintaining DNA methylation levels is somewhat manageable from youth to middle-aged adulthood. However, this maintenance of DNA methylation is not perfect. Studies have found greater variations in DNA methylation for older monozygotic twin pairs than those seen in younger monozygotic twin pairs. This phenomenon is known as epigenetic drift and could be due to small errors that stochastically accumulate whenever DNA

methylation markers are copied during successive cell divisions (Fraga et al., 2005). Another reason is the loss of epigenetic fidelity due to a decrease in DNMT1 activity that is associated with aging. A decline in DNMT1 enzymes which are are responsible for maintenance of DNA methylation would cause global demethylation of the genome and would also may lead to increase in epigenetic heterogeneity within a cell population as well as increase in variations in inter-cell gene expression within tissues (Bollati et al., 2009, Bahar et al., 2006). Such accumulation of epigenetic modifications could eventually cause a decline in tissue function and result in disease. Environmental factors are another important potentiator of variations in methylation. There is evidence that various environmental factors may also result in age-associated changes in DNA methylation. Some of these environmental factors including stress, placental insufficiency, pollution, nutrition and diet. Studies have determined that the epigenome is most susceptible to changes by such factors during early life than later in life (Godfrey et al., 2001) (Figure 1.3).

2.4.2 Cell Development Effects on DNA Methylation

DNA methylation patterns can change during the cell development. During early embryo development, DNA methylation is a very dynamic as cell-specific methylation patterns are utilized to promote cell differentiation so that different cells could possess certain structures and functions (Talens et al., 2010, Zeilinger et al., 2013). After fertilization, the zygote genome undergoes global depletion of DNA methylation until getting to its lowest levels during the pre-implantation blastocyst stage in which the pluripotent embryonic stem cells form most of the inner cell

mass of the embryo (Feng et al., 2010). Post-implantation, DNA methylation levels are re-established, stabilized and gradually become similar to the patterns found in adult somatic cells (Smith et al., 2012).

It has been shown that the erasure and re-establishment of methylation patterns that take place during early embryo development has an important role in cell differentiation. The global erasure of DNA methylation stimulates the expression of pluripotency genes which in turn activate the development of embryonic stem cells that have the ability to differentiate into any tissue in the body. At the initial stages of cell differentiation, the pluripotency genes undergo re-methylation to develop cell-specific DNA methylation patterns, thus cells formation (Straussman et al., 2009, Epsztejn-Litman et al., 2008).

2.4.3 Environmental Factors on DNA Methylation

Environmental factors play a vital role in modifying the DNA methylation level. Data from multiple research studies suggest that the DNA methylation status can be affected by a variety of external environmental factors (Dogan et al., 2014, Philibert et al., 2012, Wahl et al., 2017). Those environmental events that take place in early embryogenesis (e.g. when global erasure and re-establishment of DNA methylation occur) may produce extensive and soma-wide modifications of DNA methylation and may also play a role in fetal programing of adult disorders. The environmental changes of the epigenome during early stage of life may have great effect in metabolism and physiology as well as in influencing future disease risk (Gluckman et al., 2005). On the other hand, environmental factors that act later in life are more likely to produce less extensive and tissue-specific changes

of DNA methylation however they may contribute to tissue-specific carcinogenesis (Terry et al., 2011, Ehrlich and Lacey et al., 2013).

The potential for environmental factors to play an important role in epigenetic drift was clearly demonstrated through scientific studies showing that the difference in DNA methylation levels between monozygotic twins were greater for the pairs that spent less of their lifetime together or displayed different lifestyles (Fraga et al., 2005). Environmental exposures such as diet, stress, and smoking can alter DNA methylation at various stages of human development. DNA methylation can provide links between the fixed genome and dynamic environment altering the methylation patterns to fit the specific needs of the cell (Tammen et al., 2013). Thus, DNA methylation levels may provide evidence for the activities of genes within a particular tissue (Fragou et al., 2011).

II. Forensic Epigenetics

1. Potential Forensic Applications of DNA Methylation

DNA methylation was first introduced in forensic genetics by Naito et al. (1993) as a method for sex determination. Today, studying DNA methylation patterns has been proposed for various applications including verifying the authentication of the DNA samples (Frumkin et al., 2010), sex determination (Lee et al., 2012), differentiating between monozygotic twins (Stewart et al., 2015), identifying the source of body fluid (Madi et al., 2012), estimating body mass index (Wahl et al. 2017), and predicting ancestry (Fraser et al., 2012). Various DNA methylation biomarkers were reported recently to identify the individual's diet, (Anderson et al. 2012) alcoholic person (Philibert, et al., 2014), smoking habits, illicit drug users (Nielsen et al., 2009) and socioeconomic status (Borghol et al., 2012). Figure 1.4 shows the potential epigenetic applications that could be employed in the forensic field. The focus of this project was the forensic application of age estimation, identification of body fluid types, and inference of the smoking status.

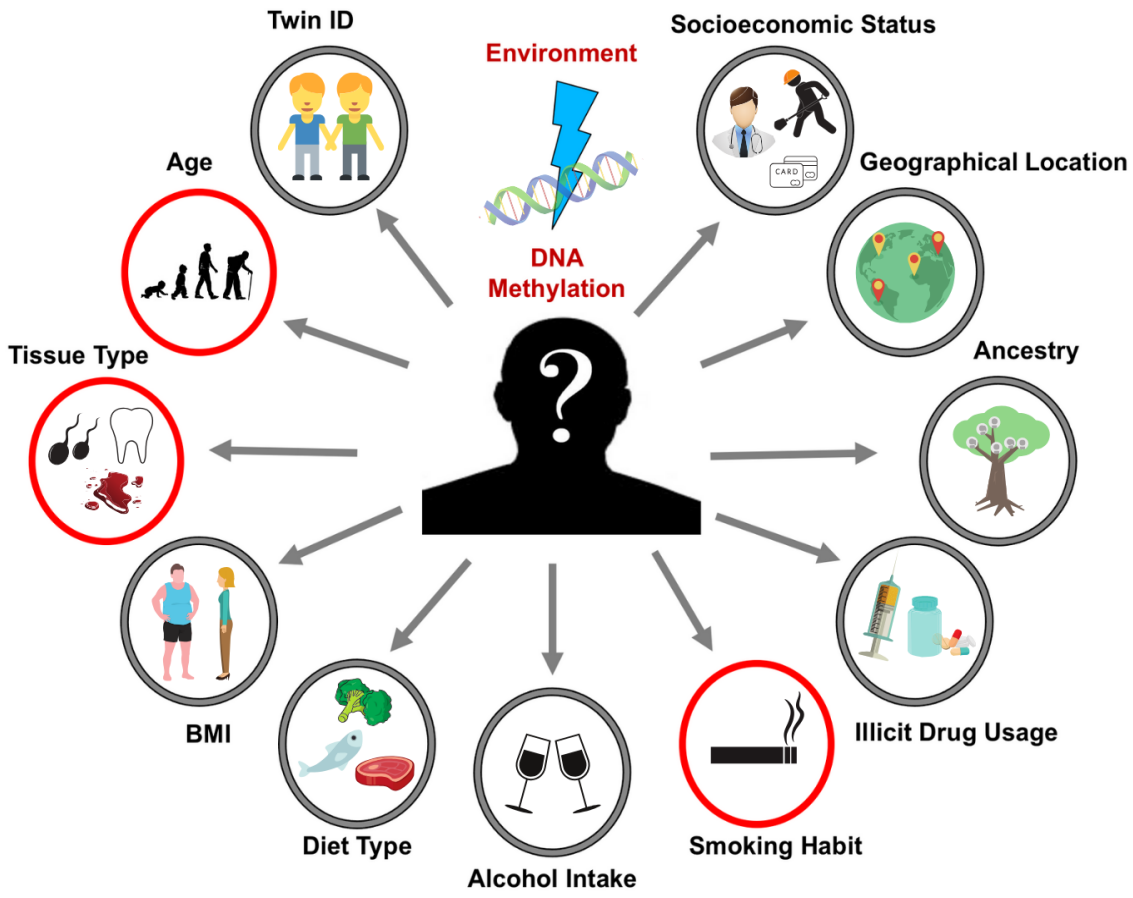


Figure 1.4: shows the potential epigenetic applications that could be employed in the forensic field. Red circle shows the applications which are the focus of this project. (regenerated from Vidaki et al., 2017)

1.1 Estimation of age

Forensic DNA phenotyping (FDP) is a field that is emerging as new tool of DNA-based intelligence information which can aid law enforcement efforts against crime. Using the DNA sample left at a crime scene to predict the external visible characteristics such as eye color, hair color, and height can narrow the list of suspects. As individuals age, the physical appearance can change drastically and thus age estimation can provide great clues to draw approximate DNA-based sketches of the unidentified suspect (Vidaki el al., 2013). Many efforts recently

have been targeting DNA methylation-based methods to identify a set of CpG sites that can estimate chronological age of individuals connected with a particular crime. Another possibility is to use DNA methylation to predict the chronological age of tissues at the time of death for cadavers (Zbieć-Piekarska et al., 2015a).

In recent years, microarray technology has enable researchers to search for specific age-dependent CpGs at genome-wide level. Thus, more specific DNA methylation changes at particular genes or CpG sites can be identified that associate highly with age. The methylation levels at these CpG sites can be tissue dependent and can be studies in various cell types (Day et al., 2013. Horvath et al., 2013). For example, in order to study the effect of aging, Bocklandt et al. (2011) used buccal swabs from 34 pairs of identical twins (21-55 years old) in a genome wide methylation study using Illumina HumanMethylation27 microarray. The authors reported finding 88 novel CpG loci that produced DNA methylation levels that were correlated with age. In this study, a subset of three CpG loci found in NPTX2, EDARADD and TOM1L1 were selected to be tested using regression analysis. Based on these three loci, the prediction model for each individual's age was estimated to provide an average accuracy of 5.2 years. Koch and Wagner et al. (2011) also used the 27K array to find age-associated CpG in various cells and tissues. They reported a prediction model for age that consist of five CpGs located in GRIA2, NPTX2, TRIM58, KCNQ1DN, and BIRC4BP. This model could be used in many types of body fluids but the average absolute difference between the predicted and chronological age was around 11 years. Using 450K array, Hannum et al. (2013) identified 71 DNA methylation sites that show correlation with age in

blood. This predictor model enhances the prediction of chronological age in blood with estimated correlation coefficient of 96% and standard error of 3.9 years. Horvath et al. (2013) made extensive study by examining 8,000 samples on 27k and 450K array. Horvath developed a mutli-tissue prediction model consist of 353 CpGs that showed high accuracy with an error of 3.6 years. Epigenome-wide studies require a large amount of DNA and significant time and effort, thus procedures are unlikely to be adopted in the forensic field. However, the development of an assay using only a few CpG sites should be more attractive to forensic geneticists, particularly if it provides similar accuracy as that provided by genome-wide DNA methylation profiling. Therefore, different studies have employed pyrosequencing analysis to target a small subset of CpG sites for age prediction. For example, in a study by Weidner et al. (2014), three different sets of age-related CpGs located in the genes ITGA2B, ASPA and PDE4C were used to establish an age predictor model. Blood samples of 151 individuals were screened by bisulfite treatment followed by PCR amplification and pyrosequencing. The selected loci showed reliable age prediction with a mean absolute deviation of about 5 years in blood.

The ELOVL2 gene is the most extensively evaluated methylated biomarker of age (Garagnani et al., 2012, Hannum et al., 2013). In a forensic study, Zbieć-Piekarska et al. (2015b) used a pyrosequencing platform to develop an age prediction model using two CpG sites on the ELOVL2 genes with a prediction error of 6.85 years. This age-production model showed promising results for blood samples which are considered to be the main source of DNA evidence analyzed

in forensic laboratories. Extending from this study, Zbieć-Piekarska et al. (2015a) further evaluated the 71 methylated DNA sites, previously reported by Hannum et al. (2013) using multiple regression statistics and selected the five most informative CpG loci which were located in *ELVOL2, C1orf132, TRIM59, KLF14, and FHL2*. The five identified CpG sites generated an estimate of chronological age with a mean absolute deviation (MAD) equal to 3.9 years for the validation set in blood. In addition, Lee et al. (2015) reported a model containing one CpG site in TTC7B gene and two CpGs located in the NOX4 gene devoted mainly for semen which is a very relevant body fluid in forensic setting. The study employed the use of 450K array followed by methylation SNaPshot analyses and displayed high accuracy for predicting age with a MAD from chronological age of around 5 years. Recently, Park et al. (2016) also proposed an age prediction model for blood from using three CpGs at ELOVL2, ZNF423, and CCDC102B genes. The three CpG sites were evaluated in more than 760 blood samples and screened by pyrosequencing analysis. The prediction model provided a high accuracy with a mean absolute deviation from the chronological age equal to 3.4 years. In another study by Bekaert et al. (2015), a model was built based on four CpGs from four different genes (ASPA, PDE4C, ELOVL2, and EDARADD). The model was designed using DNA from blood and teeth samples with MAD from chronological age of 4.9 years.

1.2 Inferring Smoking Status

The ability to infer the smoking status from DNA material found in the crime scene would be of very useful in characterizing an unknown trace donor and thus aiding the investigations. Variations in DNA methylation are one mechanism that potentially mediate the effects of tobacco smoking in individuals. Cigarette smoking is considered one of the most powerful environmental factors that cause DNA methylation changes (Breitling et al., 2011). Smoking has been associated with deferential methylation in the global methylation status (Zeilinger et al., 2013, Dogan et al., 2014) as well as with various cancer-related genes (Marsit et al., 2007, Enokida et al., 2006). Thus, it has been proposed that smoking could affect DNA methylation by mechanisms related to: 1) carcinogen-induced DNA damage and repair (Huang et al., 2012), 2) nicotine effects on gene expression by down regulating the DNMTs activity (Satta et al., 2008), 3) regulating the activity of DNA-binding factors (e.g. Sp1) (Di et al., 2012), and 4) hypoxia (Liu et al., 2011).

Several genome-wide studies investigate and identify a set of loci whose methylation pattern is associated with smoking. Breitling et al. (2011) started investigating the effect of smoking using 27K array by studying peripheral mononuclear cell pellets and identified several loci that are associated with smoking including *F2RL3*, *GPR15*, and *ORAI2*. The study reported that one CpG site at *F2RL3* gene (cg03636183) showed significant methylation differences between smoker and non-smoker (% methylation difference = 12%). The first truly genome-wide association study of the effect in smoking was reported by Monick at el. (2012). The group used Illumina HumanMethylation 450K BeadChip to study

methylation status in lymphoblast and lung macrophage DNA and identified several candidate loci with special emphasis on variation of methylation profiles on Aryl Hydrocarbon Receptor Repressor (AHRR). Two specific CpG sites (cg23576855 and cg05575921) at AHRR genes displayed hypomethylation profiles and the difference in the methylation percentage between the smokers and non-smokers in the lung was larger than in the peripheral lymphocytes (34% vs. 17%). A large study carried by Zeilinger et al. (2013) further confirmed the above noted loci and extended the list of genes to include HIVEP3 and CACNA1D. This study tested more than 2000 whole blood samples and reported a decrease in percent methylation for smokers in the AHRR gene (% methylation difference = 24%). The AHRR gene encode a protein that have a role in the aryl hydrocarbon receptor signaling cascade which mediating detoxification of environmental pollutants (Opitz et al., 2011). Thus, smoking tends to decrease the AHRR DNA methylation and as a result increase the expression of AHRR that may remove the harmful chemical substance in tobacco smoking such as the polycyclic aromatic hydrocarbon (Lee et al., 2013).

As noted before, DNA methylation changes between different tissues. For example, the study by Wu et al. (2014) demonstrated the presence of significant variations in DNA methylation patterns between saliva and blood even from one group of non-smokers. Thus, it's important to examine the effect of smoking in various type of tissues. On the other hand, the parental exposure to smoking can also affect the fetus by modifying DNA methylation during embryo development. For example, Joubert et al. (2012) studied the epigenetic effect of maternal

smoking during pregnancy on the fetus. The study found that maternal exposure to cigarette smoking can produce changes in DNA methylation at AHRR, GF1, and CY1A1 genes in cord blood and placental samples. The percent difference of DNA methylation at AHRR and CY1A1 between the exposed and non-exposed newborns was ≤10% (Joubert et al., 2012). In addition, DNA methylation effects from tobacco smoking were correlated with cumulative smoke exposure (pack-year) and with the time since quitting (Wan et al., 2012).

1.3 Identification of Body Fluids

Being able to differentiate between the different body fluids found at the crime scene evidence can provide valuable information for crime scene reconstruction, since the presence of certain body fluids imply specific types of activity (Lee et al. 2012). The methods that are currently in use by most forensic labs depend on colorimetric detection of enzymes. Because of many of these methods rely on the detection of differences in enzyme activity between various cell types, most current methods are considered only presumptive. In contrast, DNA methylation markers can be highly tissue specific and quite sensitive (Park et al. 2014). These methylation markers can also be used to determine the origins of the DNA sample, either from a single source or multiple contributors (Frumkin et al., 2011). Rakyan et al. reported a comprehensive genome wide tissue-specific DNA methylation study across 13 normal somatic tissues, placenta, sperm and an immortalized cell line. The aim was to identify tissue-specific DNA methylation regions (tDMRs) that can play a part in cellular identity and regulate tissue-specific genome function (Rakyan et al., 2004).

Frumkin et al. (2011) and his coauthors were the first to investigate the potential application of forensic tissue identification through DNA methylation approach. A panel of 15 methylation sites were selected for their tissue identification assay using methylation restriction enzymes followed by multiplex PCR. The assay as developed and tested on 50 samples including saliva skin, blood, semen, urine, vaginal secretion and menstrual blood, urine (Frumkin et al., 2011). Wasserstrom et al. adopted the method by Frumkin and developed a kit that can be used for forensic identification of seminal stain in the crime scene (Wasserstrom et al., 2013). In addition, Lee et al. 2012 explored the tDMRs that are previously reported and suggested bisulfite sequencing as a new method. The study confirmed that two previously identified tDMRs (DACT1 and (USP49) could be utilized for semen identification (Lee et al., 2012). They also demonstrated the possibility of using HOXA4 as a marker for blood and PFN3 as a marker for vaginal secretions. Several studies have shown tissue-specific patterns of DNA methylation for different forensically related tissues. The following markers have been studied: C20orf117 for blood (Madi et al., 2012), BCAS4 for saliva, ZC3H12D and FGF7 for semen (Madi et al., 2012), PFN3A and PRMT2 for vaginal secretions (Anutues et al., 2016), and SLC26A10 for menstrual blood (Lee et al., 2016). Therefore, DNA methylation can be utilized as a new biomarker for body fluid ID. In this study, the goal is to identify new DNA methylation markers to further assist in the determination of the tissue origin of the crime scene stains. The expansion of the number of tissue specific markers should prove particularly useful in mixture resolution and other applications in which enhanced selectivity is important.

2. Challenges and Considerations in Applying DNA Methylations in Forensics

Techniques based on DNA testing are preferable for forensic testing since those techniques provide high sensitivity through PCR and long term stability. In addition, unlike other serological procedures based on protein or RNA, DNA based procedures can fit well into current laboratory workflows. Thus, developing methods that employ DNA methylation markers for different forensic applications could be very advantageous. However, some recommendations should be considered when applying DNA methylation-based methods in forensics.

2.1 Selection of CpG Sites

A very important step in effective DNA-methylation assay is to identify suitable CpG markers to use. There are two approaches for picking the CpG sites. The first approach relies in investigating potential familiar regions such as CpG islands and gene promoters. The second method is to search for suitable CpG sites in a large-scale manner using genome-wide epigenome study (Bock et al., 2009). The second method is more likely to distinguish new genomic sequences that display differential DNA methylations in addition to those that are set outside the promoter regions. Using genome-wide studies often require laborious bioinformatic and statistical computational techniques. These methods are used only to screen and identify candidates of CpG sites, however additional targeted experiments are required to select and validate the most appropriate CpG markers (Bock et al., 2009).

DNA methylation is responsible for wide range of activities inside the cell, thus any exogenous factors including aging, human diseases, stress and cigarette smoking

could alter DNA methylation patterns. Therefore, it is necessary for a particular application to test a wide range of DNA samples to make sure that the selected CpG marker does not exhibit cofounding effects (Vidaki et al., 2013). For such a marker to be suitable for forensic testing, external factors need to be investigated. For instance, to propose that a specific CpG marker is a good indicator to estimate age, we need to verify that specific marker is also not environmental-dependent or affected by certain diseases (Christensen et al., 2009). When utilizing such markers, the methylation level is measured quantitatively and thus it's very important to select CpG markers that display a high percentage of methylation difference between other tissues of factors and the targeted trait. The higher the methylation difference margin, the better the discrimination power will be. For example, to associate a specific CpG with a disease such as diabetes, there should a distinct difference in methylation level between individuals or cells expressing the targeted state and those from the normal healthy population. DNA analysis of such methylation assay should be reproducible,

2.2 Assay Design

Because DNA samples that are encountered at crime scene are often degraded and may be present in low quantities, the methodology for testing the CpGs selected in the first step should be designed carefully in order that the results are suitable for forensic applications. Among the available assays that can measure DNA methylation level, bisulfite modified PCR is very effective and provides appropriate sensitivity (Madi et al., 2012, Alghanim et al., 2017). In this assay, it's very crucial to use the appropriate controls and to pay special attention to PCR

amplification and primer design. The primers need to be designed carefully as such to include some normally converted non-CpG cytosines to ensure that only the bisulfite-converted DNA will be amplified. In addition, primer sequences should not contain any CpG sites unless methylation specific PCR is desired. Amplicon length and sequence are important as differences in size between the methylated and unmethylated amplicons could leads to preferential amplification (Warnecke et al., 1999). One way to resolve the issue of preferential amplification is to normalize the data by utilizing DNA methylation controls with known levels of methylation (Moskalev et al., 2011). In addition, incomplete conversion of unmethylated cytosines has to be taken into consideration as this result in an overestimation of the amount of methylation. After performing bisulfite conversion and assuming that all non-CpG cytosines have also been converted into uracil, and following PCR, thymine, the DNA mainly will consist of three nucleotide bases. The reduced sequence complexity can limit primer design options resulting in potential issues with non-specific amplification due to mispriming events. Thus, to meet the forensic requirement of testing, the selected CpG loci must go through more stringent validation procedures to ensure compatibility with samples of low quality and/or quantity.

III. Techniques of Measuring DNA Methylation

Today, quantitative measurements of DNA methylation levels in single CpG sites, genes, or the entire methylome have become of significant value in various mecial and clinical applications. Most sequence and array-based technologies to examine DNA methylation employ three common methods for methylation detection. Those methods are: (a) methylation sensitive restriction enzymes, (b) protein interactions with 5-methylcytosine, and (c) chemical modification of the unmethylated cytosine residues (Vidaki et al., 2013).

1. Methylation Sensitive Restriction Enzymes

This method depends on the use of restriction enzymes that recognize short sequences in the genome and cleave the DNA at distinct sites within or near these recognition sequences. Some enzymes are sensitive to methylation and when cytosine nucleotides are methylated, will not cleave the DNA at a methylated CpG while other enzymes are insensitive to methylation and will cleave methylated DNA (Hua el al., 2011). McrBC is an example of the enzyme that is insensitive to methylation and cuts the methylated sequence on either one or both stands. This procedure can be used to either enrich for methylated DNA or unmethylated DNA. Then, to determine the methylation level, comparisons are to be made between the two states using different methods. In order to assess the methylation status of a specific CpG site, one method is to compare a difference in methylation level between a sample treated with an enzyme or combination of enzymes and an untreated control. Another method is to compare between a sample treated with

a methylation sensitive enzyme and a control treated with methylation insensitive enzyme. One further option is compare between two different test samples digested with the sample enzyme (Bird et al., 1986).

A Methylation-Sensitive Restriction Enzyme (MSRE) process can be used with PCR and analyzed using capillary electrophoresis (An et al., 2013). This technique involves digestion of the DNA with a methylation sensitive restriction enzyme followed by PCR amplification of surviving fragments. If no PCR product is detected after digestion, then the fragment contains unmethylated cytosine. Alternatively, if a PCR product is obtained, then the fragment said to have methylated cytosine (Figure 1.5). This method is a simple and robust procedure for determining the status of a user defined set of genes (Kader el al. 201. However, incomplete digestion or differences in enzymatic activity can inhibit the analysis. One major issue in the application of this procedure is that the use of restriction enzymes depends in the availability of specific recognition sequences that flank the targeted CpG site (Vidaki et al., 2013).

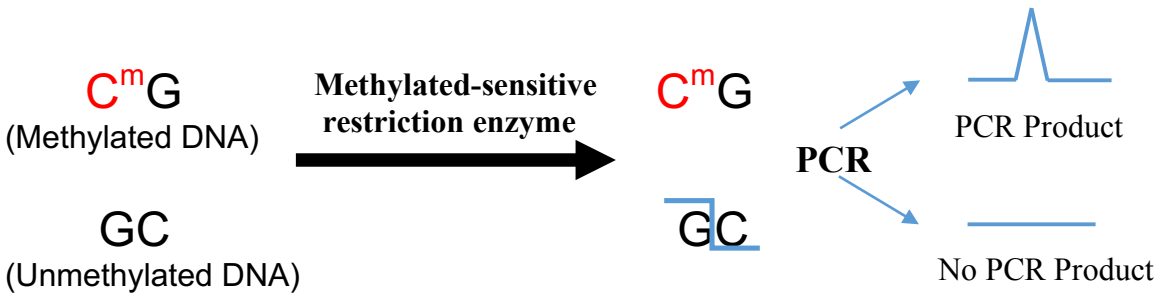


Figure 1.5: Methylation Sensitive Restriction Enzyme combined with PCR. The figure shows methylated-sensitive endonuclease that cleaves the DNA at specific restriction sites (for example: GCGC). The cleavage is possible only at the restriction sites that are not “protected” by the methyl group. Thus, the cleaved unmethylated templates give no PCR product after amplification whereas the methylated templates generate detectable amplicons.

2. Affinity Purification of Methylated DNA

The methylated DNA sites in this method can be isolated by immuno-precipitating the methylated DNA in a process known as Methyl DNA IP or MeDIP. DNA first undergoes fragmentation by restriction enzyme or sonication, followed by denaturation. Shorter fragments are essential in this method to reduce bias, and to improve efficiency and resolution. Beads containing a monoclonal antibody are used to bind 5-methylcytidine for immunoselection and immunoprecipitation. After isolation and purification, the methylated DNA can be analyzed using qPCR, amplification, microarray hybridization, or next generation sequencing (Cheong et al., 2006, Weber et al., 2007).

3. Chemical Modification of Cytosine Residues

The methylation profile of a targeted sequence in the DNA can be determined by sodium bisulfite treatment. Because standard PCR amplification does not preserve DNA methylation sites, an extra step is required to transform DNA methylation information into DNA sequence information (Bock et al., 2012). Bisulfite conversion is considered the “gold standard” for DNA methylation analysis and can provide a qualitative and quantitative measurement of DNA methylation level. Frommer et al. (1992) first introduced the method by determining that amination reactions of cytosine and 5-methylcytosine show different results when treated with sodium bisulfite. The reaction for cytosine deamination using sodium bisulfite involves three steps. The first step is sulfonation in which bisulfite is added to the 5-6 double bond of cytosine. The second step is hydrolytic deamination of

the cytosine–bisulfite derivative to give an uracil-bisulfite derivative. Finally, the alkali desulfonation step, which involves the removal of the sulfonate group producing uracil after treatment with alkali (Patterson et al., 2011). This method converts the unmethylated cytosine into uracil, whereas the methylated cytosine residue remains unchanged. Later, the uracil in the single stranded DNA will be converted to thymine upon PCR amplification and sequencing. However, the methylated cytosine will resist the conversion and will remain as cytosine (Figure 1.6). This sodium bisulfite conversion allows for qualitative and quantitative analysis of a single CpG site. However, correct determination of the tested methylation level depends mainly on the complete conversion of unmethylated cytosines. Bisulfite modification-based techniques are widely used for a variety of DNA methylation analysis methods including chip microarray, pyrosequencing, MALDI-TOF MS, high resolution melt (HRM) analysis, and Ms-SNuPE.

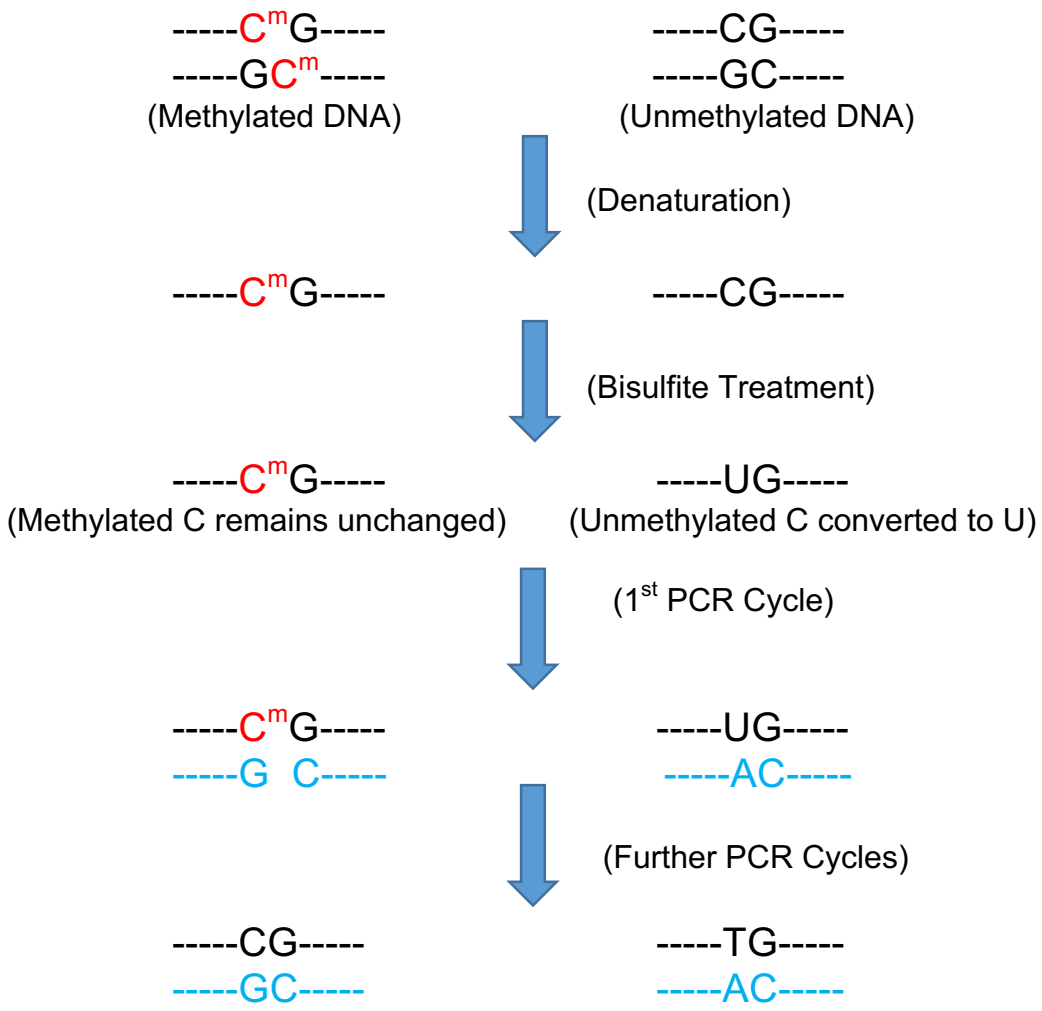


Figure 1.6: show the bisulfite modification step and how methylation information on the sequence is maintained during PCR amplification

3.1 Chip Microarray:

Illumina's Human Methylation BeadChip array is one of the methods best suited for methylation analysis using bisulfite-modified DNA. At present, this technology is capable of rapidly determining the methylation profiles of over 850,000 CpG sites in the human genome (Pidsley et al., 2016). The Illumina Beadarray assays utilize arrays of oligonucleotides attached to beads to measure certain target sequences.

After DNA is treated by sodium bisulfite, the resulting sequences can be hybridize to complementary probes on the array which are linked to fluorescent dyes. The relative methylation level in the DNA can be measured by comparing the fluorescence intensities of the probes targeting unmethylated (U) and methylated allele (M) of a CpG site. Two chemical assays known as Infinium I and Infinum II are utilized which are based on quantitative genotyping of C/T polymorphisms. The Infinium I probe contains separate beads from U and M, while the Infinium II combines U and M on the same beads but the results are detected using different sets of dyes or color channels (Bibikova et al., 2011, Heiss et al. 2019). For example, when a CpG site is fully unmethylated, the U probe displays a very high intensity while the M probe intensity will be low and vice versa. The total fluorescence intensity T, which is basically the sum of the U and M intensities, is independent of the methylation level value to a certain degree. High intensities usually show good signal-to-noise ratio and thus such probes considered to be detectable, while low intensity probes contain mainly background noise and are considered undetectable. However, the method requires large amounts of high quality DNA which is not always possible for forensic samples. In addition, the method may produce an excess of highly personal medical information and can be expensive which hinders its use in the forensic testing.

3.2 MALDI-TOF MS

Matrix-assisted laser desorption ionization time-of-flight (MALDI-TOF) mass spectrometry has been proposed as a method that could be of great use in the detection of DNA sequence variation and in DNA methylation analysis. In contrast with other DNA analysis techniques, mass spectrometry depends on direct mass determination of DNA products rather than on other indirect analyses in which radioactive or fluorescent tags are used for detection and evaluation (Gut et al., 2004). MALDI-TOF MS can provide high throughput identification of methylation sites. Because the technique can directly detect the increase in mass of methylation of DNA, it is also suited for semi-quantitative measurement of single or multiple CpG sites.

MALDI mass analysis can be obtained through two steps. The first step is to dissolved the DNA of interest (the analyte) in a solvent containing in solution of small organic molecules, known as the matrix. These molecules should have strong absorption at the laser wavelength. The mixture is dried and any excess liquid solvent used to prepare the solution is removed. The result is a solid deposit of analyte-doped matrix crystals. The second step takes place under vacuum conditions inside the source of the MS. The solid matrix containing the analyte is ablated by intense laser pulses over a short time. The laser induces rapid heating which leads to ionization reactions. An electric field can accelerate the gas phase ions towards the analyzer. Then, the ionized DNA molecules are separated due to charge and mass as they travel down a time-of-flight analyzer to the detector (Hoffmann et al., 2007). Mass spectra are then collected and analyzed to

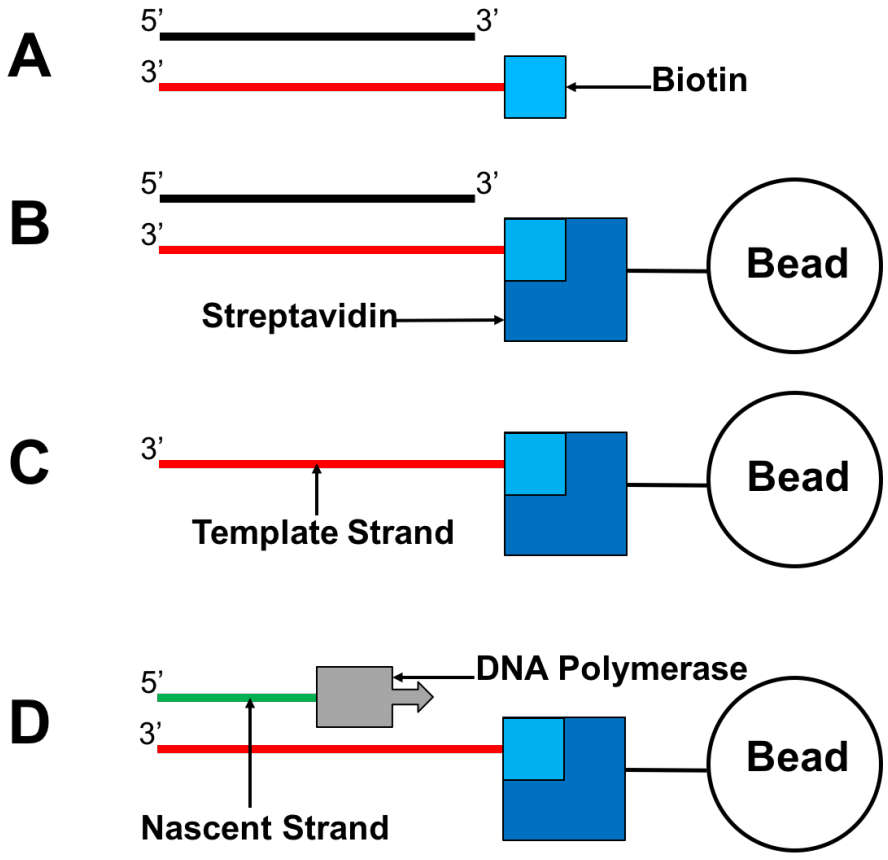
determine fragment sizes, and special software can be used to determine sequence variations. The quantity of two products (methylated versus unmethylated) after bisulfite conversion in turn will result in different fragment masses (Coolen et al., 2007). In addition, quantitative signal is produced which facilitates a measurement of the relative abundance of each fragment of the targeted genetic loci detected base on mass-per-charge ratio (m/z) (Ehrich et al., 2005, Coolen et al., 2007).

3.3 Pyrosequencing

Pyrosequencing is considered a sequencing-by-synthesis technique because the DNA synthesis is monitored in real time. The technique was first introduced by Pal Nyren in 1987. Initially, Nyren invented the method in order to continuously monitor DNA polymerase activity (Nyren et al., 1987). A year later, Hyman relied on Nyren's work to invent pyrosequencing (Hyman et al., 1988). However, it took several more years to commercialize the method and ultimately it was widely implemented (Harrington et al., 2013).

Before running a pyrosequencing reaction, the region of interest should be amplified with a primer set in which one oligonucleotide is biotinylated (Figure 1.7A). This enables the amplicons to immobilized tightly onto magnetic beads coated with streptavidin using avidin-biotin bonding (Figure 1.7B). Streptavidin is a tetrameric protein that can bind up to four biotin molecules per a streptavidin molecule (Diamandis et al., 1991). By applying a magnetic field, the magnetic beads can be immobilized against the wall of a tube. The biotinylated template strand of the double stranded PCR is then isolated via denaturation and washing

steps (Figure 1.7C). The strand captured by the bead then becomes the template strand for the sequencing primer and undergoes the four enzymatic reactions of pyrosequencing (Figure 1.7D). During this process, the complementary DNA strand is synthesized on the template strand which is attached to the bead. In standard pyrosequencing, the Klenow fragment of Escherichia coli DNA Pol I is the DNA polymerase utilized (Benkovic and Cameron et al., 1995). This enzyme is preferred as it was found to empirically decrease background signals (Klenow et al., 1971, Nordstrom et al., 2000).



Figures 1.7A-1.7D: shows the biotin-streptavidin interaction on the bead and start of the pyrosequencing reaction.

Following the PCR amplification, the sequencing primer is first hybridized to a single stranded template biotin-labeled DNA to be sequenced. The four enzymes used in pyrosequencing include DNA polymerase, ATP sulfurylase, luciferase, and apyrase. Two substrates are used: adenosine 5′ phosphosulfate (APS) and luciferin (Gharizadeh et al., 2007). DNA polymerase is used to extend the 3' end of the nascent primer strand based on the DNA sequence of the template strand. During each step of the reaction, the addition of a specific dNTP is sequentially performed into the DNA mixture. A cascade of enzymatic reactions start when the correct complementary dNTPs is incorporated by the polymerase. In the first step, an inorganic pyrophosphate (PPi) is released during the polymerization reaction as shown below:

$(DNA)_n + dNTP \rightarrow (DNA)^{n+1} + PPi + H^+$ -- (catalyzed by polymerase)

Each incorporation of nucleotide is followed by the release of a PPi in a quantity proportional (equimolar) to the amount of added nucleotide. Once generated, the PPi, in the presence of APS, is quantitatively converted to adenosine triphosphate (ATP) by ATP sulfurylase.

$PPi + APS \rightarrow ATP + SO_4^{-2}$ ---- (catalyzed by ATP sulfurylase)

The generated ATP serves as a cofactor for the enzyme luciferase to oxidize the luciferin to oxyluciferin and light:

$ATP + Luciferin + O_2 \rightarrow AMP + PPi + oxyluciferin + CO_2 + hv$ –(catalyzed by luciferase)

Therefore, amount of light emitted is proportional to the amount of PPi produced, which is directly proportional to the number of nucleotides being added. This light

signal at a wavelength of 560nm is detected by a charge coupled device (CCD) or photomultiplier. The role of Apyrase is to continuously degrade unincorporated nucleotides and ATP following the addition of each base.

ATP + dNTPs → AMP + dNMP + 4Pi --- (catalyzed by Apyrase)

A time interval (usually 65 seconds) is provided between each nucleotide dispensation to allow complete removal of any excess dNTPs and ATP from the reaction mixture before the next nucleotide dispensation starts. This is crucial to ensure that the light detected when adding a certain nucleotide only results from the incorporation of a specific nucleotide. Because the added base is known, the sequence of the template can be resolved (Ronaghi et al., 2001, Gharizadeh et al., 2007). A schematic representation of four enzymatic reactions of pyrosequencing is shown in Figure 8.

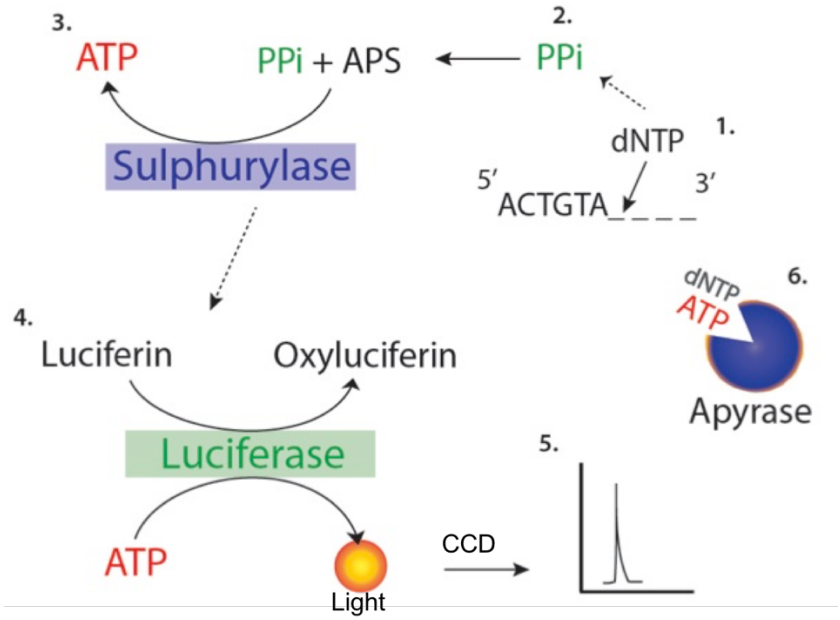


Figure 1.8: The pyrosequencing depends on four enzymatic reactions. The results will be recoded via CCD and shown as pyrogram (adopted from Website1).

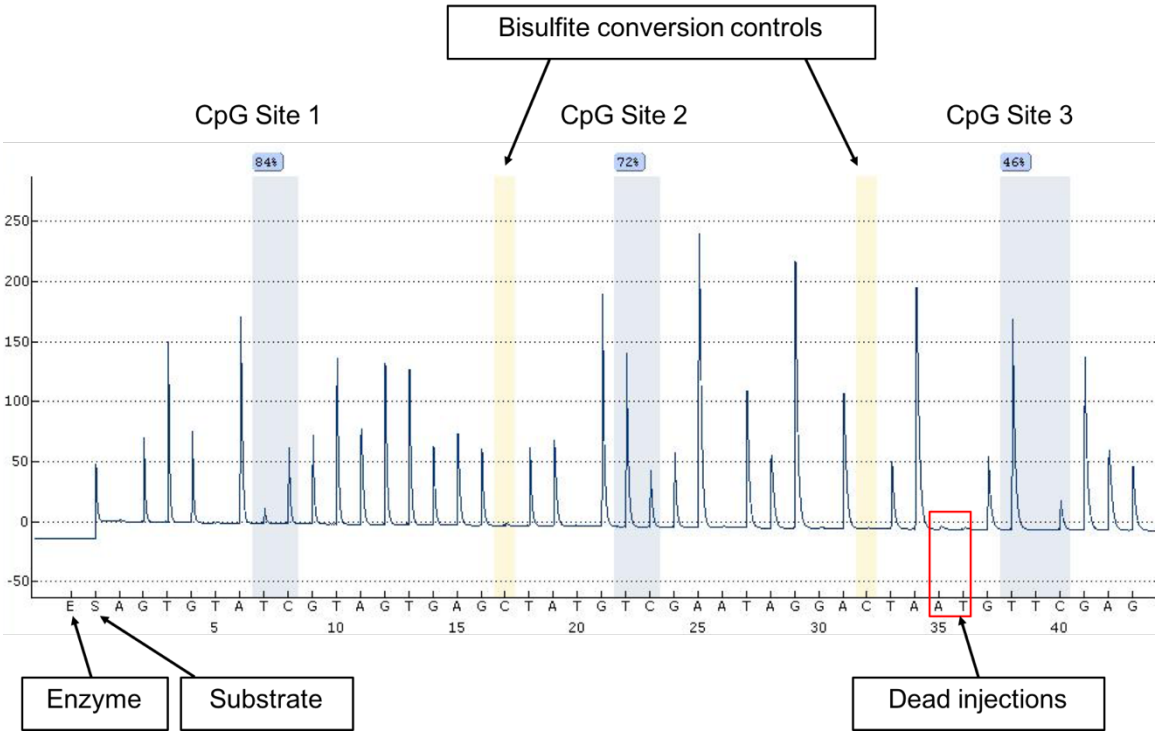


Figure 1.9: Pyrogram depicting the relative peak height for the targeted sequence. The relative peak height for two repeated nucleotides (for example, “TT” or “GG) would be two times as high as the peak produced by only single nucleotides. Yellow shaded areas indicate the bisulfite conversion control to insure full conversion. If the bisulfite conversion or the original genomic DNA is not complete, then some “C” peak will be detected here. The red box illustrates the dead injections and the blue shaded areas show the CpG sites. At each CpG site, both “T” and “C” are released and peak heights are compared to quantify the level of methylation.

Pyrosequencing is a very effective method to detect and quantify multiple CpG sites within a single reaction. The degree of methylation (% methylation) at each CpG position in the target sequence is determined from the ratio of unmethylated templates (T) and methylated templates (C) (Figure 1.9). The amplicon lengths for the pyrosequencing assays optimized for up to 200 bp (Qiagen Handbook1). However, when sequencing a long amplicon (> 50 bp), it’s common to experience

lower peak signal over time. In addition, careful assay design for pyrosequencing is needed to minimize artefacts that may arise due to biotinylated primer interference, non-specific binding of sequencing primer, or hairpin formation. The peak height of each nucleotide in the pyrogram corresponds to the number of nucleotides that are incorporated at each injection (Figure 1.9). Occasionally during an assay and especially just before the CpG site, a nucleotide or more that is not expected to be incorporated (also referred to as dead injection) may be dispensed to eliminate any possible background interference. A minimum of one bisulfite-treatment control (non-CpG cytosine) is added in the assay dispensation reaction to make sure that the DNA underwent complete bisulfite conversion (Figure 1.9). The pyrosequencing process is simple and the sequencing cost per reaction is comparable to that of the medium-throughput techniques such as Sanger sequencing. One of the main advantages is the flexibility of the method for application to wide range of applications and the ability to maintain reproducibility between different laboratories. Pyrosequencing is also suitable for forensic samples which may contain degraded and low quality DNA. In addition, the flexibility of the assay design and the analysis make the system very effective for medium-throughput genotyping and automation which meet the needs of most research laboratories (Lehmann et al., 2015).

3.4 High Resolution Melt (HRM) Analysis

High Resolution Melt analysis is a method that was first described by Wittwer et al. in 2003 and then adopted by Wojdacz and collogues for the analysis of DNA methylation (Wojdacz et al., 2007). HRM analysis can distinguish between DNA samples based on their dissociation properties. The direct melting process separates a single dsDNA molecule into its two complementary strands using a gradually increasing temperature. Variations in sequence length, GC content and strand complementarity can all be detected by monitoring the melt rate of the two strands.

DNA methylation profiling using HRM analysis starts with bisulfite treatment. Because bisulfite modified PCR transforms unmethylated cytosines into thymine, there is a difference in hydrogen bonding due to changes in GC content between methylated and unmethylated DNA. This is due to the fact that the unmenthylated template would be mainly composed of double hydrogen bonds due to base pairing of adenine to thymine, whereas the methylated amplicon would contain mainly triple hydrogen bonds between guanine and cytosine and thus higher melting temperature (Figure 1.10).

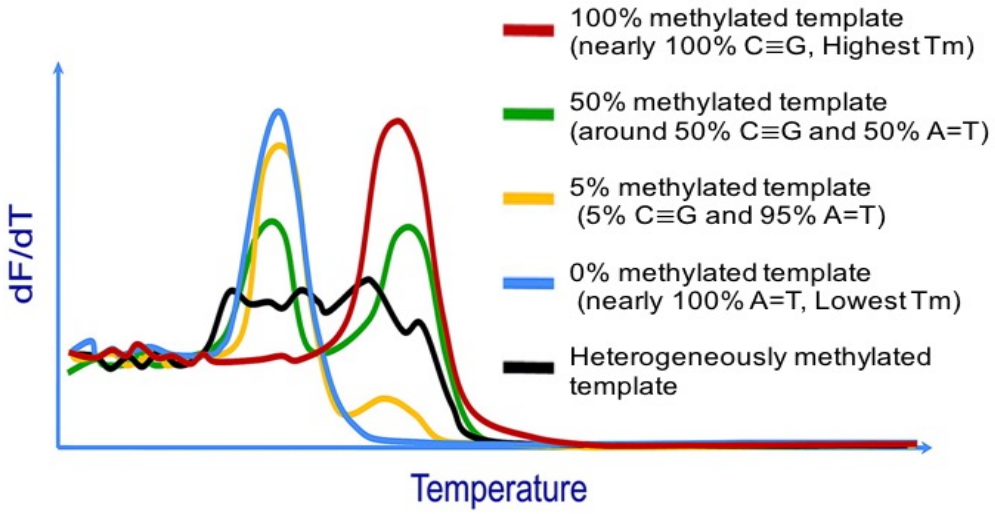


Figure 1.10: After bisulfide treatment, the methylated template is mainly GC rich (C≡G, whereas the unmethylated template would consists mostly of A=T. Thus, the methylated DNA would be more resistant to melting and would melt at higher temperature (red curve) and lower melting temperature for the unmethylated DNA (blue curve). Two peaks would be expected (green curve) when the amount of methylated and unmethylated templates is equal (nearly 50% each) (adopted from Kristensen et al., 2009).

The method is performed using a single tube process in which the DNA region containing the targeted methylation site is amplified by PCR in the presence of a fluorescence intercalating dye. This dye fluoresces only in the presence of double stranded DNA. As the template concentration increases in the reaction mixture, the fluorescence intensity exhibited by the dsDNA increase. The melt analysis typically performed in a real-time PCR instrument and carried out immediately following the end of the PCR. The melting protocol starts by heating the PCR products gradually from 60 to 95 in increments of 0.1-0.3 C. This gradual increase in temperature causes dsDNA to denature to ssDNA which in turn results in a release of the intercalating dye and a decrease in fluorescence. As the temperature continue to increase, the fluorescence decrease at a steady pace,

followed by sharp decrease around the melting point (Tm) of the DNA template (Figure 1.11). The melting temperature (Tm) is defined as the temperature at which 50% of the dsDNA has been melted (50% dissociation) which roughly corresponds to half of the fluorescence detected as showing in Figure 1.11 (Reed el al., 2007). In general, the software that analyzes the melt curve assigns the Tm value of an amplicon as the inflection point of the melt curve.

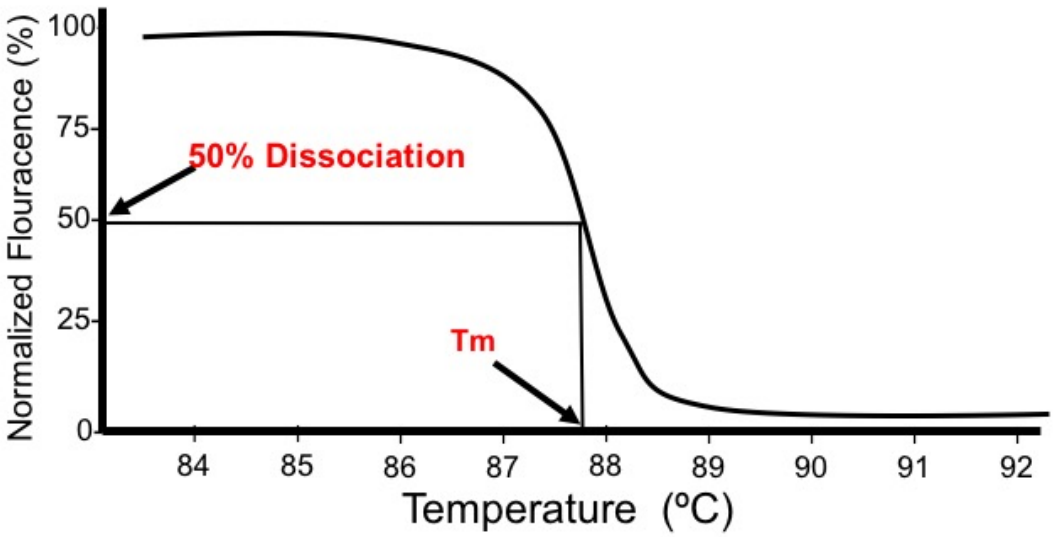


Figure 1.11: The curve illustrates a typical melt curve of dsDNA to a ssDNA. The graduate increase in temperature causes the dsDNA to denature to ssDNA and the sudden drop in temperature around the Tm. The figure show the Tm at which 50% of the dsDNA has been melted (50% dissociation) that roughly corresponds to 50% fluorescence detection of the intercalating dye (half release of the dye).

In addition, Tm can be pinpointed more efficiently by plotting the negative first derivative of melt-curve. This represents the rate of change of fluorescence in the amplification reaction over temperature (-dF/dT), observed change in slope, versus the temperature (Hernández et al., 2003, Ahmed et al., 2017, Sun et al.,

2016) (Figure 1.12). HRM is a rapid and relatively inexpensive method that only requires the use of unlabeled primer sets and one type of intercalating dye (Hanson et al., 2013). The technology employs the use of a real-time PCR, an instrument available in most forensic laboratories. HRM is a nondestructive method, thus any subsequent testing after melt analysis can still be done such as electrophoresis or sequencing (Vossen et al., 2009). HRM also has the capability to perform multiplex analysis of several amplicons in a single tube (Seipp et al., 2009). However, there are some disadvantages to HRM analysis. For example, HRM does not permit the methylation level at individual CpG sites to be qualitied. Moreover, the amplicon's DNA secondary structure following bisulfite treatment can interfere with the HRM analysis which making the results difficult to interpret (Wojdacz et al., 2012) and intermediate levels of methylation can produce non-specific results.

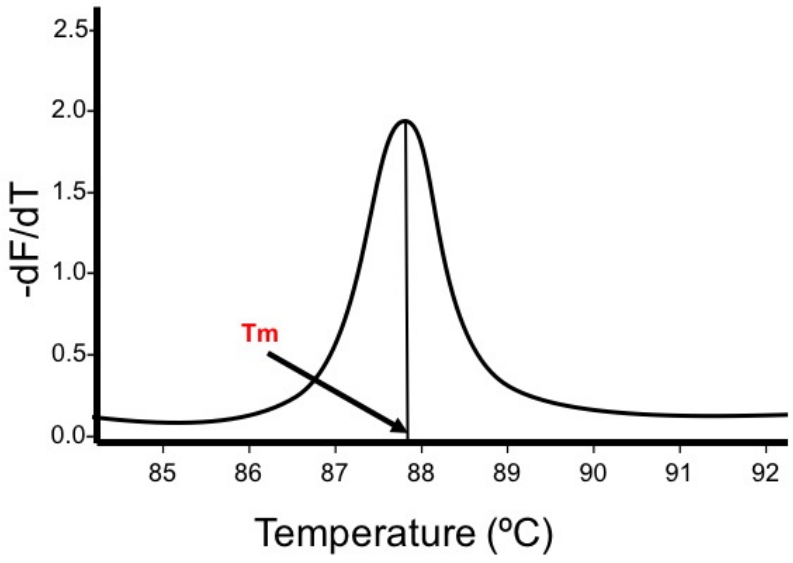


Figure 1.12: Plotting a negative first derivative of the rate of change of fluorescence over temperature (-dF/dT), observed change in slope, versus the temperature. The plot shows the inflection point on the slopes as a more easily visualized melt curve to pinpoint the Tm.

3.5 Methylation Sensitive Single Nucleotide Primer Extension (Ms-SNuPE)

The SNuPE method was originally developed to detect single nucleotide mutations (Gonzalgo et al., 1997) but later directed towards DNA methylation analyses (Gonzalgo et al., 2007, Kristensen et al., 2009). The method can be assayed on capillary electrophoresis (CE) instrumentation. The procedure is similar to the SNaPshot assay, which is widely used for SNP studies (Kristensen et al., 2009). Ms-SNuPE is a technique that can be utilized for direct quantitation of methylation at individual CpG sites (Gonzalgo et al., 2007). Following treatment of genomic DNA with sodium bisulfite, the target sequence is PCR amplified using specially designed primers specific for bisulfite-modified DNA. The amplified product is then purified to remove any excess unused primers and dNTPs. After that, an oligonucleotide is added that is designed to anneal to the PCR product upstream from the CpG site being investigated and terminates immediately 5' to it (Gonzalgo et al., 2007). Once the primer has annealed, a single base extension reaction is performed in the presence of DNA polymerase and dideoxy-modified nucleotides (ddGTP and ddATP). Since these dideoxy nucleotides do not allow for further extension and can be fluorescently labeled, the fragments can be separated by capillary electrophoresis instrument (Figure 1.13). The CE instrument can distinguish the unmethylated template fragments which end with 'T' from the

methylated strands that end with 'C'. The peak height provides relative quantitation of each fragments because each nucleotide is labeled with different fluorophores. This technology is very practical since it requires small amounts of DNA and utilizes instrumentation that is familiar to the forensic analyst. However, the methodology involves two separate reactions - the initial PCR amplification and the single base extension reaction - which may hinder throughput. In addition, the assay is limited to a relatively low number of CpG sites that can be detected and quantified (up to four sites) per reaction (Gonzalgo et al., 2007).

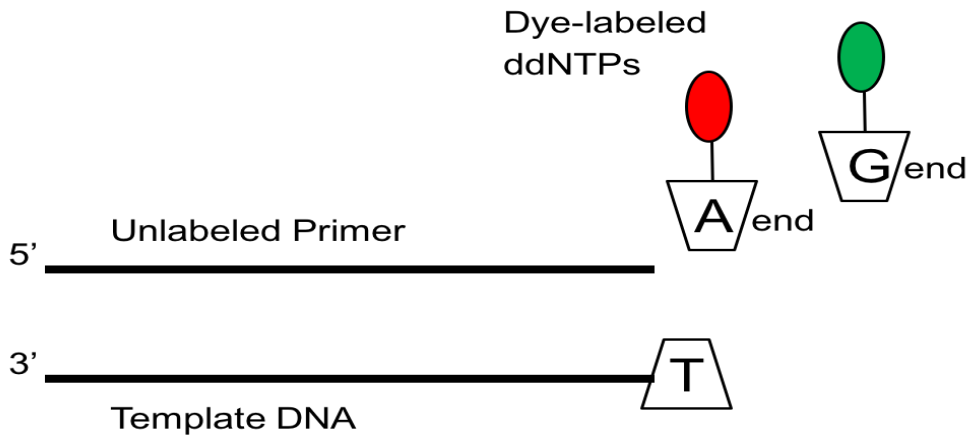


Figure 1.13: Single base extension with dye-labeled dideoxy-modified nucleotides (ddNTPs).

List of References

Aguilera O, Fernandez AF, Munoz A, Fraga MF. Epigenetics and environment: A complex relationship. J Appl Physiol. 2010;109(1):243-51.

Ahmed FE, Gouda MM, Hussein LA, Ahmed NC, Vos PW, Mohammad MA. Role of melt curve analysis in interpretation of Nutrigenomics' MicroRNA expression data. Cancer Genomics-Proteomics, 2017;14(6), 469-481.

Alghanim, H., Antunes, J., Silva, D. S. B. S., Alho, C. S., Balamurugan, K., & McCord, B. Detection and evaluation of DNA methylation markers found at SCGN and KLF14 loci to estimate human age. Forensic Sci. Int. Genet., 2017;31, 81-88.

An JH, Choi A, Shin KJ, Yang WI, & Lee, HY. DNA methylation-specific multiplex assays for body fluid identification. Int. J. Legal Med., 2013;127(1), 35-43.

Antunes, J., Silva, D. S., Balamurugan, K., Duncan, G., Alho, C. S., & McCord, B. (2016). Forensic discrimination of vaginal epithelia by DNA methylation analysis through pyrosequencing. Electrophoresis, 37(21), 2751-2758.

Bahar R, Hartmann CH, Rodriguez KA, Denny AD, Busuttil RA, Dollé ME, et al. Increased cell-to-cell variation in gene expression in ageing mouse heart. Nature. 2006;441(7096):1011-4.

Bekaert B, Kamalandua A, Zapico SC, Van de Voorde W, Decorte R. Improved age determination of blood and teeth samples using a selected set of DNA methylation markers. Epigenetics. 2015;10(10):922-30.

Bell JT, Pai AA, Pickrell JK, Gaffney DJ, Pique-Regi R, Degner JF, et al. DNA methylation patterns associate with genetic and gene expression variation in HapMap cell lines. Genome Biol. 2011;12(1):R10.

Benkovic SJ. and Cameron CE. Kinetic analysis of nucleotide incorporation and misincorporation by Klenow fragment of Escherichia coli DNA polymerase I. Methods Enzymol. 1995: 262: 257-269.

Berdyshev GD, Korotaev GK, Boiarskikh GV, Vaniushin BF. Nucleotide composition of DNA and RNA from somatic tissues of humpback and its changes during spawning. Biokhimiia. 1967 Sep-Oct;32(5):988-93.

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;98(4):288-95.

Bird A. DNA methylation patterns and epigenetic memory. Genes. Dev. 2002;16(1):6-21.

Bird, AP. CpG-rich islands and the function of DNA methylation. Nature, 1986;321 (6067); 209.

Bjornsson HT, Sigurdsson MI, Fallin MD, Irizarry RA, Aspelund T, Cui H, et al. Intra-individual change over time in DNA methylation with familial clustering. JAMA. 2008;299(24):2877-83.

Bock C. Analysing and interpreting DNA methylation data. Nat. Rev. Genet. 2012;13(10):705-19.

Bocklandt S, Lin W, Sehl ME, Sánchez FJ, Sinsheimer JS, Horvath S, et al. Epigenetic predictor of age. PloS one. 2011;6(6):e14821.

Bollati V, Schwartz J, Wright R, Litonjua A, Tarantini L, Suh H, et al. Decline in genomic DNA methylation through aging in a cohort of elderly subjects. Mech Ageing Dev. 2009;130(4):234-9.

Borghol N, Suderman M, McArdle W, Racine A, Hallett M, Pembrey M, et al. Associations with early-life socio-economic position in adult DNA methylation. Int J Epidemiol. 2012;1:62–74.

Brait M, Sidransky D. Cancer epigenetics: Above and beyond. Toxicol. Mech. Metheods. 2011;21(4):275-88.

Breitling LP, Yang R, Korn B, Burwinkel B, Brenner H. Tobacco-smoking-related differential DNA methylation: 27K discovery and replication. Am. J. Hum. Genet. 2011;88(4):450-7.

Casillas MA, Lopatina N, Andrews LG, Tollefsbol TO. Transcriptional control of the DNA methyltransferases is altered in aging and neoplastically-transformed human fibroblasts. Mol. Cell Biochem. 2003;252(1):33-43.

Cheong J, Yamada Y, Yamashita R, Irie T, Kanai A, Wakaguri H, ... & Suzuki Y. Diverse DNA methylation statuses at alternative promoters of human genes in various tissues. DNA Rer., 2006;13(4), 155-167.

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. PLoS Genet. 2009;5(8):e1000602.

Coolen, MW, Statham AL, Gardiner-Garden M, & Clark SJ. Genomic profiling of CpG methylation and allelic specificity using quantitative high-throughput mass spectrometry: critical evaluation and improvements. Nucleic Acids Res. 2007;35(18), e119.

Dahl C, Grønbæk K, Guldberg P. Advances in DNA methylation: 5-hydroxymethylcytosine revisited. Clinica Chimica. Acta. 2011;412(11):831-6.

Day, K., Waite, L. L., Thalacker-Mercer, A., West, A., Bamman, M. M., Brooks, J. D., ... & Absher, D. (2013). Differential DNA methylation with age displays both common and dynamic features across human tissues that are influenced by CpG landscape. Genome Biol., 2013;14(9), R102.

Di YP, Zhao J, Harper R. Cigarette smoke induces MUC5AC protein expression through the activation of Sp1. J. Biol. Chem. 2012 Aug 10;287(33):27948-58.

Diamandis EP, Christopoulos TK. The biotin-(strept)avidin system: principles and applications in biotechnology. Clin. Chem. 1991;37(5):625–636.

Dogan MV, Shields B, Cutrona C, Gao L, Gibbons FX, Simons R, et al. The effect of smoking on DNA methylation of peripheral blood mononuclear cells from african american women. BMC Genomics. 2014;15(1):151.

Eckhardt F, Lewin J, Cortese R, Rakyan VK, Attwood J, Burger M, et al. DNA methylation profiling of human chromosomes 6, 20 and 22. Nat Genet. 2006;38(12):1378-85

Ehrich, M, Nelson MR, Stanssens P, Zabeau M, Liloglou T, Xinarianos G, ... & van den Boom D. Quantitative high-throughput analysis of DNA methylation patterns by base-specific cleavage and mass spectrometry. PNAS, 2005;102(44), 15785-15790.

Ehrlich, M., and Lacey, M. (2013). DNA hypomethylation and hemimethylation in cancer. Adv. Exp. Med. Biol. 754, 31–56

Enokida H, Shiina H, Urakami S, Terashima M, Ogishima T, Li L, et al. Smoking influences aberrant CpG hypermethylation of multiple genes in human prostate carcinoma. Cancer. 2006;106(1):79-86.

Epsztejn-Litman, S., Feldman, N., Abu-Remaileh, M., Shufaro, Y., Gerson, A., Ueda, J., et al. (2008). De novo DNA methylation promoted by G9a prevents reprogramming of embryonically silenced genes. Nat. Struct. Mol. Biol. 15, 1176–1183.

Feng S, Jacobsen SE, Reik W. Epigenetic reprogramming in plant and animal development. Science. 2010 Oct 29;330(6004):622-7.

Ficz G, Branco MR, Seisenberger S, Santos F, Krueger F, Hore TA, et al. Dynamic regulation of 5-hydroxymethylcytosine in mouse ES cells and during differentiation. Nature. 2011;473(7347):398-402.

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

Fragou D, Fragou A, Kouidou S, Njau S, Kovatsi L. Epigenetic mechanisms in metal toxicity. Toxicol. Mech. Metheods., 2011;21(4):343-52.

Fraser HB, Lam LL, Neumann SM, Kobor MS. Population-specificity of human DNA methylation. Genome Biol. 2012;13(2):R8.

Frigola J, Song J, Stirzaker C, Hinshelwood RA, Peinado MA, Clark SJ. Epigenetic remodeling in colorectal cancer results in coordinate gene suppression across an entire chromosome band. Nat Genet. 2006;38(5):540-9.

Frommer M, McDonald LE, Millar, DS, Collis CM, Watt F, Grigg GW, ... & Paul CL. A genomic sequencing protocol that yields a positive display of 5-methylcytosine residues in individual DNA strands. PNAS, 1992;89(5), 1827-1831.

Frumkin D, Wasserstrom A, Budowle B, Davidson A, DNA methylation-based forensic tissue identification. Forensic Sci. Int. Genet. 2011;5:517–524

Frumkin D, Wasserstrom A, Davidson A, Grafit A. Authentication of forensic DNA samples. Forensic Sci. Int. Genet. 2010;4(2):95-103.

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;11(6):1132-4.

Gharizadeh B., Ghaderi M. and Nyren P. Pyrosequencing technology for short DNA sequencing and whole genome sequencing. Technology. 2007: 47: 129-132.

Gluckman PD, Hanson MA, Pinal C. The developmental origins of adult disease. Matern. Child Nutr., 2005;1(3):130-41.

Godfrey KM, Barker DJ. Fetal programming and adult health. Public Health Nutr. 2001;4(2b):611-24.

Goldberg AD, Allis CD, Bernstein E. Epigenetics: A landscape takes shape. Cell. 2007;128(4):635-8.

Gonzalgo ML, Jones PA. Rapid quantitation of methylation differences at specific sites using methylation-sensitive single nucleotide primer extension (Ms-SNuPE). Nucleic Acids Res., 1997;25:2529-2531.

Gonzalgo, ML, & Liang, G. Methylation-sensitive single-nucleotide primer extension (Ms-SNuPE) for quantitative measurement of DNA methylation. Nature protocols, 2007;2(8), 1931.

Gut, IG. DNA analysis by MALDI- TOF mass spectrometry. Hum. Mutat., 2004;23(5), 437-441.

Han L, Su B, Li W, Zhao Z. CpG island density and its correlations with genomic features in mammalian genomes. Genome Biol. 2008;9(5):R79.

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. Mol Cell. 2013;49(2):359-67.

Hanson EK, & Ballantyne J. Rapid and inexpensive body fluid identification by RNA profiling-based multiplex High Resolution Melt (HRM) analysis. F1000Research, 2013;2.

Harrington CT, Lin EI, Olson MT, Eshleman JR (2013) Fundamentals of pyrosequencing. Arch Pathol Lab Med 137:1296–1303

Heiss, JA, & Just AC. Improved filtering of DNA methylation microarray data by detection p values and its impact on downstream analyses. Clin. epigenetics, 2019;11(1), 15.

Hernandez M, Rodriguez-Lazaro D, Esteve T, Prat S, & Pla M. (2003). Development of melting temperature-based SYBR Green I polymerase chain reaction methods for multiplex genetically modified organism detection. Anal. Biochem., 2003;323(2), 164-170.

Hoffmann E, and Stroobant V. Mass spectrometry: principles and applications. England: West Sussex (2007).

Horvath S. DNA methylation age of human tissues and cell types. Genome Boil., 2013;14(10): 3156.

Hua D, Hu Y, Wu YY, Cheng ZH, Yu J, Du X, & Huang, ZH. Quantitative methylation analysis of multiple genes using methylation-sensitive restriction enzyme-based quantitative PCR for the detection of hepatocellular carcinoma. Experimental and molecular pathology, 2011;91(1), 455-460.

Huang J, Okuka M, Lu W, Tsibris JC, McLean MP, Keefe DL, et al. Telomere shortening and DNA damage of embryonic stem cells induced by cigarette smoke. Reproductive toxicology. 2013;35:89-95

Hyman ED. A new method of sequencing DNA. Anal. Biochem. 1988;174(2):423–436.

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. 2012; 120(10), 1425-1431.

Jung M, Pfeifer GP. Aging and DNA methylation. BMC biology. 2015;13(1):7.

Kader F, Ghai M. DNA methylation and application in forensic sciences. Forensic Sci Int. 2015; 249:255-65.

Klenow H, Overgaard-Hansen K, Patkar SA. Proteolytic cleavage of native DNA polymerase into two different catalytic fragments: influence of assay conditions on the change of exonuclease activity and polymerase activity accompanying cleavage. Eur J Biochem. 1971;22(3):371–381.

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

Kristensen, LS, & Hansen LL. PCR-based methods for detecting single-locus DNA methylation biomarkers in cancer diagnostics, prognostics, and response to treatment. Clin. Chemistry, 2009; 55(8), 1471-1483.

Lee HY, Jung S, Oh YN, Choi A, Yang WI, Shin K. Epigenetic age signatures in the forensically relevant body fluid of semen: A preliminary study. Forensic Sci Int Genet. 2015;19:28-34.

Lee HY, Park MJ, Choi A, An JH, Yang WL, and Shin KJ. Potential forensic application of DNA methylation profiling to body fluid identification. Int. J. Legal Med.,: 2012;126,(1): 55-62.

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

Lee, H. Y., Jung, S. E., Lee, E. H., Yang, W. I., & Shin, K. J.. DNA methylation profiling for a confirmatory test for blood, saliva, semen, vaginal fluid and menstrual blood. Forensic Sci. Int. Genet, 2016;24, 75-82.

Lehmann U, Tost J. Pyrosequencing-methods and Protocols. Second Edition ed. Springer Protocols; 2015.

Levin HL, Moran JV. Dynamic interactions between transposable elements and their hosts. Nature Rev. Genet., 2011;12(9):615-27.

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

Lienert F, Wirbelauer C, Som I, Dean A, Mohn F, and Schübeler D. Identification of genetic elements that autonomously determine DNA methylation states. Nature Genet. 2011; 43(11): 1091.

Lillycrop KA, Hoile SP, Grenfell L, Burdge GC. DNA methylation, ageing and the influence of early life nutrition. Proc Nutr Soc. 2014;73(03):413-21.

Lim JP, Brunet A. Bridging the transgenerational gap with epigenetic memory. Trends Genet. 2013;29(3):176-86.

Liu Q, Liu L, Zhao Y, Zhang J, Wang D, Chen J, et al. Hypoxia induces genomic DNA demethylation through the activation of HIF-1alpha and transcriptional upregulation of MAT2A in hepatoma cells. Mol. Cancer Ther. 2011 Jun;10(6):1113-23.

Madi T, Balamurugan K, Bombardi R, Duncan G, McCord B. The determination of tissue- specific DNA methylation patterns in forensic biofluids using bisulfite modification and pyrosequencing. Electrophoresis. 2012;33(12):1736-45.

Marsit CJ, Houseman EA, Schned AR, Karagas MR, Kelsey KT. Promoter hypermethylation is associated with current smoking, age, gender and survival in bladder cancer. Carcinogenesis. 2007 Aug;28(8):1745-51.

Maunakea AK, Nagarajan RP, Bilenky M, Ballinger TJ, D'Souza C, Fouse SD, et al. Conserved role of intragenic DNA methylation in regulating alternative promoters. Nature. 2010;466(7303):253-7.

Medvedeva YA, Fridman MV, Oparina NJ, Malko DB, Ermakova EO, Kulakovskiy IV, et al. Intergenic, gene terminal, and intragenic CpG islands in the human genome. BMC Genomics. 2010;11(1):48.

Monick MM, Beach SR, Plume J, Sears R, Gerrard M, Brody GH, et al. Coordinated changes in AHRR methylation in lymphoblasts and pulmonary macrophages from smokers. Ame. J. Med. Genet. B Neuropsychiatr. Genet., 2012;159B(2):141-51.

Moskalev EA, Zavgorodnij MG, Majorova SP, Vorobjev IA, Jandaghi P, Bure IV, et al. Correction of PCR-bias in quantitative DNA methylation studies by means of cubic polynomial regression. Nucleic Acids Res. 2011 Jun;39(11):e77.

Naito E, Dewa K, Yamanouchi H, Takagi S, Kominami R. Sex determination using the hypomethylation of a human macro-satellite DXZ4 in female cells. Nucleic Acids Res. 1993 May 25;21(10):2533-4.

Nielsen DA, Yuferov V, Hamon S, Jackson C, Ho A, Ott J, et al. Increased OPRM1 DNA methylation in lymphocytes of methadone-maintained former heroin addicts, Neuropsychopharmacology, 2009;34: 867-873.

Nordstrom T, Nourizad K, Ronaghi M, Nyren P. Method enabling pyrosequencing on double-stranded DNA. Anal Biochem. 2000;282(2):186–193.

Nyren P. Enzymatic method for continuous monitoring of DNA polymerase activity. Anal. Biochem., 1987;167(2):235-238.

Okano M, Bell DW, Haber DA, Li E. DNA methyltransferases Dnmt3a and Dnmt3b are essential for de novo methylation and mammalian development. Cell. 1999;99(3):247-57.

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. 2011;478(7368):197-203.

Park J, Kim JH, Seo E, Bae DH, Kim S, Lee H, et al. Identification and evaluation of age-correlated DNA methylation markers for forensic use. Forensic Sci. Int. Genet. 2016;23:64-70.

Park JL, Kwon OH, Kim JH, Yoo HS, Lee HC, Woo KM, Kim SY, Lee SH, Kim YS. Identification of body fluid-specific DNA methylation markers for use in forensic science. Forensic Sci Int Genet 2014;13:147–153

Patterson K, Molloy L, Qu W, & Clark S. DNA methylation: bisulphite modification and analysis. JoVE, 2001;(56), e3170.

Philibert R, Plume JM, Gibbons FX, Brody GH, Beach S. The impact of recent alcohol use on genome wide DNA methylation signatures. Front. Genet. 2012;3:54.

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 Biol., 2016;17(1), 208.

Portela A, Esteller M. Epigenetic modifications and human disease. Nat. Biotechnol. 2010;28(10):1057-68.

Qiagen Handbook1: PyroMark Q48 Autoprep User Manual, 6/2016

Rakyan VK, Down TA, Balding DJ, Beck S. Epigenome-wide association studies for common human diseases. Nature Rev. Genet.. 2011;12(8):529-41

Rakyan VK, Hildmann T, Novik KL, Lewin J, Tost J, Cox AV, et al. DNA methylation profiling of the human major histocompatibility complex: A pilot study for the human epigenome project. PLoS Biol. 2004;2(12):e405.

Rando, O. J., & Verstrepen, K. J. Timescales of genetic and epigenetic inheritance. Cell. 2007;128(4), 655-668.

Reed, GH, Kent JO, & Wittwer CT. High-resolution DNA melting analysis for simple and efficient molecular diagnostics. Pharmacogenomics. 2007;8:597-608.

Ronaghi, M. (2001). Pyrosequencing sheds light on DNA sequencing. Genome Res., 11(1), 3-11.

Satta R, Maloku E, Zhubi A, Pibiri F, Hajos M, Costa E, et al. Nicotine decreases DNA methyltransferase 1 expression and glutamic acid decarboxylase 67 promoter methylation in GABAergic interneurons. PNAS. 2008 Oct 21;105(42):16356-61.

Seipp MT, Durtschi JD, Voelkerding KV, et al.: Multiplex amplicon genotyping by high-resolution melting. J Biomol Tech. 2009; 20(3): 160–164

Smith ZD, Chan MM, Mikkelsen TS, Gu H, Gnirke A, Regev A, et al. A unique regulatory phase of DNA methylation in the early mammalian embryo. Nature. 2012;484(7394):339-44.

Stewart L, Evans N, Bexon KJ, van der Meer, Dieudonne J, Williams GA. Differentiating between monozygotic twins through DNA methylation-specific high-resolution melt curve analysis. Anal. Biochem. 2015;476:36-9.

Straussman R, Nejman D, Roberts D, Steinfeld I, Blum B, Benvenisty N, et al. Developmental programming of CpG island methylation profiles in the human genome. Nature Struct. Mol. Boil., 2009;16(5):564-71.

Sun W, Li JJ, Xiong C, Zhao B, & Chen SL. The potential power of Bar-HRM technology in herbal medicine identification. Front. Plant Sci., 2016;7, 367.

Talens RP, Boomsma DI, Tobi EW, Kremer D, Jukema JW, Willemsen G, et al. Variation, patterns, and temporal stability of DNA methylation: Considerations for epigenetic epidemiology. FASEB J. 2010 Sep;24(9):3135-44.

Tammen SA, Friso S, Choi S. Epigenetics: The link between nature and nurture. Mol Aspects Med. 2013;34(4):753-64.

Terry MB, Delgado-Cruzata L, Vin-Raviv N, Wu HC, Santella RM. DNA methylation in white blood cells: Association with risk factors in epidemiologic studies. Epigenetics. 2011;6(7):828-37.

Tsukahara S, Kobayashi A, Kawabe A, Mathieu O, Miura A, Kakutani T. Bursts of retrotransposition reproduced in arabidopsis. Nature. 2009;461(7262):423-6.

Vanyushin B, Nemirovsky L, Klimenko V, Vasiliev V, Belozersky A. The 5-methylcytosine in DNA of rats. Gerontology. 1973;19(3):138-52.

Vidaki, A, and Kayser M. From forensic epigenetics to forensic epigenomics: broadening DNA investigative intelligence. Genome Biol., 2017;18(1): 238.

Vidaki, A., Daniel, B., & Court, D. S. Forensic DNA methylation profiling—potential opportunities and challenges. Forensic Sci. Int. Genet., 2013;7(5), 499-507.

Voet, D., Voet J. G. Biochemistry. John Wiley & Sons Inc. Fourth Edition. 2011;1246- 1247.

Vossen RH, Aten E, Roos A, & den Dunnen JT. High- Resolution Melting Analysis (HRMA)—More than just sequence variant screening. HuM. Mutat., 2009;30(6), 860-866.

Waddington, C.H. Waddington. Endeavour, 1942,1; pp. 18-20

Wahl S, Drong A, Lehne B, Loh M, Scott WR, Kunze S, Tsai PC et al. Epigenome-wide association study of body mass index, and the adverse outcomes of adiposity. Nature; 2017;541(7635): 81

Wan ES, Qiu W, Baccarelli A, Carey VJ, Bacherman H, Rennard SI, et al. Cigarette smoking behaviours and time since quitting are associated with differential DNA methylation across the human genome. Hum. Mol. Genet. 2012;21:3073–82.

Warnecke PM, Stirzaker C, Melki JR, Millar DS, Paul CL, Clark SJ. Detection and measurement of PCR bias in quantitative methylation analysis of bisulphite-treated DNA. Nucleic Acids Res. 1997 Nov 1;25(21):4422-6.

Wasserstrom A, Frumkin D, Davidson A, Shpitzen M, Herman Y, & Gafny R. Demonstration of DSI-semen—a novel DNA methylation-based forensic semen identification assay. Forensic Sci. Int. Genet., 2013; 7(1), 136-142.

Weber M, Hellmann I, Stadler MB, Ramos L, Paabo S, Rebhan M, & Schubeler D. Distribution, silencing potential and evolutionary impact of promoter DNA methylation in the human genome. Nature Genet., 2007;39(4), 457.

Website1:mmg-233-2014-genetics-genomics.fandom.com/wiki/Pyrosequencing

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 Biol. 2014;15(2):R24.

Wilson VL, Smith RA, Ma S, Cutler RG. Genomic 5-methyldeoxycytidine decreases with age. J. Biol Chem. 1987 Jul 25;262(21):9948-51.

Wittwer, CT, Reed, GH, Gundry CN, Vandersteen JG, & Pryor RJ. High-resolution genotyping by amplicon melting analysis using LCGreen. Clin. Chemistry, 2003;49(6), 853-860.

Wojdacz TK, & Dobrovic A. Methylation-sensitive high resolution melting (MS-HRM): a new approach for sensitive and high-throughput assessment of methylation. Nucleic Acids Res., 2007;35(6), e41.

Wojdacz TK. Methylation-sensitive high-resolution melting in the context of legislative requirements for validation of analytical procedures for diagnostic applications. Expert Rev. Mol. Diagn., 2012;12(1), 39-47.

Wu H, Wang Q, Chung WK, Andrulis IL, Daly MB, John EM, et al. Correlation of DNA methylation levels in blood and saliva DNA in young girls of the LEGACY girls study. Epigenetics. 2014;9(7):929-33.

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 Sci. Int. Genet.. 2015b;14:161-7.

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 Sci. Int. Genet., 2015a;17:173-9.

Zeilinger S, Kühnel B, Klopp N, Baurecht H, Kleinschmidt A, Gieger C, et al. Tobacco smoking leads to extensive genome-wide changes in DNA methylation. PloS one. 2013;8(5):e63812.

Chapter IV: Detection and evaluation of DNA methylation markers found at SCGN and KLF14 loci to estimate human age

This is the pre-peer reviewed version of the following article: Alghanim, H., Antunes, J., Silva, D. S. B. S., Alho, C. S., Balamurugan, K., & McCord, B. (2017). Detection and evaluation of DNA methylation markers found at SCGN and KLF14 loci to estimate human age. Forensic Science International: Genetics, 31, 81-88. which has been published in final form at https://www.fsigenetics.com/article/S1872-4973(17)30153-9/fulltext

1. Abstract

Recent developments in the analysis of epigenetic DNA methylation patterns have demonstrated that certain genetic loci show a linear correlation with chronological age. It is the goal of this study to identify a new set of epigenetic methylation markers for the forensic estimation of human age. A total number of 27 CpG sites at three genetic loci, SCGN, DLX5 and KLF14, were examined to evaluate the correlation of their methylation status with age. These sites were evaluated using 72 blood samples and 91 saliva samples collected from volunteers with ages ranging from 5 to 73 years. DNA was bisulfite modified followed by PCR amplification and pyrosequencing to determine the level of DNA methylation at each CpG site. In this study, certain CpG sites in SCGN and KLF14 loci showed methylation levels that were correlated with chronological age, however, the tested CpG sites in DLX5 did not show a correlation with age.

Using a 52-saliva sample training set, two age-predictor models were developed by means of a multivariate linear regression analysis for age prediction. The two models performed similarly with a single-locus model explaining 85% of the age variance at a mean absolute deviation of 5.8 years and a dual-locus model explaining 84% of the age variance with a mean absolute deviation of 6.2 years. In the validation set, the mean absolute deviation was measured to be 8.0 years and 7.1 years for the single- and dual-locus model, respectively. Another age predictor model was also developed using a 40-blood sample training set that accounted for 71% of the age variance. This model gave a mean absolute deviation of 6.6 years for the training set and 10.3 years for the validation set. The

results indicate that specific CpGs in SCGN and KLF14 can be used as potential epigenetic markers to estimate age using saliva and blood specimens. These epigenetic markers could provide important information in cases where the determination of a suspect's age is critical in developing investigative leads.

2. Highlights

- Single- and dual-locus age predictors for saliva were developed from CpG sites in *KLF14* and *SCGN* using a multivariate linear regression analysis.
- Single-locus model explained 85% of age variance with a MAD of 5.8 years and dual-locus model explained 84% with a MAD of 6.2 years in the training set.
- Single-locus model was efficient for younger subjects correctly predicting age of 78.9% of the samples with a MAD of 5.1 years in the validation set.
- The age predictor for blood was based on CpG sites in *SCGN* and *KLF14* and accounted for 71% of age variance with a MAD of 6.6 years in the training set.

3. Introduction

In order to identify individuals following the commission of a crime, DNA profiles generated from crime scene samples are compared either to the suspect's DNA or to existing local and national databases. If no match is reported, then the case remains unsolved. For such cases, where a suspect cannot be identified based on a DNA match with a database, the investigation depends on tools that can provide investigative information regarding the evidence found at the crime scene. For this reason, forensic scientists have been investigating ancestry and phenotypic traits available in short tandem repeats (STRs), insertion/deletion

polymorphisms (InDel), single nucleotide polymorphisms (SNPs), and epigenetic markers [1,2]. Investigative leads generated by these genetic markers can be used to direct police to a probable suspect list. Recently our laboratory has become interested in developing epigenetic markers to serve as new type of “DNA witness”. Epigenetic modifications have been examined in various diseases, especially cancer [3,4]. There have also been studies linking these changes with age-related illnesses such as Alz-heimer’s disease [5]. Interestingly epigenetic modifications can also be generally associated with chronological age. DNA methylation is the most widely studied form of the epigenetic modifications [3]. In eukaryotic DNA, 5-methylcystosine, which signifies the presence of a methyl group at the 5’ position of the cytosine nucleotide, commonly occurs in base sequences when the ‘C’ is followed by ‘G’ which are known as CpG sites [6]. The upstream regions of many genes contain CpG dinucleotides clusters known as CpG islands which have a role in gene regulation. In general, the presence of DNA methylation is believed to inhibit gene expression by affecting the chromatin structure [6,7]. While these locations are critical to cellular function, there are also epigenetic loci that are less stable, or no longer utilized. Such loci are more prone to changes in the methylation status over time due to environmental factors and spontaneous epigenetic modification [8]. Different studies have also demonstrated that DNA methylation can produce fixed changes in genomic information coded into the DNA sequence as a result of dynamic environmental factors [9,10]. It has also been demonstrated that DNA methylation is influenced by a variety of complex exogenous and endogenous factors including early life experiences, nutrition and

diets, aging, exposure to pollutants, smoking, ethnicity, and social environment [6-10].

DNA methylation is a useful and accessible epigenetic marker for quantitative measurements in human populations, thus there has been great interest in exploring the associations between DNA methylation with different diseases and phenotypic variations. Over time, specific CpG dinucleotides at certain locations can become hyper- or hypo-methylated [8,11]. The first study which related DNA methylation with aging investigated life stages in spawning humpback salmon demonstrating that global DNA methylation decreased significantly with age (during ontogenesis) [12]. Subsequently, various studies have reported similar decreases in DNA methylation with age in rats, mice, and humans [13-15]. More recently a number of studies have appeared which examined specific CpG sites that demonstrated linear correlations with chronological age [16-23]. Such information is useful in age-related research and can also have significant value in forensic identification of unknown individuals.

Interest in such applications has resulted in a number of genome wide association studies using large scale epigenetic arrays. Based on these studies, different research groups have reported finding a large number of CpG sites that show linear correlation with age [8,16-19]. Additional studies have examined the use of more specific subsets of age-related CpG sites ranging from 1-5 CpGs [20-23]. We have taken a similar approach by studying methylation sites identified by various studies [8,17,24-26] and summarized by Steegenga et al. [27]. From this report, we identified a set of potential age-related epigenetic loci and studied these

loci using quantitative pyrosequencing. These sites are located in the vicinity of three genes, *SCGN* (Secretagogin), *DLX5* (distal-less homeobox 5 gene), and *KLF14* (Kruppel-Like Factor 14). We next investigated CpG sites in these loci at two different types of body fluids with the potential to be found at crime scenes (blood and saliva). The main objective of this study was to develop epigenetic methylation markers to estimate the age of individuals from DNA samples found at crime scenes.

4. Materials and methods

4.1 Sample collection

Blood and saliva (buccal) samples (n=72 and 91 respectively) were collected from donors with ages ranging from 6 to 73 years for saliva and 5 to 72 for blood. Whole blood samples were collected onto a cotton swab. The saliva (buccal) samples were collected by rubbing a cotton swab along the inside of the cheek of volunteers. From the 72 people that donated blood samples, 58 volunteers also donated saliva samples. The remaining volunteers only donated one of the two body fluids. All biological samples were collected according to the protocol approved by the Institutional Review Board at Florida International University under IRB-13-0555 and IRB-16-0021; the University of Southern Mississippi protocol # 12010303; Pontifícia Universidade Científico do Rio Grande do Sul (CONEP #723.619/ CEP #845.747); and the General Headquarter of Dubai Police (approval letters #365259/11/33/3583 and #410126/11/33/3583). Prior to sample collection, all participants signed the informed consent forms.

4.2 DNA extraction and bisulfite conversion

DNA was extracted from blood and saliva samples using either the EZ1® DNA Investigator Kit on the BioRobot® EZ1 automated extraction workstation (Qiagen Inc., CA) or by organic extraction method using phenol-chloroform-isoamyl alcohol (Fisher Scientific, NJ) [28]. A total DNA volume of 50 μL was recovered and quantified using Alu-based real-time PCR method with a Rotor-Gene 6000 (Corbett Research, Sydney, Australia, now Qiagen Inc., CA) [29]. Two-hundred to five hundred nanograms of extracted DNA were bisulfite-modified using the EpiTect® Fast DNA Bisulfite Kit (Qiagen Inc., CA) to convert the unmethylated cytosines to uracil.

4.3 Assay design

In previous Genome Wide Association Studies (GWAS), *SCGN*, *DLX5*, and *KLF14* loci were reported to contain age-related CpG sites [8,17,24-26]. In these studies, however, only one or two CpG sites were tested per locus. Here, we examined these loci further by investigating a broad range of CpG sites in each genetic locus. Specific PCR primers were designed using PyroMark Assay Design 2.0 software (Qiagen Inc. CA) to amplify the bisulfite modified target region. The designed assays of *SCGN*, *DLX5*, and *KLF14* loci targeted between seven to ten CpG sites in each locus (Tables 2.1 and 2.2). One of the PCR primers was biotin labeled to produce biotinylated PCR amplicons needed for the pyrosequencing reaction.

Table 2.1: Assays design and primer sequences for each assay to evaluate CpG sites in three different genetic loci. Chr.: chromosome *: biotinylated primer Amp.: Amplicon

Locus	Chr./ Gene ID	Amp. Size	CpG sites analyzed (**bold**) (sequence to analyze)
KLF14	7/136259	137	TGG**Y**GTTTGGTAGTA GGTGTGATAGATTTT TTT**Y**GGGGG**Y**GTTTGA TT**Y**G**Y**GG**Y**GGGGG**Y** GGGGTTTGTTTTTAG GGTTTTTTTAG
DLX5	7/1749	190	**R**ATAAACC**R**CAAAC**R** AC**R**AAAAAAC**R**CCAA ACAAAACTAAACC**R**C CCAAAAATCC**R**AAAA TACCTCCAATC**R**CC**R** CCCC**R**AAAAAAAC
SCGN	6/10590	261	TT**Y**G**Y**GT**Y**GGTGTTT GGTTTT**Y**GT**Y**GTTAA TATTATGGATAGTTT T**Y**GGGAAT**Y**GATTTT GGGG**Y**GTTTGGA**Y**G T**Y**GTTGGTTTTTGGT AGGTTTGGTAG

4.4 PCR and pyrosequencing

PCR reactions were carried out in a singleplex fashion using the PyroMark® PCR kit (Qiagen Inc., CA) and utilizing the GeneAmp® PCR system 9700 (Applied Biosystems, Foster City, CA). The PCR reaction was modified to utilize half reaction volumes based on the total volume specified by the manufacturer's protocol. Amplifications were performed using 2.5μL (10-25 ng) of bisulfite modified DNA, 7.5μL ProMark PCR master mix, 1.5μL of CoralLoad concentrate, 1.5μL of primers and 2μL of distilled water in a total reaction volume of 15μL. The

pyrosequencing procedure was carried out using the Pyromark® Q24 pyrosequencer (Qiagen Inc., CA) according to the recommended manufacturer's instructions. After pyrosequencing, the percent methylation was calculated by Pyromark® Q24 software and the results were displayed as a pyrogram with the methylation values for each CpG site. The repeatability of the pyrosequencing method is 0.2-0.7% of methylation according to the PyroMark® Q24 validation oligo handbook which can be downloaded from the vendor's website at https://www.qiagen.com/.

4.5 Statistical analysis

The correlation between age and methylation status of the 27 tested CpG sites was examined across saliva and blood samples using linear regression analysis. Initially a bivariate correlation was performed for each CpG site. In addition, multivariate regression analysis was executed for all tested CpG sites as well as for different combinations of CpG sites located at *SCGN* and *KLF14* loci. The multivariate regression analysis was based on multiple CpGs from single and dual genetic loci. The accuracy of age prediction associated with particular CpG sites and the developed multivariate regression model were assessed using an adjusted R^2 value. This adjusted R^2 parameter calculates the proportion of age variance that can be accounted for using certain CpG markers (CpG predictor sites). All the analyses were performed using SPSS statistics software ver. 23.

5. Results

5.1 Developing the age prediction model for saliva

In order to develop a prediction model for age in saliva, the 91 collected saliva samples were divided into a 52 sample training set used for the initial building of the model and a 39 sample validation set used to evaluate the prediction performance. DNA methylation data were obtained from 52 training samples collected from individuals aged between 9 and 73 years. The methylation status for a total of 27 CpG sites were examined in the three genetic loci. In particular, 10 CpGs from *SCGN*, 10 CpGs from *DLX5*, and 7 CpGs from *KLF14* were examined (Table 2.2). The percent DNA methylation values obtained for the saliva samples were used to evaluate the correlation of the 27 CpG sites with chronological age. Bivariate correlation was used to test the relationship between chronological age and each CpG site. The results indicated that 9 CpGs from *SCGN* and 6 CpGs from *KLF14* produced a methylation status that suggested significant association with age (correlation coefficient $r > 0.5$). The strongest correlation was seen at position CpG1 in *KLF14* ($r = 0.876$) explaining 77% of the age variance in the training set ($R^2 = 0.768$). The second best was at position CpG3 in *KLF14* ($r = 0.862$), followed by CpG2 in *KLF14* ($r = 0.839$) and then CpG3 in *SCGN* ($r = 0.838$). The *DLX5* genetic locus failed to show correlation with age in the initial trial testing period and thus no further analysis was performed at this locus.

Age-predictor models were developed using a multivariate linear regression analysis based on the CpG sites tested in this study. In particular, a simple, cost effective, and relatively accurate single-locus age predictor can be obtained by utilizing CpG1 and CpG2 from *KLF14* locus (adjusted R^2= 0.851) with standard error of 7.2 (Table 2.3). The mean absolute deviation (MAD) calculated for the single-locus model in the training set was 5.8 years. In addition, we also developed a multivariate regression model combining CpGs from the two genetic loci. By doing so, we can compare results and see which model presents a more accurate age predictor. We chose the CpG site that presented the strongest correlation with age from each of the two genetic loci. Thus, the simplest and the most informative dual-locus assay was obtained by including CpG1 from *KLF14* and CpG3 from *SCGN*. This dual-locus age prediction model can explain 84% of the age variance (adjusted R^2=0.840) with a standard error of 7.5 years and a MAD of 6.2 years (Table 2.3).

Table 2.2: Information about the 27 CpG sites evaluated as age methylation markers in this study.

Locus	Chromosome location (GRCh37)	CpG position number	Illumina ID
KLF14	Chr7:130418281	CpG1	
	Chr7:130418311	CpG2	
	Chr7:130418316	CpG3	cg04528819
	Chr7:130418325	CpG4	cg20426994
	Chr7:130418327	CpG5	
	Chr7:130418330	CpG6	
	Chr7:130418336	CpG7	
SCGN	Ch6:25652601	CpG1	
	Ch6:25652603	CpG2	cg06493994
	Ch6:25652606	CpG3	
	Ch6:25652620	CpG4	
	Ch6:25652623	CpG5	
	Ch6:25652645	CpG6	
	Ch6:25652652	CpG7	
	Ch6:25652663	CpG8	
	Ch6:25652671	CpG9	
	Ch6:25652674	CpG10	
DLX5	Chr7:96650517	CpG1	
	Chr7:96650509	CpG2	cg00503840
	Chr7:96650503	CpG3	
	Chr7:96650500	CpG4	
	Chr7:96650492	CpG5	
	Chr7:96650474	CpG6	
	Chr7:96650462	CpG7	
	Chr7:96650446	CpG8	
	Chr7:96650443	CpG9	
	Chr7:96650438	CpG10	

Table 2.3: Single- and dual-locus multivariate saliva age prediction model for the training and validation sets. Adj.: Adjusted Stand.: Standard

Locus	CpG sites involved	Training set				Validation set			
		R	Adj. R^2	Stand. error	MAD	*R*	Adj. R^2	Stand. error	MAD
KLF14	CpG1 CpG2	0.926	0.851	7.2	5.8	0.900	0.810	7.3	8.0
KLF14 +*SCGN*	CpG1-*KLF14* CpG3-*SCGN*	0.920	0.840	7.5	6.2	0.861	0.741	8.5	7.1

The developed age prediction models based on the training set of saliva samples are presented as follows:

1. Single-locus model: Estimated age (in years) = - 24.884 + (1.703 * CpG1 from *KLF14*) + (1.963 * CpG2 from *KLF14*)

2. Dual-locus model: Estimated age (in years) = - 14.969+ (1.710* CpG1 from *KLF14*) + (2.064 * CpG3 from *SCGN*)

5.2 Developing the age prediction model for blood

Using the same procedure that was applied to the saliva samples, the 72 collected blood samples were divided into a 40 sample training set and a 32 sample validation set in order to develop an age-predictor model for blood. The methylation profiles for the two genetic loci were tested using the 40 blood samples of the training set which were collected from donors with ages 5 to 72 years old. By evaluating the methylation levels in blood, the best age related methylation sites were CpG1 in *SCGN* which showed correlation with an increase in age (*r*= 0.832) followed by CpG2 (r= 0.643) and CpG3 (*r*= 0.534) in *KLF14*. Based on these three most informative CpG sites, a multivariate linear regression model was developed

to predict the age of the donors from blood samples. This dual-locus model could explain 71% of the age variance (adjusted R^2= 0.708) with a standard error of 9.4 and a MAD of 6.6 years (Table 2.4). All methylation sites tested during the initial trial testing period in blood samples for *DLX5* locus also failed to give a significant correlation with age (r < 0.5).

Table 2.4: Dual-locus multivariate blood age prediction model for the training and validation sets. Adj.: Adjusted Stand.: Standard

Locus	CpG sites involved	Training set				Validation set			
		R	Adj. R^2	Stand. error	MAD	*R*	Adj. R^2	Stand. error	MAD
KLF14 + *SCGN*	CpG2-*KLF14*	0.855	0.708	9.4	6.6	0.912	0.813	7.9	10.3
	CpG3-*KLF14*								
	CpG1-*SCGN*								

5.3 Testing the accuracy of the age prediction model

Using a validation set of 39 independent saliva samples, the performance of the prediction models obtained from the training set was evaluated. Based on the multivariate linear regression model, the standard error of estimate in the validation set for the single- and the dual-locus model was 7.3 and 8.5 years, respectively. The MAD from chronological age calculated for the validation set using the single- and dual-locus model was 8.0 and 7.1 years, respectively (Table 2.3). To further validate the single- and dual-locus models, ages were predicted using the methylation values for the saliva samples in the validation set. The predicted ages were compared with chronological ages and considered correct when the

predicted age was within ±8 years of the chronological age. From the 39 validation samples, the single-locus formula gave a total number of 64.1% correct predictions whereas the dual-locus formula predicted age correctly for 66.7% of the samples.

It was also observed that age prediction is more accurate for younger subjects in the combined category 1 and 2 (≤40 years) when compared to older subjects in the combined category 3 and 4 (>40 years) (Table 2.5). In the single-locus model correct predictions of age occurred in 78.9 % of the samples for younger subjects (≤40 years) and only in 50.0% of the samples for the older subjects (>40 years) (Table 2.5). For the dual-locus model a similar situation occurred with 78.9% of the samples from younger subjects providing correct age predictions, whereas only 55.0% of the samples were correctly predicted for older subjects. MAD increased with age, showing values in the single-locus model of 5.1 years for younger subjects and 10.8 for the older subjects. For the dual-locus model the MAD values were equal to 5.7 and 8.5 for the younger and older subjects, respectively (Table 2.5). The higher accuracy for predicting age in the younger subjects compared to the older subjects was also demonstrated in a plot of chronological age *versus* predicted age of the entire data set for the single- and the dual-locus model as shown in Figures 2.1 and 2.2, respectively. To test the accuracy of the age predictor model for blood, 32 independent blood samples from the validation set were used on a multivariate linear regression analysis and the standard error of estimate was calculated to be 7.9 years with a MAD of 10.3 years.

Table 2.5: MAD and % of correct prediction in four age categories. Combining category 1 and 2 represent younger individuals ≤ 40 years old and combining category 3 and 4 represent older individuals > 40 years of age. Correct prediction was assumed when the predicted age was within ±8 years of the chronological age.

Age category	Single-locus Model		Dual-locus Model	
	MAD	% Correct prediction	MAD	% Correct prediction
1 (6-22 years)	4.4	100.0	5.0	83.3
2 (23-40 years)	5.6	69.2	5.8	76.9
3 (41-54 years)	12.2	45.5	8.4	54.9
4 (55-67 years)	9.0	55.6	8.6	55.6
Category 1 & 2	5.1	78.9	5.7	78.9
Category 3 & 4	10.8	50.0	8.5	55.0
Overall	8.0	64.1	7.1	66.7

Figure 2.1: Chronological age *versus* predicted age of the entire data set of the 91 saliva samples using the single-locus prediction model (CpG1 and CpG2 from *KLF14*)

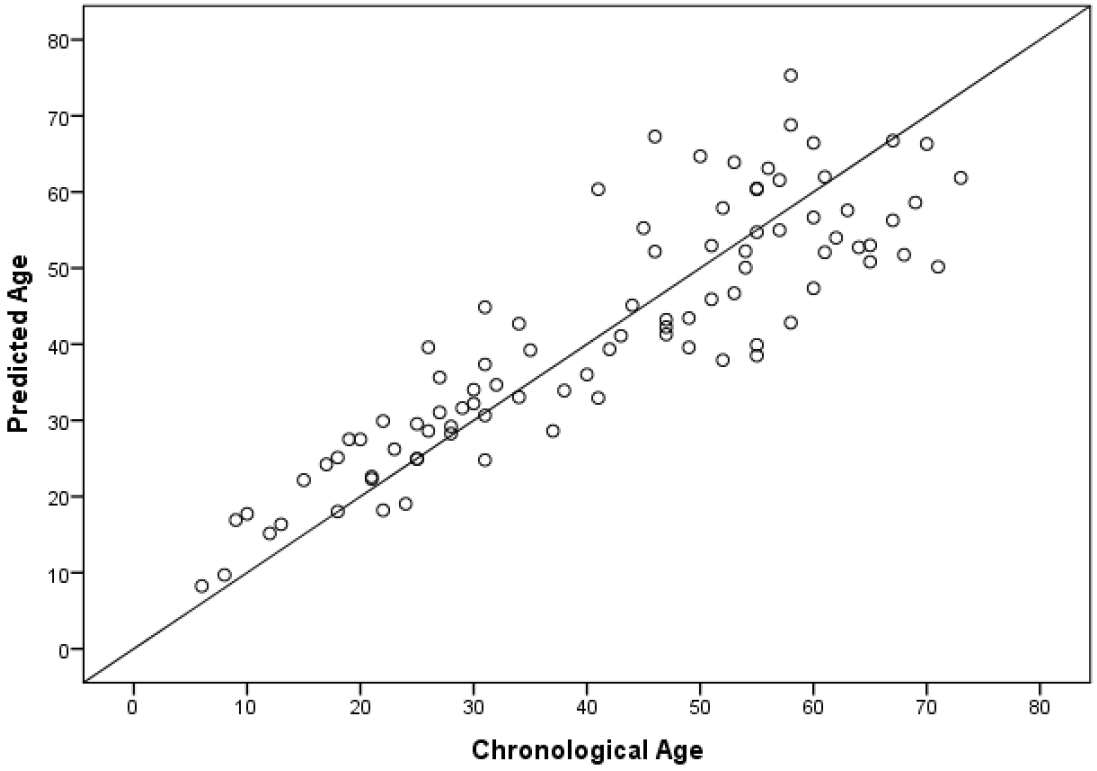
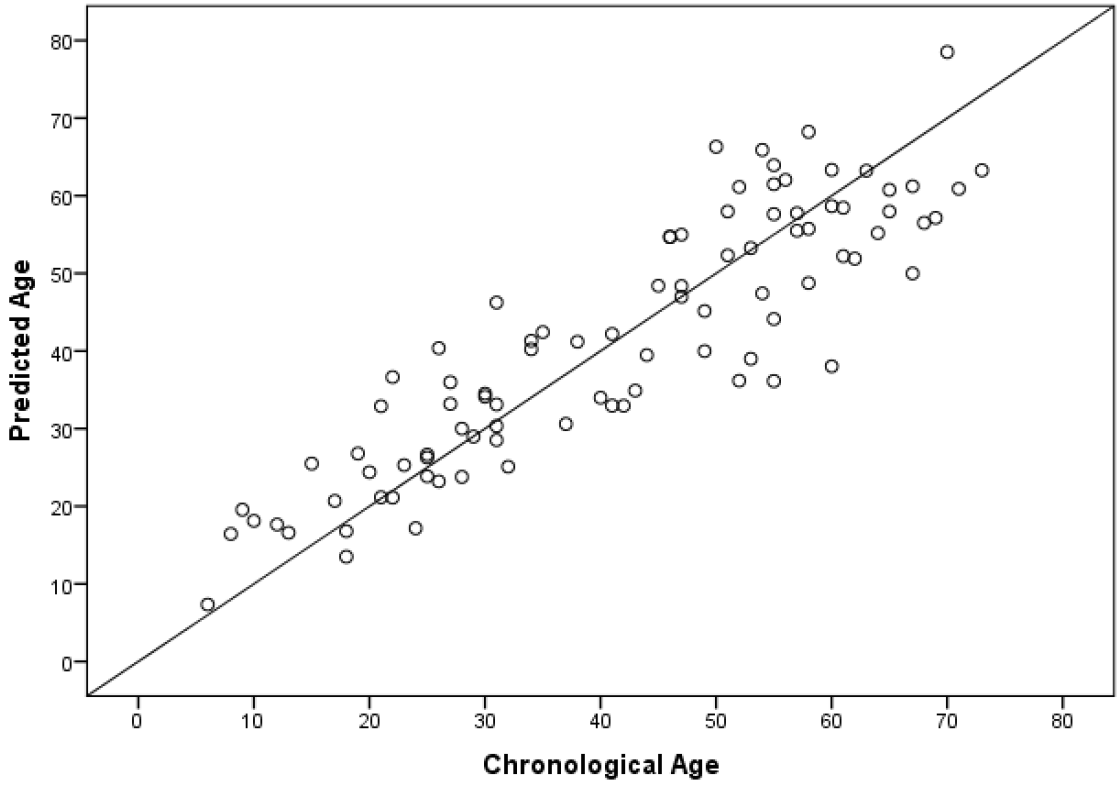


Figure 2.2. Chronological age *versus* predicted age of the entire data set of the 91 saliva samples using the dual-locus prediction model (CpG1 from *KLF14* and CpG3 from *SCGN*)



6. Discussion

To date, DNA methylation is the most widely studied epigenetic process in aging studies. Research in the field of aging utilizing DNA methylation started with a study performed by Berdyshev et al. in 1967 [12]. However, the field has experienced rapid growth during the last 8 years due to advancements in technology that allowed genome-wide analysis of the DNA methylome. The global status of DNA methylation usually decreases with aging in human tissues [30], however certain genetic loci linked with specific genes show an increase in DNA

methylation status upon aging [31]. In forensics, age estimation could be a very useful tool in determining the potential identity of a suspect by narrowing the range of those who could have been the source of the recovered samples. Different genetic markers in the human genome have been examined to predict age, including accumulation of mitochondrial DNA deletions, telomere shrinkage, and, most recently, epigenetic DNA methylation [2,32]. Overall, DNA methylation has been shown to be more suitable for age estimation due to its stability, especially in degraded biological samples. This could be a particular advantage when examining biological samples found at crime scenes [33].

The goal in this study was to identify novel methylation sites associated with age that have not been previously examined using pyrosequencing. In previous studies, we have demonstrated that single CpG sites identified by genome wide array studies commonly contain a multiplicity of nearby CpG sites which can be identified once their level of methylation is determined using pyrosequencing [11,34]. Therefore, a survey was performed in order to select individual epigenetic loci based on loci's potential to predict age and the number of times a particular CpG site was reported across different GWAS. These results as well as published data on age related methylation compiled from different GWAS by Steegenga et al. [27], indicated that three age-related genes, *SCGN* (Secretagogin), *DLX5* (distal-less homeobox 5 gene), and *KLF14* (Kruppel-Like Factor 14) might prove useful for pyrosequencing based age prediction. To predict age using DNA methylation, it is important to understand the effect of different tissue sources [22], therefore we tested both blood and saliva samples as these are commonly

encountered in forensic casework. The gene *SCGN* on chromosome 6 encodes for secretagogin protein, a secreted calcium-binding protein found in the cytoplasm. The protein is believed to be involved in potassium chloride stimulated calcium flux as well as cell proliferation [35]. The gene *DLX5*, on chromosome 7 (at 7q22) [36], encodes a member of homeobox transcription factor gene family. The *DLX* family plays a role in appendage development. The encoded protein has an important role in bone development and fracture healing [37]. Finally, the *KLF14* gene, on chromosome 7, is a member of the Kruppel-like factor family of transcription factors that tend to regulate the transcription of various genes. Studies have shown that *KLF14* seems to be a master regulator of gene expression in adipose tissues [38] and appears to be associated with type 2 diabetes [39]. It is of course important to note that locations which correlate with chronological age are unlikely to be involved in active transcriptional events as the gradual buildup or loss of methylation status is most likely the result of a slow stochastic drift [8], and is less likely associated with biological function.

A correlation of the methylation patterns with chronological age had been previously detected at the cg06493994 in *SCGN*, cg00503840 in *DLX5*, and cg04528819 in *KLF14* using GWAS [17,24-26]. This previous work on the three CpG sites was based on array studies which require large amounts of DNA and laborious bioinformatic analysis. Thus, we decided to investigate specific CpG sites in these loci using quantitative pyrosequencing. The use of bisulfite DNA conversion followed by pyrosequencing permits the identification and quantification of methylation for clusters of CpG sites associated with a single

epigenetic locus. Pyrosequencing permits the relative methylation at each CpG site to be measured at high accuracy [40]. Furthermore, pyrosequencing-based techniques utilize minimal DNA starting material permitting downstream STR testing as well.

KLF14 is an important locus that is utilized in various age predictor models for various cell types including saliva [20,41]. In previous work by Zbiec-Piekarska et al. [20] and Hong et al. [41], researchers evaluated several CpGs in the *KLF14* gene located around the CpG at cg14361627 probe site using pyrosequencing. In this report, we showed the results for a new set of CpGs located near the cg04528819 probe site of *KLF14*. In the study by Eipel et al. [22], an age predictor was developed for buccal epithelial cells based on three CpG sites using three different genetic loci with a MAD of 4.3 years in the training set and 7.03 years in the validation set using pyrosequencing technique. This study reported that the composition of the buccal epithelial cells could also be determined by measuring the DNA methylation at two additional CpG sites. The study concluded that combining the cell type specific and the age-associated CpGs into one model can improve the age estimation in saliva [22]. In addition, Hong et al. [41] reported a multiplex age prediction model for saliva that is composed of seven CpGs located at seven different genetic loci exhibiting a MAD of 3.1 years in the training set and 3.2 years in the validation set using SNaPshot. In our study, we wanted to develop age predictors that require only one or two PCR reactions for quick age estimation using pyrosequencing. To develop such age prediction models, the methylation values obtained from the saliva samples of the training set were first evaluated by

performing a bivariate correlation analysis at each methylated site. CpG1 in *KLF14* showed the strongest correlation with age followed by CpG3 in *KLF14*, CpG2 in *KLF14* and then CpG3 in *SCGN*. These results demonstrate the importance of assessing the influence of each CpG site in a multiple regression model. Based on the multivariate linear regression method, the simplest and the most informative single-locus age predictor was developed using CpG1 and CpG2 from *KLF14.* The second age predictor model for saliva was developed based on two separate loci; hence the name "dual-locus" in which CpG1 from *KLF14* is combined with CpG3 from *SCGN*. The methylation level of CpG sites used in the single-locus and dual-locus prediction models increased proportionally with increase in age. The single-locus and the dual-locus models performed similarly in the saliva training set giving adjusted R^2 of 0.851 with a standard error of 7.2 years and adjusted R^2 of 0.840 with a standard error of 7.5 years, respectively (Table 2.3). The accuracy of both age-predictor models was tested using an independent validation set of 39 saliva samples. In the validation set, the standard error of estimate and the mean absolute deviation were estimated as 7.3 and 8.0 years, respectively, for the single-locus model and 8.5 and 7.1 years, respectively, for the dual-locus model (Table 2.3). In addition, when the methylation values of the samples in the validation set were used with the age prediction formula, we observed that the dual-locus gave a total number of 66.7% correct predictions within ±8 years *versus* 64.1% correct predictions using the single-locus formula indicating that both models are comparable in term of accuracy. In summary, the single-locus age predictor provides similar accuracy and is faster to perform since it only requires

one PCR amplification and a single pyrosequencing reaction. The use of single-locus for age estimation is an efficient technique which allow the user to preserves sample, an important factor in forensic analysis which is often sample limited.

When studying the 40 blood samples in the training set, all CpG sites in *KLF14* and *SCGN* presented a lower correlation with age when compared to the results obtained from the saliva samples. The multivariate age-predictor model for blood was developed using the three most informative CpG sites at *KLF14* and *SCGN* detected in the training set. The combination of these three CpGs provided an adjusted R^2 of 0.708 and a standard error of 9.4 years (Table 2.4). To test the accuracy of the age predictor model for blood, the methylation values for the validation set of 32 blood samples were used to perform multivariable linear regression analysis, which had a lower standard error (7.9 years) when compared to the 40 blood samples of the training set. None of the tested CpG sites in *DLX5* showed a linear association with aging in saliva or blood samples.

Another interesting observation was that the age of the younger individuals (≤40 years) can be better predicted than the age of the older subjects (>40 years). For example, the single-locus age predictor had high accuracy, predicting the correct age (±8 years) of 78.9% of the younger subjects with a MAD equal to 5.1 years compared to a total number of correct predictions of 50.0% in the older subjects with a MAD of 10.8 years in the validation set (Table 2.5). This trend can be clearly observed when plotting the chronological age *versus* predicted age for the single-locus model for saliva in the entire data set as shown in Figure 2.1. These results agree with the previous studies by Horvath et al. [18], Zbiec-Piekarska et al. [20],

and Park et al. [42]. Overall, when evaluating the accuracy of the developed age-predictors models for saliva, the high number of correct predictions (78.9%) observed in the younger subjects confirms the accuracy and reliability of both saliva models to estimate age for individuals less or equal to 40 years old. In addition, we decided to test the difference of the DNA methylation values between saliva and blood samples originating from the same donor for the three most informative CpGs at each locus (Supplementary Figures 2.S1-2.S6). The results indicate a consistent trend in which the blood CpG values are higher than saliva with the exception observed for CpG2 and CpG3 at *KLF14* in which the methylation values were very similar. As a result, when testing the saliva age predictors in blood samples from the validation set, more than 96% of the samples were biased toward giving a higher age prediction value compared to the actual age for the single- and the dual-locus models. However, when testing the blood age predictor in saliva samples, no bias in the predicted age was detected. Therefore, the determination of body fluid type may be a prerequisite for an accurate estimate of age using the developed age prediction models.

In humans, age-related changes in DNA methylation status have been reported in blood [8], saliva [19], brain [43], and other tissues [44,45]. One of the challenges facing DNA methylation, particularly when used in forensic casework, is the reproducibility of the results against different cell types. The available literature suggests that changes in DNA methylation are tissue specific [22, 46]. Koch and Wagner et al. [44] believed that it may be difficult to find universal DNA methylation markers which can be used to predict age using any body fluid, however certain

markers like *ELOVL2* have been shown to be relatively stable age predictors for certain cell types [8,23,47]. In this study, we also observed a difference in DNA methylation patterns between saliva and blood from the same donors. Thus, further studies are required to test cross-correlation between multiple cell types in order to identify a universal set of CpG sites for age prediction. Fortunately, the same epigenetic techniques used to determine age can also assist in identifying the body fluid type [34]. Thus, we envision a future multiplex assay that determines the type of body fluid as well as the suspect's predicted age.

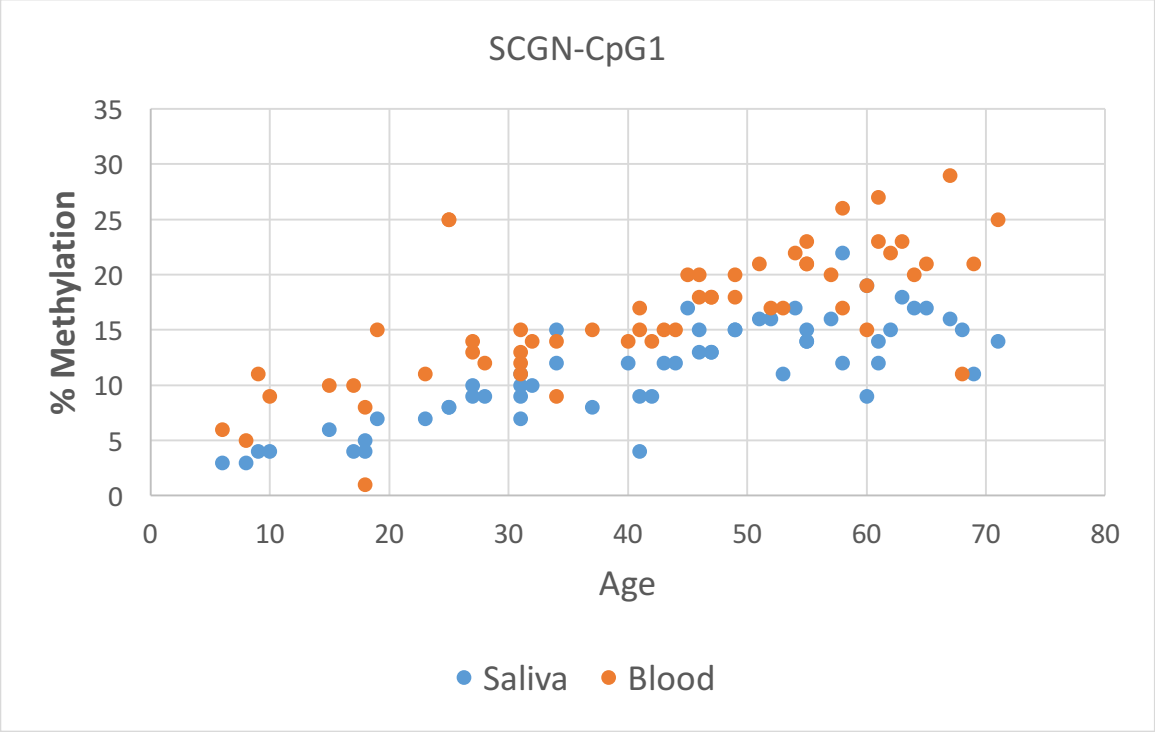
7. Conclusion

In this study, the main goal was to identify and evaluate a set of CpG sites to serve as novel and quick DNA methylation predictors of chronological age. A total of 27 CpG sites located in three different genetic loci, *SCGN*, *DLX5* and *KLF14*, were investigated using quantitative pyrosequencing. We have developed single- and dual-locus models to determine age based on saliva samples. The single locus age-predictor model for saliva sample is more cost-effective than the dual locus age-predictor model, especially for forensic routine. This single-locus model could be very useful to estimate age especially for younger subjects (≤40 years) predicting age (±8 years) with accuracy of 78.9% and a MAD of 5.1 years in the validation set. However, the developed age-prediction model used for blood cells is less informative having an adjusted R^2= 0.708 with a MAD of 6.6 years in the training set and 10.3 years in the validation set. Overall, we find that specific CpG

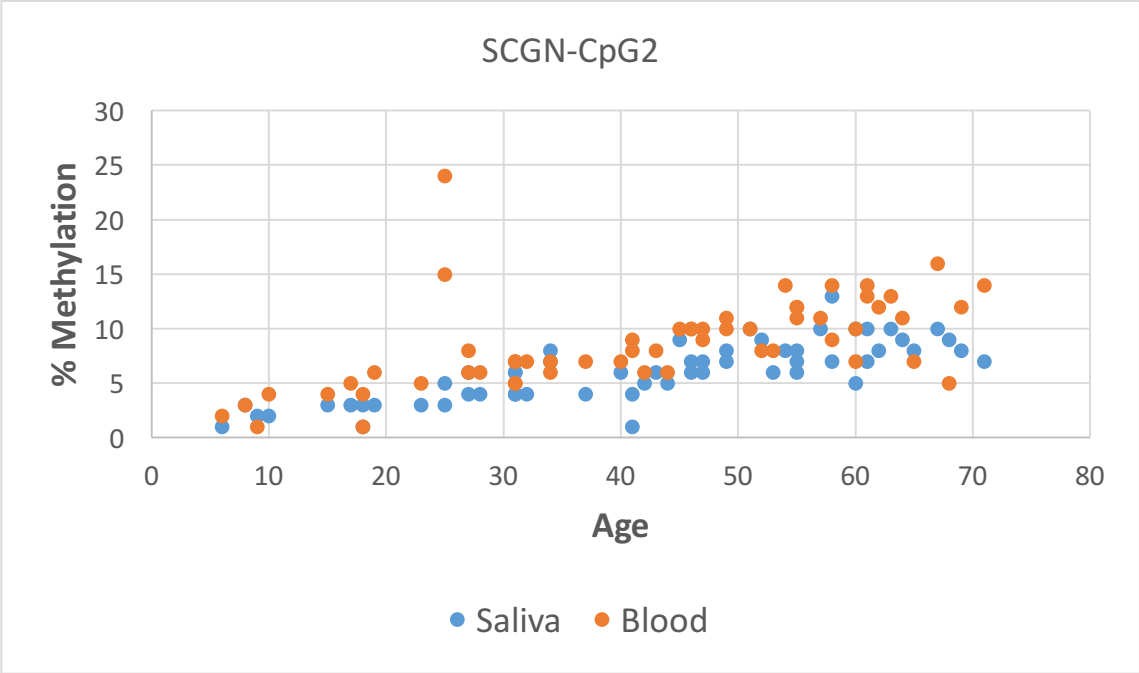
sites within *KLF14* and *SCGN* can serve as good age predictors. The results demonstrate the utility of pyrosequencing in the determination of age related methylation sites, and indicate that several of the investigated loci should prove to be useful for forensic evaluation of the age of unknown suspects recovered from body fluids left behind at crime scenes.

8. Supplementary

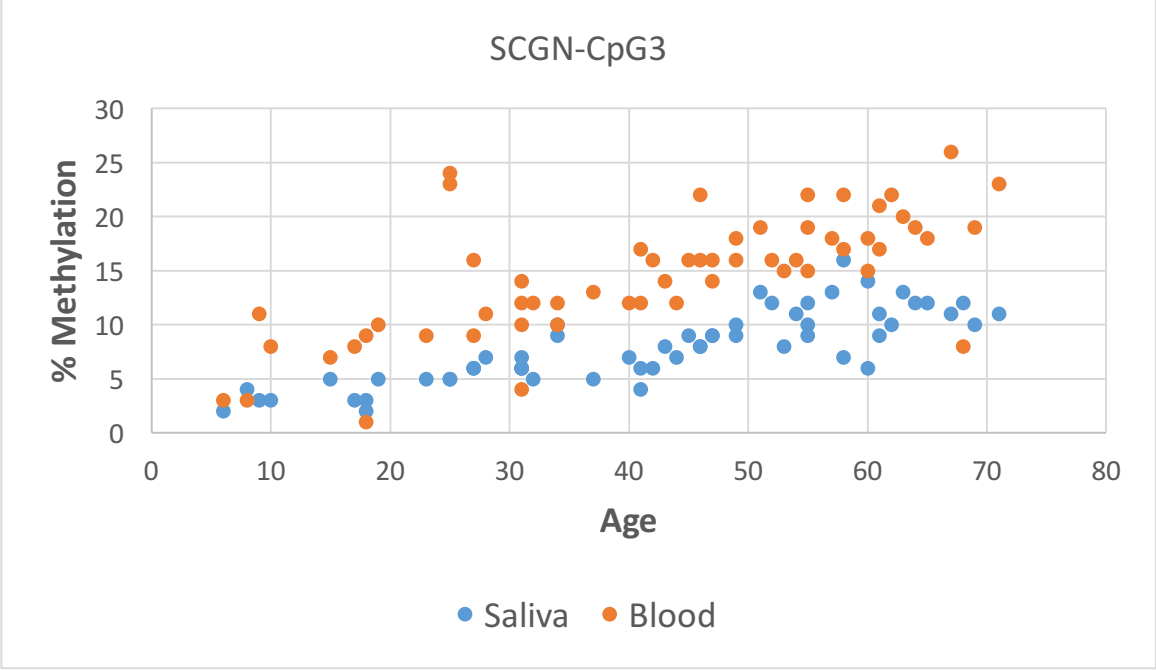
Supplementary Figure 2.S1: The graph shows the methylation percent for saliva and blood samples collected from the same donors at CpG1 in *SCGN*.



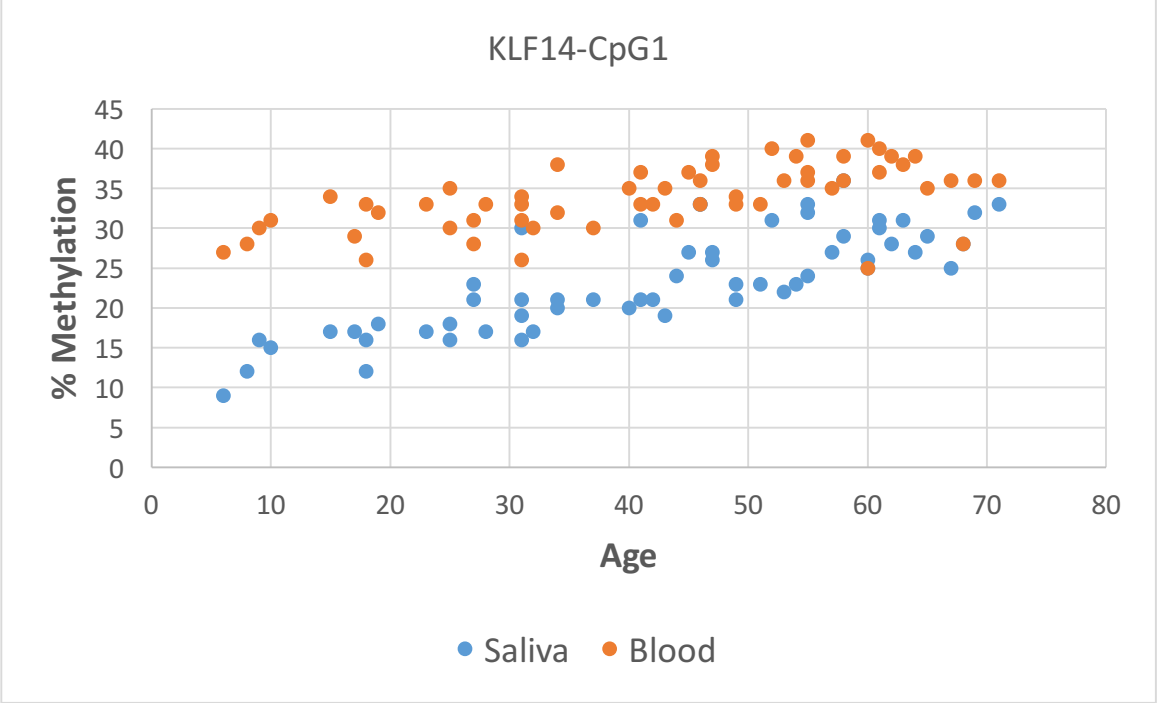
Supplementary Figure 2.S2: The graph shows the methylation percent for saliva and blood samples collected from the same donors at CpG2 in SCGN.



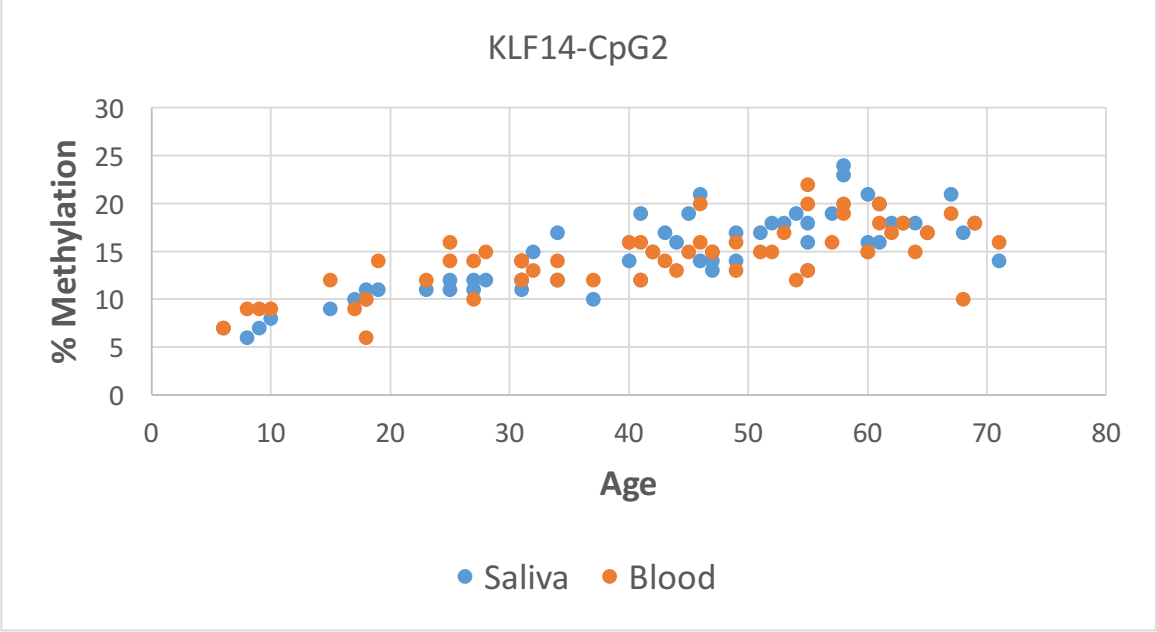
Supplementary Figure 2.S3: The graph shows the methylation percent for saliva and blood samples collected from the same donors CpG3 in SCGN.



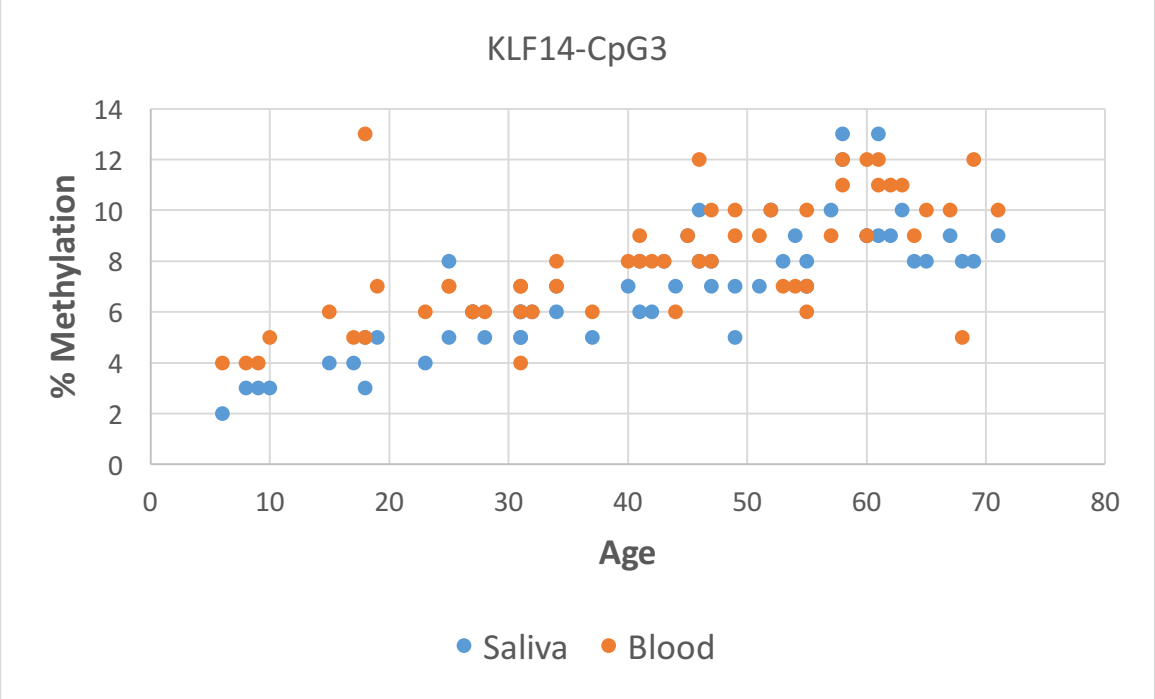
Supplementary Figure 2.S4: The graph shows the methylation percent for saliva and blood samples collected from the same donors CpG1 in KLF14.



Supplementary Figure 2.S5: The graph shows the methylation percent for saliva and blood samples collected from the same donors CpG2 in KLF14.



Supplementary Figure 2.S6: The graph shows the methylation percent for saliva and blood samples collected from the same donors CpG3 in *KLF14*.



9. References

[1] M. Kayser. Forensic DNA phenotyping: predicting human appearance from crime scene material for investigative purposes, Forensic Sci. Int. Genet., 18 (2015), pp. 33-48

[2] M. Kayser, P. de Knijff, Improving human forensics through advances in genetics, genomics and molecular biology, Nat. Rev. Genet., 12 (2011), pp. 179-192

[3] V.K. Rakyan, T.A. Down, D.J. Balding, S. Beck, Epigenome-wide association studies for common human diseases. Nat. Rev. Genet., 12 (2011), pp. 529-541

[4] J. Madrigano, A.A. Baccarelli, M.A. Mittleman, D. Sparrow, P.S. Vokonas, L. Tarantini, et al. Aging and epigenetics: longitudinal changes in gene-specific DNA methylation, Epigenetics, 7 (2012), pp. 63-70

[5] S. Wang, B. Oelze, A. Schumacher. Age-specific epigenetic drift in late-onset Alzheimer's disease. PLoS One, 3 (2008), p. e2698

[6] A. Vidaki, B. Daniel, D.S. CourtForensic DNA methylation profiling-Potential opportunities and challenges. Forensic Sci. Int. Genet., 7 (2013), pp. 499-507

[7] F. Kader, M. GhaiDNA methylation and application in forensic sciences. Forensic Sci. Int., 249 (2015), pp. 255-265

[8] G. Hannum, J. Guinney, L. Zhao, L. Zhang, G. Hughes, S. Sadda, et al. Genome-wide methylation profiles reveal quantitative views of human aging rates. Mol. Cell., 49 (2013), pp. 359-367

[9] H. Heyn, S. Moran, I. Hernando-Herraez, S. Sayols, A. Gomez, J. Sandoval, et al. DNA methylation contributes to natural human variation. Genome Res., 23 (2013), pp. 1363-1372

[10] S.A. Tammen, S. Friso, S. Choi. Epigenetics: the link between nature and nurture. Mol. Aspects Med., 34 (2013), pp. 753-764

[11] D.S.B.S Silva, J. Antunes, K. Balamurugan, G. Duncan, C. Sampaio Alho, B. McCord. Evaluation of DNA methylation markers and their potential to predict human aging. Electrophoresis, 36 (2015), pp. 1775-1780

[12] G.D. Berdyshev, G.K. Korotaev, G.V. Boiarskikh, B.F. Vaniushin. Nucleotide composition of DNA and RNA from somatic tissues of humpback and its changes during spawning. Biokhimiia, 32 (1967), pp. 988-993

[13] H.T. Bjornsson, M.I. Sigurdsson, M.D. Fallin, R.A. Irizarry, T. Aspelund, H. Cui, et al. Intra-individual change over time in DNA methylation with familial clustering. JAMA, 299 (2008), pp. 2877-2883

[14] B. Vanyushin, L. Nemirovsky, V. Klimenko, V. Vasiliev, A. Belozersky. The 5-methylcytosine in DNA of rats. Gerontology, 19 (1973), pp. 138-152

[15] V.L. Wilson, R.A. Smith, S. Ma, R.G. Cutler. Genomic 5-methyldeoxycytidine decreases with age. J. Biol. Chem., 262 (1987), pp. 9948-9951

[16] H. Heyn, N. Li, H.J. Ferreira, S. Moran, D.G. Pisano, A. Gomez, et al. Distinct DNA methylomes of newborns and centenarians. Proc. Natl. Acad. Sci. U. S. A., 109 (2012), pp. 10522-10527

[17] A.E. Teschendorff, U. Menon, A. Gentry-Maharaj, S.J. Ramus, D.J. Weisenberger, H. Shen, et al. Age-dependent DNA methylation of genes that are suppressed in stem cells is a hallmark of cancer. Genome Res., 20 (2010), pp. 440-446

[18] S. Horvath. DNA methylation age of human tissues and cell types. Genome Biol., 14 (2013), p. 1

[19] S. Bocklandt, W. Lin, M.E. Sehl, F.J. Sánchez, J.S. Sinsheimer, S. Horvath, et al. Epigenetic predictor of age. PLoS One, 6 (2011), p. e14821

[20] R. Zbieć-Piekarska, M. Spólnicka, T. Kupiec, A. Parys-Proszek, Ż. Makowska, A. Pałeczka, et al. Development of a forensically useful age prediction method based on DNA methylation analysis. Forensic Sci. Int. Genet., 17 (2015), pp.173-179

[21] C.I. Weidner, Q. Lin, C.M. Koch, L. Eisele, F. Beier, P. Ziegler, et al. Aging of blood can be tracked by DNA methylation changes at just three CpG sites. Genome Biol., 15 (2014), p. 1

[22] M. Eipel, F. Mayer, T. Arent, M.R. Ferreira, C. Birkhofer, U. Gerstenmaier, et al. Epigenetic age predictions based on buccal swabs are more precise in combination with cell type-specific DNA methylation signatures. Aging (Albany NY)., 8 (2016), pp. 1034-1048

[23] R. Zbieć-Piekarska, M. Spólnicka, T. Kupiec, Ż. Makowska, A. Spas, A. Parys-Proszek, et al. Examination of DNA methylation status of the ELOVL2 marker may be useful for human age prediction in forensic science. Forensic Sci. Int. Genet., 14 (2015), pp. 161-167

[24] J.T. Bell, P. Tsai, T. Yang, R. Pidsley, J. Nisbet, D. Glass, et al. Epigenome-wide scans identify differentially methylated regions for age and age-related phenotypes in a healthy ageing population. PLoS Genet., 8 (2012), p. e1002629

[25] V.K. Rakyan, T.A. Down, N.P. Thorne, P. Flicek, E. Kulesha, S. Graf, et al. An integrated resource for genome-wide identification and analysis of human tissue-specific differentially methylated regions (tDMRs). Genome Res., 18 (2008), pp. 1518-1529

[26] Z. Xu, J.A. Taylor. Genome-wide age-related DNA methylation changes in blood and other tissues relate to histone modification, expression and cancer. Carcinogenesis, 35 (2014), pp. 356-364

[27] W.T. Steegenga, M.V. Boekschoten, C. Lute, G.J. Hooiveld, P.J. de Groot, T.J. Morris, et al. Genome-wide age-related changes in DNA methylation and gene expression in human PBMCs Age, 36 (2014), pp. 1523-1540

[28] B.W. Koons, C.A. Sobieralski, C.T. Comey, E.S. Baechtel, J.B. Smerick, K.W. Presley, et al. DNA extraction strategies for amplified fragment length polymorphism analysis. J. Forensic Sci., 39 (1994), pp. 1254-1269

[29] J.A. Nicklas, E. Buel. Development of an Alu-based, real-time PCR method for quantitation of human DNA in forensic samples. J. Forensic Sci., 48 (2003), pp. 936-944

[30] M.A. Casillas, N. Lopatina, L.G. Andrews, T.O. Tollefsbol. Transcriptional control of the DNA methyltransferases is altered in aging and neoplastically-transformed human fibroblasts. Mol. Cell. Biochem., 252 (2003), pp. 33-43

[31] C. Murgatroyd, Y. Wu, Y. Bockmuhl, D. Spengler. The Janus face of DNA methylation in aging. Aging (Albany NY)., 2 (2010), pp. 107-110

[32] S.H. Yi, L.C. Xu, K. Mei, R.Z. Yang, D.X. Huang. Isolation and identification of age-related DNA methylation markers for forensic age-prediction. Forensic Sci. Int. Genet., 11 (2014), pp. 117-125

[33] N. Vilahur, A.A. Baccarelli, M. Bustamante, S. Agramunt, H. Byun, M.F. Fernandez, et al. Storage Conditions and Stability of Global DNA Methylation in Placental Tissue. (2013)

[34] T. Madi, K. Balamurugan, R. Bombardi, G. Duncan, B. McCord. The determination of tissue-specific DNA methylation patterns in forensic biofluids using bisulfite modification and pyrosequencing. Electrophoresis, 33 (2012), pp. 1736-1745

[35] K. Skovhus, R. Bergholdt, C. Erichsen, T. Sparre, J. Nerup, A. Karlsen, et al. Identification and characterization of secretagogin promoter activity Scand. J. Immunol., 64 (2006), pp. 639-645

[36] A. Simeone, D. Acampora, M. Pannese, M. D'Esposito, A. Stornaiuolo, M. Gulisano, et al. Cloning and characterization of two members of the vertebrate Dlx gene family. Proc. Natl. Acad. Sci. U. S. A., 91 (1994), pp. 2250-2254

[37] R.F. Robledo, L. Rajan, X. Li, T. LufkinThe Dlx5 and Dlx6 homeobox genes are essential for craniofacial, axial, and appendicular skeletal development. Genes Dev., 16 (2002), pp. 1089-1101

[38] K.S. Small. Å.K. Hedman, E., Grundberg, A.C., Nica, G., Thorleifsson, A. Kong, et al., Corrigendum: identification of an imprinted master trans regulator at the KLF14 locus related to multiple metabolic phenotypes. Nat.Genet, 43 (2011) 1040–1040

[39] B.F. Voight, L.J. Scott, V. Steinthorsdottir, A.P. Morris, C. Dina, R.P. Welch, et al. Twelve type 2 diabetes susceptibility loci identified through large-scale association analysis. Nat. Genet., 42 (2010), pp. 579-589

[40] D.S. Silva, J. Antunes, K. Balamurugan, G. Duncan, C.S. Alho, B. McCord. Developmental validation studies of epigenetic DNA methylation markers for the detection of blood, semen and saliva samples. Forensic Sci. Int. Genet., 23 (2016), pp. 55-63

[41] S.R. Hong, S. Jung, E.H. Lee, K. Shin, W.I. Yang, H.Y. Lee. DNA methylation-based age prediction from saliva: high age predictability by combination of 7 CpG markers. Forensic Sci. Int. Genet., 29 (2017), pp. 118-125

[42] J. Park, J.H. Kim, E. Seo, D.H. Bae, S. Kim, H. Lee, et al. Identification and evaluation of age-correlated DNA methylation markers for forensic use. Forensic Sci. Int. Genet., 23 (2016), pp. 64-70

[43] D.G. Hernandez, M.A. Nalls, J.R. Gibbs, S. Arepalli, M. van der Brug, S. Chong, et al. Distinct DNA methylation changes highly correlated with chronological age in the human brain. Hum. Mol. Genet., 20 (2011), pp. 1164-1172

[44] C.M. Koch, W. Wagner. Epigenetic-aging-signature to determine age in different tissues. Aging (Albany NY)., 3 (2011), pp. 1018-1027

[45] S. Bork, S. Pfister, H. Witt, P. Horn, B. Korn, A.D. Ho, et al. DNA methylation pattern changes upon long-term culture and aging of human mesenchymal stromal cells. Aging Cell, 9 (2010), pp. 54-63

[46] E. Schilling, M. Rehli. Global, comparative analysis of tissue-specific promoter CpG methylation. Genomics, 90 (2007), pp. 314-323

[47] P. Garagnani, M.G. Bacalini, C. Pirazzini, D. Gori, C. Giuliani, D. Mari, et al. Methylation of ELOVL2 gene as a new epigenetic marker of age. Aging cell, 11 (2012), pp. 1132-1134

Chapter V: DNA methylation assay based on pyrosequencing for determination of smoking status

This is the pre-peer reviewed version of the following article: Alghanim, H., Wu, W., & McCord, B. (2018). DNA methylation assay based on pyrosequencing for determination of smoking status. Electrophoresis, 39(21), 2806-2814. This article has been published in final form at:
https://onlinelibrary.wiley.com/doi/abs/10.1002/elps.201800098

1. Abstract

The goal of this study was to utilize pyrosequencing to identify CpG sites indicative of tobacco smoking using DNA sequences surrounding ten frequently reported smoking- related CpGs. Initially, six genetic loci were investigated including AHRR, 2q37, 6p21.33, GFI1, F2RL3, and MYO1G in order to detect novel CpG sites associated with tobacco smoking. The methylation data revealed a set of 23 consecutive CpG sites in blood (Chr5:373,115-Chr5:373,653) that were significantly hypomethylated in current smokers. In addition, 10 of these 23 CpGs were also significantly hypomethylated in the saliva of current smokers. The most significant CpG sites were located at Chr5:373,490 in blood and Chr5:373,476 in saliva with a decrease in methylation in current smokers of 42.3% and 21.3% respectively. In the model-building steps of this study, a quick 4-CpG assay was developed. The assay consisted of the top ranked CpG sites in blood and saliva. The assay was applied in a leave-one-out approach to test its ability to infer an individual's self-identified history of smoking habits. A multinomial logistic regression model (MLR) containing all 4 CpG sites gave the most accurate results in blood and saliva. In blood, the model correctly predicted 90.0% of current smokers, 66.7% of former smokers, and 84.9% of never smokers. In addition, the MLR model correctly predicted 86.9% of current smokers, 54.5% of former smokers, and 77.8% of never smokers in saliva. These results demonstrate that this pyrosequencing-based assay can provide an effective tool for identifying individuals who smoke tobacco, particularly when using epigenetic markers in blood.

2. Introduction

DNA methylation is an epigenetic modification that permits regulation of gene expression. Generally, this occurs through the methylation of cytosine residues in the CpG dinucleotide sites by DNA methyl transferases (DNMT). The methylation process can inhibit gene transcription by recruitment of chromatin remodeling factors that influence the accessibility of DNA during transcription [1]. Recently, it has been recognized that environmental factors play an important role in modifying the levels of DNA methylation at these sites. For example, differences in DNA methylation levels between monozygotic twins have been observed to be greater for the pairs that spend less of their lifetime together or exhibit different lifestyles [2,3]. Environmental exposures such as diet, stress, and smoking can alter DNA methylation at various stages of human development. Amongst these drivers, tobacco smoking is considered one of the most powerful environmental factors that cause DNA methylation changes [4].

There are several different mechanisms by which tobacco smoking can alter DNA methylation [5]. First, smoking can modulate methylation patterns through carcinogen-induced DNA damage and repair. Various carcinogenic materials in tobacco smoking, such as arsenic, nitrosamines, polycyclic aromatic hydrocarbons, and formaldehyde, can cause double-stranded DNA breaks. Such breaks require DNA repair which is mediated by DNA methyltransferase 1 (DNMT1) to methylate the CpGs adjacent to the repaired sites [6]. Second, smoking can also alter DNA methylation levels through a nicotine effect on gene

expression [7]. Nicotine has the ability to alter DNMT1 activity and affect protein expression [8]. Third, smoking can modify DNA methylation by affecting the expression and activity of DNA-binding factors such as Sp1. Smoking is proposed to increase Sp1 expression which binds to GC-rich motifs in gene promoters and subsequently prevent de novo methylation of CpGs at these motifs [9,10]. Fourth, hypoxia is another way by which tobacco smoking may alter DNA methylation. Tobacco smoke contains carbon monoxide that binds to hemoglobin and reduces the oxygen levels in the tissue. In turn, hypoxia may upregulate methionine adenosyltransferase 2A that is responsible for S-adenosylmethionine synthesis, a key methyl donor for any DNA methylation [11]. Thus, variations in DNA methylation are one mechanism that can potentially mediate the effects of tobacco smoking in individuals.

A variety of chip array-based platforms (i.e. Illumina 27K and 450K) have been developed to permit a much broader investigation and identification of differentially methylated loci across the genome. Several genetic loci have appeared as robust indicators of tobacco smoking as a result of investigations utilizing these arrays based platforms. The first consistent locus to be discovered was the coagulation factor II (thrombin) receptor-like 3 (*F2RL3*), by Breitling et al. [4]. The group used a 27K array to study the effect of smoking on peripheral mononuclear cell pellets and identified several loci that were associated with smoking including *F2RL3*, *GPR15*, and *ORAI2*. The second important locus to emerge was the aryl hydrocarbon receptor repressor (*AHRR*) uncovered by Monick el al. [12]. In this study, the effect of smoking in lymphoblast and lung macrophage DNA was

examined using the Illumina Human Methylation 450K BeadChip. The results demonstrated that tobacco smoking can cause significant changes in DNA methylation patterns at various genomic loci and especially at *AHRR*. A large study carried by Zeilinger et al. [13] further confirmed the above noted loci and extended the list of genes to include *HIVEP3* and *CACNA1D*. Other genetic loci have also emerged to contain smoking-specific CpG sites including 2q37, 6p21.33, growth factor independent 1 transcription repressor (*GFI1*), myosin IG (*MYO1G*), *CPOX*, *GPR15*, *CYP1A1*, and many others [4,12-18].

The DNA methylation signatures of candidate sites have been shown to serve as useful biomarkers for various traits. Interest in such applications has resulted in several genome wide association studies using large scale epigenetic arrays. However, because DNA methylation analysis is mainly performed by array studies which require laborious bioinformatic analysis, applying DNA methylation is still difficult in the clinical and forensic regimes due to the complexity of the instrumentation and the need for relatively large sample quantities. We decided to examine and sequence some of these locations using bisulfite modified PCR followed by pyrosequencing. Pyrosequencing is a technique that can measure the relative methylation level at each CpG site at high accuracy [19]. This technique is quick, easy and has the potential to readily screen large numbers of samples.

The goal of this project was to identify novel CpG sites indicative of tobacco smoking by investigating the vicinity around some of the previously reported smoking-specific CpG probe sites [4,12-16] using quantitative pyrosequencing. The second goal of the project was to develop an inexpensive and accurate assay

to measure the DNA methylation patterns in blood and saliva to be used in differentiating between the smoking status of various individuals.

3. Materials and methods

3.1 Sample collection

Blood and saliva (buccal swab) samples (n = 161 each) were collected from volunteers. Whole blood and saliva samples were collected on a cotton swab. In the present study, the subjects were categorized into three groups based on their self-reported smoking history: never smokers, former smokers, and current smokers as described in Table 3.1. Participants in the study were aged from 6–87 years and whose smoking status was determined based on standardized self-administered questionnaires. The category included individuals who smoked a variety of tobacco products, predominately cigarettes and few used pipes. Table 3.1 provides a description of the sample types collected in this study. The entire samples were randomly divided into three sets: discovery, training, and validation. All biological samples were collected according to the protocol approved by the Institutional Review Board at Florida International University under IRB-16-0341 and General Headquarters of Dubai Police (approval letter #410126/11/33/3583). All participants were asked to sign the informed consent forms prior to sample collection.

Table 3.1. Demographic characteristics of the populations used in this study. *: for blood samples, #: for saliva samples

Description	Never Smoker	Former Smoker	Current Smoker	Total
	Lifetime never smoke	Abstinent from smoking for the past 1+ years, but used to smoke tobacco ≥ 5+ times every day continuous for the past 3+ years	Smoking tobacco ≥ 5+ times every day continuous for the past 3+ years	
Blood samples	n=59	n=42	n=60	n= 161
Saliva samples	n=58	n=41	n=62	n= 161
Ages, Mean (SD)	30* (±13)*, 29# (±13)#	55* (±14)*, 54# (±15)#	40* (±14)*, 41# (±14)#	
Pack-year, Mean (SD)		24.4* (±17)*, 24.3# (±18)#	14.9*# (±14)*#	
Time since quitting, Mean (SD)		14.8*# (±12)*#		
Sampling Location	Dubai=51*, Miami=8*, Dubai=51#, Miami=7#	Dubai=42*, Miami=0*, Dubai=41#, Miami=0#	Dubai=54*, Miami=6*, Dubai=55#, Miami=7#	Dubai= 148*#, Miami= 14*#
Sex	M=47*, F=12*, M=48#, F=10#	M=42*, F=0*, M=41#, F=0#	M=58*, F=2*, M=60#, F=2#	M=148*, F=13*, M=149#, F=12#

3.2 Screening strategies

In the discovery step, two screening steps were performed in order to test the association with tobacco smoking. The two screening steps were executed using random subsets consisting of blood and saliva samples (n = 6-12 per sample type) for each set of current and never smokers. The first round of screening targeted the methylation profiles at six genetic loci in order to identify their relative

correlation with tobacco smoking. Top candidate loci were then further examined in a second round of screening. The genetic loci tested, CpG sites examined, sample size, and mean methylation levels obtained in the first and second rounds of screening are summarized in Supporting Information Tables 3.1 and 3.2, respectively. This discovery step was important to identify the top ranked methylation sites and the best candidates to be used as biomarkers for tobacco smoking. Table 3.2 contains the list of the top ranked CpG sites identified from the discovery step.

Table 3.2: The top ranked and the significant CpG sites identified based on Benjamini-Hochberg method used to control false discovery rate at a level of 0.05.

Rank number	In Blood		In Saliva	
	Locus	Chromosome location (GRCh37)/ (Illumina ID)	Locus	Chromosome location (GRCh37)/ (Illumina ID)
1	AHRR	Chr5:373,490	AHRR	Ch5:373,476
2	2q37.3	Chr2:233,284,675	AHRR	Ch5:373,494
3	AHRR	Chr5:373,423	AHRR	Ch5:373,423
4	AHRR	Chr5:373,476	AHRR	Ch5:373,490
5	AHRR	Chr5:373,378 (cg05575921)	AHRR	Ch5:373,398
6	AHRR	Chr5:373,494	AHRR	ch5:395,488
7	AHRR	Chr5:373,315	AHRR	Ch5:374,018
8	AHRR	Chr5:373,299 (cg23576855)	AHRR	Ch5:373,250
9	AHRR	Chr5:373,651	AHRR	Ch5:373,147
10	AHRR	Chr5:373,398	AHRR	Ch5:373,989
11	AHRR	Chr5:373,653		
12	AHRR	Chr5:373,555		
13	GFI1	Chr1:92,947,588 (cg09935388)		
14	2q37.3	Chr2:233,284,662 (cg21566642)		
15	F2RL3	Chr19:17,000,553		

3.3 DNA extraction and bisulfite conversion

DNA was extracted from blood and saliva samples using an organic extraction method involving proteinase digestion followed by phenol-chloroform-isoamyl alcohol extraction (Fisher Scientific, NJ) [20]. A total DNA volume of 50 μL was recovered and quantified using an Alu-based real-time PCR method with a Rotor-Gene 6000 (Corbett Research, Sydney, Australia, now Qiagen Inc., CA) [21]. Two-hundred to five hundred nanograms of extracted DNA were bisulfite-modified using the EpiTect® Fast DNA Bisulfite Kit (Qiagen Inc., CA) to convert the unmethylated cytosines to uracils.

3.4 Assay design

Different epigenome association studies have reported various genetic loci in which methylation levels were associated with tobacco smoking. We focused on ten CpG sites at six genetic loci that show significant association with the smoking effect. These six genetic loci (AHRR, 2q37, 6p21.33, GFI1, F2RL3, MYO1G) were frequently reported in earlier studies as being differentially methylated according to tobacco smoke exposure [4, 12–16]. However in these previous reports, only a single CpG site indicated by the array probe was tested at each locus. Here, we examined these loci further by investigating a broad range of CpG sites in each genetic locus with special emphasis in the vicinity around cg05575921 probe site at AHRR. Specific sets of PCR primers were designed using PyroMark Assay Design 2.0 software (Qiagen Inc., CA) to amplify the bisulfite modified target region. The designed assays targeted between two to fourteen CpG sites. One of

the PCR primers was biotin labeled to produce biotinylated PCR amplicons needed for the pyrosequencing reaction. Supporting Information Table 3.3 shows all the primer sequences used to examine the selected CpG sites in this study. Finally, we designed a primer set targeting four consecutive CpG sites at AHRR to serve as a biomarker assay for tobacco smoking (Supporting Information Table 3.4).

3.5 PCR and pyrosequencing

PCR reactions were carried out in a singleplex fashion by utilizing the PyroMark® PCR kit (Qiagen Inc., CA) on the GeneAmp® PCR system 9700 (Applied Biosystems, Foster City, CA). The PCR reaction was modified to utilize a 15 μL reaction volume [22]. The pyrosequencing was performed using a Pyromark Q24 pyrosequencer (Qiagen Inc., CA) following the manufacturer's recommended protocols. Pyromark® Q24 software was used to calculate the percent methylation for each CpG site. The results were displayed as a pyrogram with the methylation percentage. Supporting Information Figure 3.S (1A–1C) shows examples of pyrogram results which represent the difference in methylation level between current, former, and never smokers, respectively.

3.6 Statistical analysis and model building

A test of normality was performed using a Shapiro-Wilks test. Because non-normal distributions were detected, smoking-dependent differences in median methylation at single CpG sites between two smoking groups (current smokers versus never

smokers) in blood and saliva were tested using a non-parametric Mann-Whitney U-test. For the entire sample set, a nonparametric Kruskal-Wallis test was used to compare the differences in median between each of the three smoking groups. Box plots were used to demonstrate the distribution of methylation patterns for the CpG sites across the three smoking groups. P-values of 0.05 were considered to be significant. The Benjamini-Hochberg method was used to control false discovery rate at a level of 0.05 when significant CpGs were claimed among many [23]. After the discovery set, the remaining samples collected (n = 143 for blood and n = 139 for saliva) consisting of the three smoking groups were used to build a biomarker assay, and to test the performance of the model. The training set was used for model-building followed by a validation step using an independent set of samples. A receiver operating characteristic (ROC) analysis was used to test the performance of each CpG predictive model using a five-fold cross validation at each genomic locus tested in the assay. Based on combined and stepwise multinomial logistic regression (MLR) analysis, a model was constructed using methylation data for the 4-CpG assay with the training set, and then the accuracy of each predictive model was examined with a validation set using a leave-one-out (LOO) approach. This stepwise MLR model used an iterative process in order to select a biomarker based on the four CpGs that would provide optimal discrimination power. All the analyses were performed using SPSS statistics software ver. 23.

In this study, potential confounding effects of age and sex in the methylation data were tested and found to be insignificant. Although there are differences in mean age between the smoking groups, the effect of age on the level of methylation at the four CpGs becomes insignificant in blood (p-value 0.238) when controlled by smoking status. The age effect in the saliva DNA methylation level was also insignificant except for CpG1 due to the presence of outliers in the data. When excluding these outliers at CpG1 in saliva from the methylation data, the age effect at this locus was also insignificant when controlled by smoking status (p-value 0.1). The effect of sex, location, and type of tobacco smoking used were tested using a nonparametric Mann-Whitney test in blood and saliva samples. The effect of sex in the methylation pattern at the four CpG sites was found to be insignificant in blood and saliva samples (p-value 0.082). In addition, the samples were collected from two different geographical locations (Dubai and Miami). The confounding effect due to sampling location (environment) was also examined and eliminated (p-value 0.082). Finally, we tested whether the type of smoking (cigarettes vs. pipes) would have an effect in DNA methylation patterns. The results also indicate insignificant correlation between smoking type and the DNA methylation pattern in blood and saliva samples (p-value 0.207). This means that the difference in mean methylation of these CpGs is directly due to smoking status. Age, sex, locations, and smoking type had no significant influence. Multiple studies have als confirmed that age and sex have no effect in DNA methylation at various CpG sites in various loci including AHRR [13, 24].

4. Results

4.1 Discovery step

4.1.1 Screening areas near previously reported smoking-specific CpG sites

As a first step in this study, we explored the DNA methylation profiles in current and never smokers for ten of the most frequently reported CpG sites in recent epidemiological studies (cg05575921 in AHRR, cg03636183 in F2RL3, cg09935388 in GFI1, cg01940273 in 2q37, cg25648203 in AHRR, cg21161138 in AHRR, cg06126421 in 6p21.33, cg21566642 in 2q37, cg23576855 in AHRR and cg12803068 in MYO1G). All ten loci were previously found to be differentially methylated upon tobacco smoking in blood [4, 12–16]. We started by examining the methylation status of these 10 selected probe sites along with 22 nearby CpG sites. From the total samples collected, we randomly selected 6–12 participants for each group of current and never smokers and the results of the mean methylation profiles for each of the groups in blood and saliva samples were used to evaluate the association of these 32 CpGs with tobacco smoking (Supporting Information Table S3.1). In this part of the discovery step, significant differences in methylated CpG sites (U-test p-value 0.05) could be detected between current and never smokers in 19 CpG sites in blood and 5 CpG sites in saliva (Supporting Information Table S3.1). The majority of tested sites exhibited smoking-induced methylation changes when tested using blood. Among all the genomic loci tested at this step, we found that AHRR gene and in particular the area located around cg05575921 contained the highest number of smoking-specific CpG sites.

Therefore, we decided to further investigate a wider range of CpG sites located around this particular probe site in AHRR.

4.1.2 Epigenomic screening around cg05575921

Continuing with the discovery step, a second round of screening was used to investigate additional CpG sites near the cg05575921 probe in AHRR. This part of the analysis was focused on identifying the most significant differentially methylated CpGs as a result of tobacco smoking in the selected region. A total of 56 CpGs near cg05575921 probe in AHRR were examined by recording the DNA methylation values of current and never smokers in blood and saliva (n = 6-9, for each group per sample type). We designed 11 primer sets to screen the methylation levels at the selected CpG sites. Supporting Information Table 3.2 shows the results of the mean methylation percentage at the 56 CpG

sites for current smokers versus never smokers. From the methylation data obtained by pyrosequencing, a new set of CpG sites was identified showing significant hypomethylation with current smokers in blood and saliva. Combining the results from Supporting Information Tables 3.1 and 3.2, it can be observed that a cluster of 23 consecutive CpG sites starting from the CpG site at position Chr5:373,115 to the CpG site at Ch5:373,653 was significantly hypomethylated with current smokers in blood (p-value 0.05). In addition, 10 out of the 23 consecutive CpGs were also significantly associated with tobacco smoking in saliva. The Benjamini-Hochberg method at a 0.05 false discovery rate was used to detect and rank the most significant CpGs tested (Table 3.2). The Benjamini-Hochberg method identified 15 and 10 CpG sites showing significant decrease in

methylation level with current smokers in blood and saliva, respectively. The most striking and significant CpG site identified was at Chr5:373,490 with a decrease in mean methylation of 42.3% (p-value = 3.39×10^{-4}) in the blood of current smokers. In saliva, the CpG sites at Chr5:373,476 showed the greatest decrease in mean methylation for current smokers equal to 21.3% (p-value = 2.98×10^{-4}).

4.2 Model-building and validation steps

Another important target of this study was to find an optimal set of biomarkers for tobacco smoking in blood and saliva. The aim was to develop a quick and easy assay that could distinguish current, former, and never smokers. We designed an assay that consisted of the top ranked CpG sites identified in the discovery analyses for blood and saliva. The assay was composed of four consecutive CpG sites at AHRR located at Chr5:373,476 (CpG1), Chr5:373,490 (CpG2), Chr5:373,494 (CpG3), and Chr5:373,529 (CpG4) that were highly smoking specific (Table 3.3). This assay is capable of screening methylation profiles utilizing a singleplex bisulfite modified PCR followed by pyrosequencing. Box plots in Figures 3.1 and 3.2 show the significant difference in methylation levels between never, former, and current smokers for the 4 CpG sites in blood and saliva, respectively.

Table 3.3: Summary of average ROC area under the curve (AUC) of the fivefold cross validation for each of the 4 CpG sites in the biomarker assay in blood and saliva.

Genomic locus	CpG position number	AUC for current versus never smokers		AUC for current versus former smokers		AUC for never versus former smokers	
		In blood	In saliva	In blood	In saliva	In blood	In saliva
Chr5:373,476	CpG1	0.981	0.959	0.941	0.875	0.732	0.701
Chr5:373,490	CpG2	0.969	0.943	0.919	0.804	0.751	0.695
Chr5:373,494	CpG3	0.972	0.963	0.927	0.850	0.720	0.779
Chr5:373,529	CpG4	0.951	0.830	0.864	0.721	0.722	0.567

Figure 3.1: Box plots of distribution for methylation at each CpG sites included in the 4-CpG assay for never, former, and current smokers in blood. The boxes in the plots represent the 25% and 75% percentile, whiskers represent the non-outlier range, dots indicate outliers, and stars show extreme outliers.

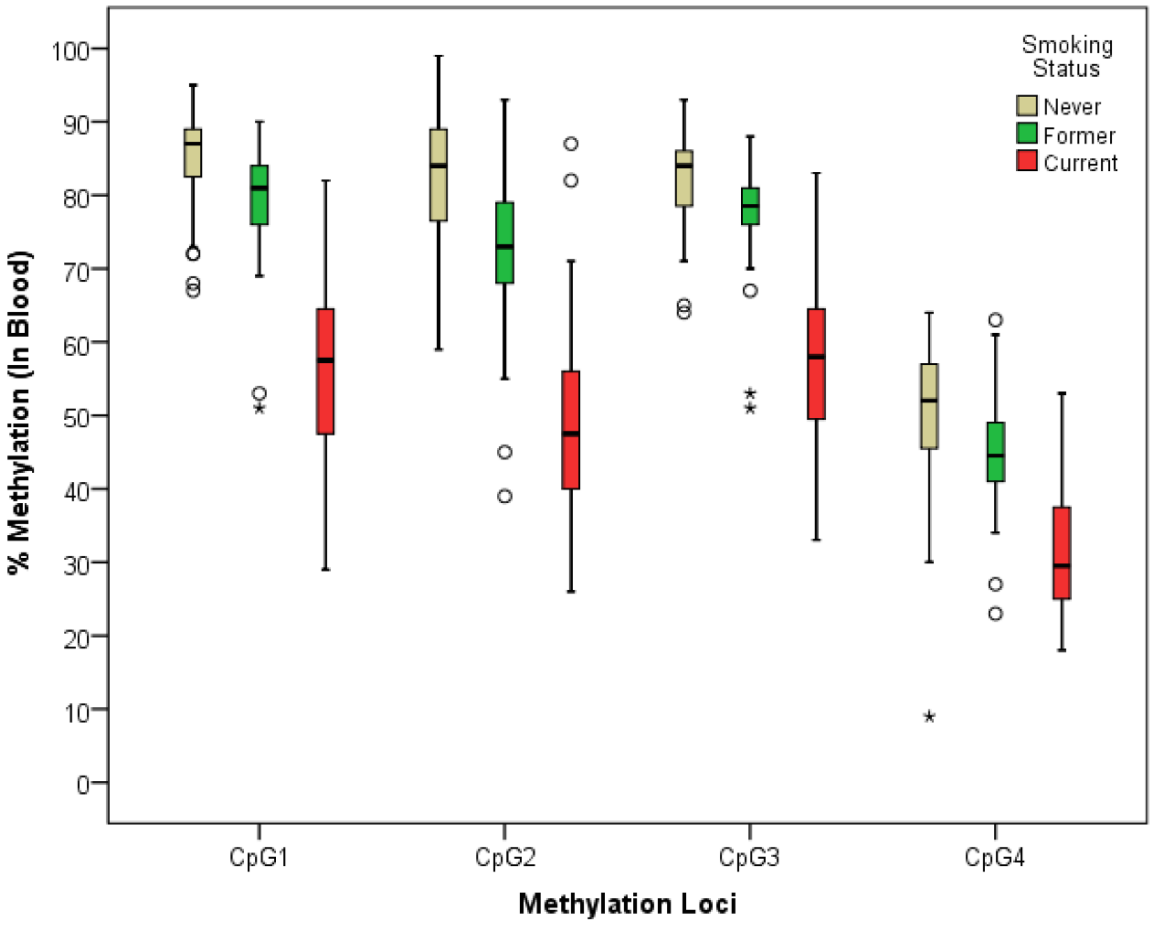
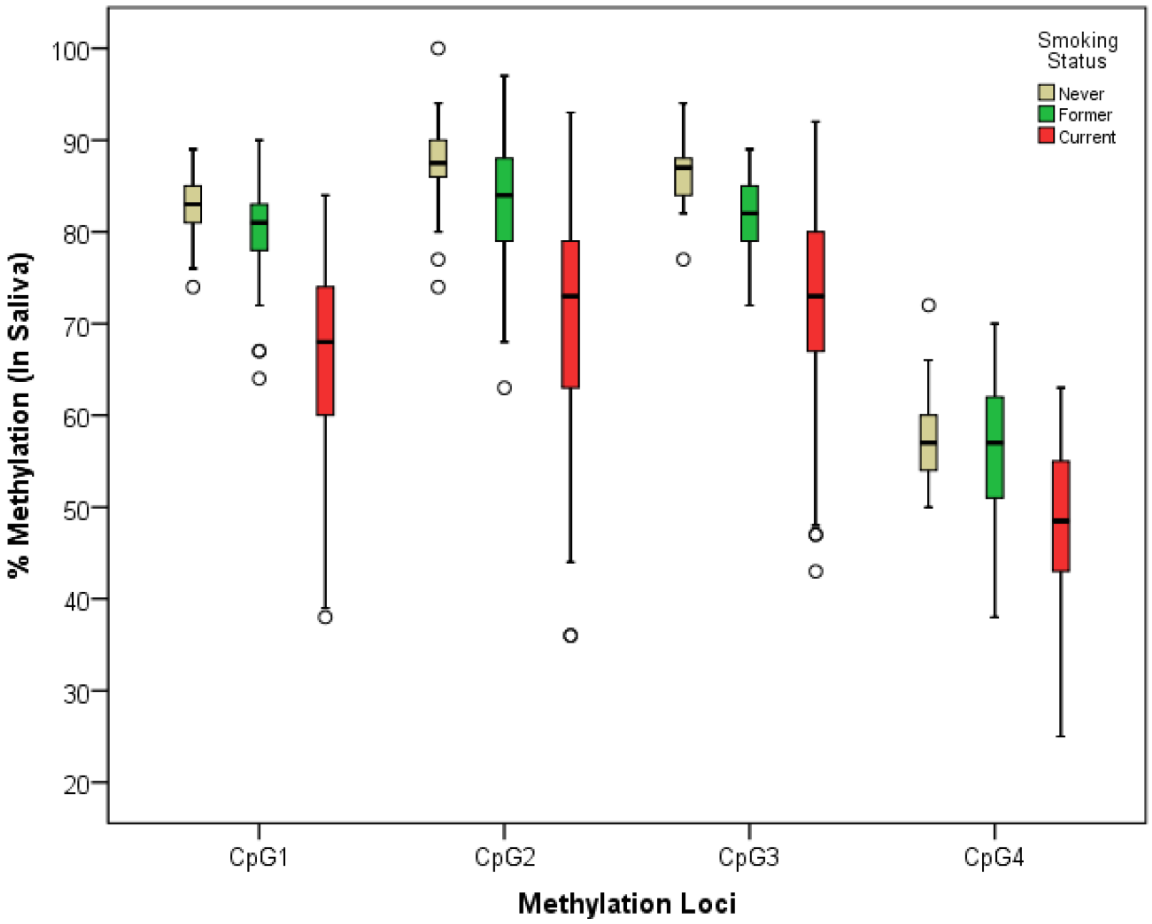


Figure 3.2: Box plots of distribution for methylation at each CpG sites included in the 4-CpG assay for never, former, and current smokers in saliva. The boxes in the plots represent the 25% and 75% percentile, whiskers represent the non-outlier range, dots indicate outliers, and stars show extreme outliers.



Finally, the assay was built and tested using a different sample set from those used in the discovery set (n = 141 for blood and n = 139 for saliva) in which the subjects were categorized into three groups based on self-reported smoking history (Table 3.1). Based on ROC analyses in a fivefold cross validation, the average areas under the receiver operating characteristic curve (AUCs) were calculated for each CpG site used to discriminate current versus never smokers, current versus former smokers, and never versus former smokers (Table 3.3). The AUC values indicate

that these CpGs worked the best to discriminate current from never smokers and current from former smokers. CpG1, CpG2, and CpG3 individually were the most significant and provided similar discrimination power in distinguishing current smokers from never or former smokers with AUC values ranging from 0.981 to 0.919 in blood and 0.963 to 0.804 in saliva. Among all CpG sites in the assay, CpG4 gave the lowest AUC values in blood and saliva. Data generated from the ROC analyses demonstrated a lower discrimination power to distinguish former smokers from never smokers in blood and in saliva (Table 3.3). As a last step in our analyses, the performance of each of the four CpG sites in the assay were tested by performing a fivefold cross validation ROC analysis in the training sets (n = 114 for blood and n = 111 for saliva). Two cut off values were selected (Tables 3.4 and 3.5), one to separate current from former smokers, and another to separate never smokers from former smokers. Thus, any samples that have methylation values below the first cut off (between current and former) would be classified as current smokers and any samples above the second cut off (between never and former) would be classified as never smokers. Finally, those samples that fell between the two cut off values were classified as former smokers. The cut off values were selected in order to achieve the highest sum of specificity and sensitivity. Using the LOO approach, the accuracy of an independent validating set (n = 29 for blood and n = 28 for saliva) of each group was recorded and averaged (Tables 3.4 and 3.5). The performance of individual CpG site using ROC analyses was also compared with models of multiple CpG sites using the MLR model based on two methods: a combined method using all four CpG sites and a

stepwise method using a subset of the four CpG sites. The MLR models were built using a training set (n = 114 for blood and n = 111 for saliva) and then the accuracies of the models were tested using an independent set (n = 29 for blood and n = 28 from saliva). Overall, the combined MLR model using the four CpG sites provided the best accuracy for the smoking status correctly predicting 90.0% of current smokers, 66.7% of former smokers, and 84.9% of never smokers in blood (Table 3.6). In saliva, the combined MLR method gave the best accuracy overall correctly predicting 86.9% of current smokers, 54.5% of former smokers, and 77.8% of never smokers (Table 3.6).

Table 3.4: Average accuracy of prediction for each of the four CpGs of the assay in blood using fivefold cross validation.

Genomic locus	CpG position number	Cutoff value for current versus former smoker	Cutoff value for never versus former smoker	Accuracy of prediction in blood		
				Current smoker	Former smoker	Never smoker
Chr5:373,476	CpG1	68.5%-69 %	83.5-87.5%	86.7%	69.0%	53.7%
Chr5:373,490	CpG2	62.5%-63%	80.5%	89.6%	72.3%	57.0%
Chr5:373,494	CpG3	69.5%-70.5%	81.5%-82.5%	88.14%	73.5%	59.2%
Chr5:373,529	CpG4	39.5%-40%	51.5%	84.1%	62.2%	55.5%

Table 3.5: Average accuracy of prediction for each of the four CpGs of the assay in salvia using fivefold cross validation.

Genomic locus	CpG position number	Cutoff value for current versus former smoker	Cutoff value for never versus former smoker	Accuracy of prediction in saliva		
				Current smoker	Former smoker	Never smoker
Chr5:373,476	CpG1	74.5%	81.5%	81.7%	40.4%	75.4%
Chr5:373,490	CpG2	78.5%-81.5%	83.5%	80.0%	18.3%	82.7%
Chr5:373,494	CpG3	75.5%-81.5%	83.5%	69.7%	42.8%	84.8%
Chr5:373,529	CpG4	52.5%	52.5%	52.4%	17.5%	93.6%

Table 3.6: Multinomial logistic regression (MLR) models for the 4-CpG assay based on Leave-one-out approach.

Type of Model	CpG utilized	Accuracy of prediction in blood			
Combined MLR	All 4 CpGs	Current smoker	Former smoker	Never smoker	Total
		90.0%	66.7%	84.9%	82.7%
Stepwise MLR	CpG2 + CpG4	81.8%	%60.0	84.6%	79.3%
Type of Model	CpG utilized	Accuracy of prediction in saliva			
Combined MLR	All 4 CpGs	Current smoker	Former smoker	Never smoker	Total
		86.9%	54.5%	77.8%	71.4%
Stepwise MLR	CpG1+ CpG3+ CpG4	87.4%	54.5%	66.7%	67.9%

5. Discussion

Several genome-wide association studies have investigated and identified loci whose methylation patterns were associated with tobacco smoking. Furthermore, Monick et al. [12] made an important finding by identifying that smoking produces

strong changes in DNA methylation at AHRR. Since then the changes of AHRR methylation in response to smoking have been the most consistently replicated findings in various studies, suggesting that this locus is the most probative location for the detection of tobacco smoking in the human genome [5, 25]. The AHRR gene is located on chromosome 5p15.33. The gene encodes a protein that has an important role in the aryl hydrocarbon receptor (AHR) pathway which is used for the detoxification of toxins such as polyatomic hydrocarbons and dioxins present in burnt products such as tobacco smoking [26]. Toxic substances from tobacco smoking enter the bloodstream through the alveolar capillary system which can then alter the methylation patterns of the white blood cells [14]. In particular, the most smoking-specific CpG site reported within the AHRR gene is found at the cg05575921 array probe (Chr5:373,378) located within intron 3. This region in AHRR consists of an enhancer motif that upon demethylation promotes recruitment of DNA complex C2 and C3 and a subsequent increase in AHRR mRNA production [12, 13, 25]. Another important CpG probe in AHRR is cg23576855 located 79 bp downstream (Chr5:373,299) from cg05575921 [14, 16]. In the previous work, we have demonstrated that single CpG sites identified by epigenome wide array studies commonly contain a cluster of nearby CpG sites that show similar behavior toward a particular trait [22, 27, 28]. These nearby CpG sites can be identified, and their level of methylation can be quantified using pyrosequencing. The first objective of this study was to identify novel methylation sites associated with tobacco smoking that had not been previously examined using pyrosequencing. In this study, we examined a range of methylation sites at

six genetic loci including AHRR, 2q37 intergenic region, 6p21.33 intergenic region, growth factor independent 1 transcription repressor (GFI1), and myosin IG (MYO1G). The methylation sites in and around ten of the most significantly associated and replicated CpG probe sites related to tobacco smoking on these six genetic loci were examined (Supporting Information Table S3.1). In this part of the discovery step, novel CpG sites that were hypomethylated in current smokers were detected in the nearby sequences at the majority of the probes tested, especially in blood. For example, new clusters of CpG sites associated strongly with tobacco smoking in blood were observed near cg09935388, cg21566642, and cg12803068. In particular, CpG sites in sequences surrounding the cg05575921 probe in AHRR were found to contain the most significant number of smoking-specific CpG clusters. The next part of the discovery step was to further investigate the CpG sites around this particular probe site. An additional 56 CpG sites were selected and evaluated within the AHRR locus from which 20 CpGs in blood and 13 CpGs in saliva were observed to be significantly associated with tobacco smoking (p-values 0.05) (Table S3.2). Interestingly, by combining the results from Supporting Information Tables S3.1 and S3.2, a cluster of 23 CpG sites (Chr5:373,115-Chr5:373,653, http://genome.ucsc.edu/GRCh37/hg19) was found to have the highest number of top-ranked CpG sites associated with tobacco smoking, including the two probe sites cg05575921 and cg23576855 (Table 2). Among the 88 total CpG sites examined in the discovery step, the CpG sites at Ch5:373,490 and Ch5:373,476 were determined to rank the most significant in blood and saliva, respectively (Table 3.2).

The second aim of this study was to develop a quick and easy DNA methylation assay to predict smoking status. This assay was arranged to include the two most significant CpG sites for tobacco smoker status that were detected in the discovery analyses along with two adjacent smoking-specific CpG sites. The biomarker assay consisted of four consecutive CpG sites labeled CpG1 to CpG4 located at Chr5:373,476, Chr5:373,490, Chr5:373,494, and Chr5:373,529 in AHRR, respectively. All four CpG sites in the assay were examined using blood and saliva samples through bisulfite-modified PCR and pyrosequencing in which the participants were divided according to their self-reported smoking behavior into current, former, or never smokers (Table 3.1). All CpG sites showed significant differences in methylation levels between the three smoking groups in blood and saliva as shown in Figures 3.1 and 3.2, respectively. Based on the methylation data generated, it was also clear that tobacco smoking increased the demethylation of these CpG sites in AHRR. This decrease in methylation is concordant with the results observed from a number of other CpG sites detected at various genetic loci [12, 13, 29, 30]. As shown in Figures 3.1 and 3.2, current and former smokers had low and intermediate methylation levels for all CpG sites respectively, when compared with the data from never smokers. In all four CpGs in the assay, the methylation levels were significantly lower in current smokers than in never smokers and smaller differences were found between former smokers and never smokers in blood and saliva.

Data generated from using this 4-CpG assay was next examined to determine whether a single CpG or combination of some or all CpGs was optimal for

distinguishing smoking status. Using a training set in a fivefold cross validation approach, ROC analyses were performed to build a model using two threshold values that were specifically selected to discriminate between the three smoking groups. CpG1, CpG2, and CpG3 were all capable of distinguishing current smokers from former or never smokers in blood samples (AUC values = 0.981-0.919) and in saliva samples (AUC values = 0.963- 0.804) (Table 3.3). When testing the accuracy of each CpG site using a validation set, CpG1, CpG2, and CpG3 also had the highest accuracies. In addition, these CpG sites performed better in blood DNA than in saliva (Tables 3.4 and 3.5). Finally, the accuracy of various combinations of the four CpGs was tested by the LOO approach. Overall, the combined MLR model containing all four CpG sites was determined to be the best indicator for smoking status in blood and saliva in the validation set (Table 6). The combined MLR method considerably improved the prediction accuracies over that obtained through individual ROC analyses. In the validation sets, the combined MLR model could correctly predict a total of 82.7% in blood samples and gave a total accuracy of 71.4% in saliva samples (Table 3.6). The assay is very effective and provides high accuracy to predict current smokers and never smokers.

Several studies have shown a correlation between time since quitting smoking, with methylation levels eventually reverting to those of never smokers [29]. This trend of methylation profiles seen in former smokers was observed in genetic loci such as F2RL3 locus [20, 29]. Several studies have reported that the methylation level of former smokers at various genetic loci tends to regenerate and return to

the level close to that of never smokers with increasing time from cessation [31–33]. This could be the main reason that the AUC values and the accuracies to distinguish former smokers from never smokers using the 4-CpG assay were low. Overall, in this study, the data of former smokers produced an intermediate mean methylation level at the AHRR loci (Figures 3.1 and 3.2).

In summary, our study has identified novel smoking-specific CpG sites in different genes with special emphasis on a 23 CpG cluster located in AHRR that shows a strong association with tobacco smoking. Overall, 15 CpGs in blood and 10 CpGs in saliva showed a significant decrease in methylation level in current smokers. A quick and inexpensive biomarker assay consisting of 4 novel CpG sites at AHRR was developed and determined to be a useful predictor of smoking behavior using bisulfite conversion followed by pyrosequencing.

6. Supplementary

Supplementary Table 3.S1: First set of preliminary data showing mean methylation profiles in current versus never smokers for ten of the most frequently reported CpG sites and some 22 additional CpG sites in the nearby vicinity. the P-value is calculated using Mann-Whitney U-test.

Locus/ Primer set	Chromosome location (GRCh37)/ (Illumina ID)	CpG position number	Smoking status (mean methylation percentage)/ Body fluid					
			Current/ Blood	Never/ Blood	*P*-value	Current/ Saliva	Never/ Saliva	*P*-value
			n=(8-12)	n=(8-12)		n= (8-12)	n=(8-12)	
AHRR/ Set 1	Chr5:373,378 (cg05575921)	CpG1	57.7	84.2	0.000477	43.5	51.9	0.060405
	Chr5:373,398	CpG2	55.8	77.6	0.000777	45.2	58.3	0.001986
	Chr5:373,423	CPG3	56.1	87.2	0.000398	67.4	80	0.001413
AHRR/ Set 2	Chr5:395,445 (cg25648203)	CpG1	78.7	89.7	0.004569	85.2	89.4	0.007540
	Chr5:395,464	CpG2	64.9	70.8	0.170877	75.1	81.3	0.037348
	Chr5:395,488	CpG3	43.2	52.4	0.063280	50.5	67.1	0.002745
AHRR/ Set 3	Chr5:399,361 (cg21161138)	CpG1	72.2	75.6	0.013714	45.1	42.4	0.677356
AHRR/ Set 4	Chr5:373,299 (cg23576855)	CpG1	61	86.3	0.000735	59.4	65.9	0.116239
	Chr5:373,315	CpG2	59.6	88.4	0.000520	56	68.5	0.104571
2q37.1/ Set 5	Chr2:233,284,950	CpG1	12.4	16.8	0.010843	11.4	10	0.435542
	Chr2:233,284,935 (cg01940273)	CpG2	64.4	71.2	0.052912	49.1	46.3	0.303086
2q37.1/ Set 6	Chr2:233,284,662 (cg21566642)	CpG1	50.9	63.7	0.002468	28.2	24.7	0.269080
	Chr2:233,284,672	CpG2	11.8	17.5	0.028904	7.1	6.6	0.617813
	Chr2:233,284,675	CpG3	12.2	21.5	0.000348	6	6.1	0.835531
	Chr2:233,284,691	CpG4	28.1	31.5	0.061651	22.1	19	0.118861
	Chr2:233,284,693	CpG5	17.4	17.5	0.475598	12.1	10	0.263966
	Chr2:233,284,703	CpG6	29.7	32	0.098470	20.7	18.7	0.344980
6p21.33/ Set 7	Chr6:30,720,081 (cg06126421)	CpG1	67.5	81.2	0.025558	21.2	20.1	1
	Chr6:30,720,109	CpG2	32.3	23.3	0.197158	6.1	7.4	0.542298
GFI1/ Set 8	Chr1:92,947,559	CpG1	40.6	54.4	0.040960	10	8.3	0.542606
	Chr1:92,947,567	CpG2	45.3	55.7	0.054694	9.9	9.1	0.909142
	Chr1:92,947,571	CpG3	47.3	54.5	0.139895	15.7	13.9	0.403371
	Chr1:92,947,581	CpG4	53.1	61.1	0.119841	19.1	16.1	0.254865
	Chr1:92,947,586	CpG5	65.2	80.4	0.007772	26.3	19.7	0.240258
	Chr1:92,947,588 (cg09935388)	CpG6	61.4	74.9	0.002464	22.6	17.1	0.305114
F2RL3/ Set 9	Chr19:17,000,553	CpG1	74.3	89.9	0.002787	68.1	73.3	0.785796
	Chr19:17,000,568	CpG2	87.2	87.5	0.884529	82.7	84.3	0.731783
	Chr19:17,000,586 (cg03636183)	CpG3	77.9	80.5	0.594372	66.9	70.8	0.447808
	Chr19:17,000,597	CpG4	72.7	73.6	0.884954	68.2	69.5	0.648789
MYO1G/ Set 10	Chr7:45,002,914	CpG1	74.9	65	0.008538	51.7	46.3	0.145834
	Chr7:45,002,919	CpG2	79.6	68.1	0.026743	54.8	51.1	0.449475

	(cg12803068)							
	Chr7:45,002,931	CpG3	62.9	53.8	0.040605	30.4	27.4	0.523707

Supplementary Table 3.S2: Second set of preliminary data showing mean methylation profiles in current versus never smokers for 56 CpG sites in the nearby vicinity of cg05575921 and cg23576855 probe sites. P-value is calculated using Mann-Whitney U-test.

Locus/ Primer set	Chromosome location (GRCh37)/ (Illumina ID)	CpG position number	Smoking status (mean methylation percentage)/ Body fluid					
			Current/ Blood	Never/ Blood	P-value	Current/ Saliva	Never/ Saliva	P-value
			n=6-9	n=6-9		n=6-9	n=6-9	
AHRR/ Set 11	Chr5:373,115	CpG1	89.3	94.3	0.003639	94.5	96	0.252135
	Chr5:373,119	CpG2	90.3	95.5	0.005938	92.5	95.5	0.035064
	Chr5:373,147	CpG3	85	90.8	0.005938	85.7	92.2	0.004922
AHRR/ Set 12	Chr5:373,193	CpG1	77.3	89.5	0.003700	82.5	87.8	0.292834
	Chr5:373,199	CpG2	88.2	90.8	0.006947	91.7	93	0.093696
	Chr5:373,203	CpG3	77.5	87.8	0.003639	80.2	85.2	0.075046
	Chr5:373,248	CpG4	60.7	71.5	0.003823	64.7	71.7	0.009745
	Chr5:373,250	CpG5	64.2	75.2	0.003885	67.5	74.2	0.004847
AHRR/ Set 13	Chr5:373,353	CpG1	54.6	76	0.003135	55.5	52.2	0.935622
	Chr5:373,355	CpG2	47.8	69	0.004678	50.7	47.2	0.871663
AHRR/ Set 14	Chr5:373,476	CpG1	50.8	87	0.000407	60.7	82	0.000298
	Chr5:373,490	CpG2	41.8	84.1	0.000339	63.3	87.2	0.001933
	Chr5:373,494	CpG3	52.8	83.7	0.000480	69.2	85.8	0.001382
	Chr5:373,529	CpG4	24.8	50.3	0.011805	44	56.7	0.014916
	Chr5:373,555	CpG5	40.8	69.2	0.001249	79.8	83	0.522446
AHRR/ Set 15	Chr5:373,609	CpG1	15.5	20.6	0.030210	50.5	51.4	0.635256
	Chr5:373,651	CpG2	27.4	48.2	0.000771	58.8	66.1	0.635256
	Chr5:373,653	CpG3	20.5	37.1	0.001112	50.6	58.8	0.343139
AHRR/ Set 16	Chr5:373,698	CpG1	10	10.5	0.915927	25.5	26.1	0.915989
	Chr5:373,709	CpG2	9.8	11	0.358243	23.6	24.2	0.916236
AHRR/ Set 17	Chr5:373,783	CpG1	2.3	1.9	0.430556	2.8	3.3	0.332084
	Chr5:373,788	CpG2	2.8	3.8	0.053291	8.9	9.3	0.915677
	Chr5:373,824	CpG3	6.1	8	0.015368	10.5	11.1	0.395197
AHRR/ Set 18	Chr5:373,991	CpG1	4.5	7.2	0.470753	3.2	1.5	0.045092
	Chr5:373,989	CpG2	4.3	6.4	0.348434	5.8	1.2	0.004922
	Chr5:373,985	CpG3	7	7.6	0.942259	6	6	1
	Chr5:373,978	CpG4	4.5	7.6	0.278631	3.2	1.8	0.035791
	Chr5:373,972	CpG5	5	7	0.507195	1.2	1.5	0.588919
	Chr5:373,966	CpG6	4.5	6.8	0.344347	0.83	1.5	0.092674
	Chr5:373,962	CpG7	4	6.6	0.273862	0.83	1.5	0.339556
	Chr5:373,958	CpG8	6.3	6.2	1	5	5.2	0.598161
	Chr5:373,949	CpG9	4.2	6.4	0.088688	1.7	1.7	0.858586
	Chr5:373,931	CpG10	3.5	5.4	0.342968	2.3	2	0.528336
	Chr5:373,929	CpG11	5.3	5	0.662521	3.8	4	0.669764
	Chr5:373,922	CpG12	3.7	4.8	0.557806	1.3	1.2	0.4898
	Chr5:373,915	CpG13	3.7	6.4	0.065438	2	2	0.718646
	Chr5:373,913	CpG14	4.3	5.8	0.717238	0.83	0.83	1
AHRR/ Set 19	Chr5:374,011	CpG1	6.2	3.5	0.463245	2.5	1.7	0.03028
	Chr5:374,013	CpG2	4.2	2.8	0.415063	1.8	1.3	0.092601
	Chr5:374,018	CpG3	11.5	3.2	0.018733	5.3	1.8	0.003926
	Chr5:374,021	CpG4	4.3	3.5	0.86788	2.2	2.5	0.240955
	Chr5:374,024	CpG5	5.3	3.3	0.284957	2.3	1.7	0.055511

	Chr5:374,027	CpG6	2.7	1.5	0.672203	1	0.82	0.651748
	Chr5:374,033	CpG7	4.8	3.2	0.604479	2.3	1.8	0.09169
AHRR/ Set 20	Chr5:375,020	CpG1	90.3	90.7	0.80503	91.5	90.8	0.731428
	Chr5:375,024	CpG2	95.7	96.8	0.515153	97.2	97	0.871208
	Chr5:375,046	CpG3	85.5	86.7	0.246011	82.3	80.2	0.256429
	Chr5:375,066	CpG4	84.3	84.5	1	84.2	82.7	0.286728
	Chr5:375,133	CpG5	84.3	83.8	0.368235	80.3	77.7	0.193029
AHRR/ Set 21	Chr5:375,562	CpG1	95.2	95.7	0.557421	86.5	84.5	0.374269
	Chr5:375,564	CpG2	94.3	93.2	0.063664	95.3	94.8	0.557421
	Chr5:375,580	CpG3	97.5	97.5	1	96	94.7	0.413366
	Chr5:375,594	CpG4	92.2	92.5	0.80684	92.5	91.7	0.346201
	Chr5:375,610	CpG5	90.8	90.3	0.445392	86.7	84.8	0.071929
	Chr5:375,613	CpG6	90	89	0.243345	88.2	86.8	0.116073
	Chr5:375,621	CpG7	97.5	95	0.111293	97.7	94	0.225666

Supplementary Table 3.S3: 21 primer sets used to target the all 88 CpG sites investigated in this study. *:biotinylated primer

Locus	Set #	Sequence	
AHRR	Set 1	Forward	AGGGGTTGTTTAGGTTATAGAT
		Reverse*	AACCCTACCAAAACCACTC
		Sequencing	GGTTTTGGTTTTGTTTTGTA
AHRR	Set 2	Forward	ATAGAGGGGGTTTGGGAGATA
		Reverse*	AAATTCCCCTACTCTAAACTAATAAATCAA
		Sequencing	GTGGTGGGATGTAGTTA
AHRR	Set 3	Forward*	GGGTTGGTGGTGTAGGATATA
		Reverse	AACCCATCCTACCCAAATCCTAATAATTAA
		Sequencing	ATAATTAAAAAACCACCCCTA
AHRR	Set 4	Forward	GTTGGTAGAGTGTTGGTAGGATATA
		Reverse*	CCTCCAAAACCCCAAAAACCAACCTATC
		Sequencing	GGGGTTGTTTAGGTTA
2q37.1	Set 5	Forward*	TTTATGGGAAGGGGGAGG
		Reverse	CCCCACCCCACTTAACCTT
		Sequencing	CCCACTTAACCTTAACT

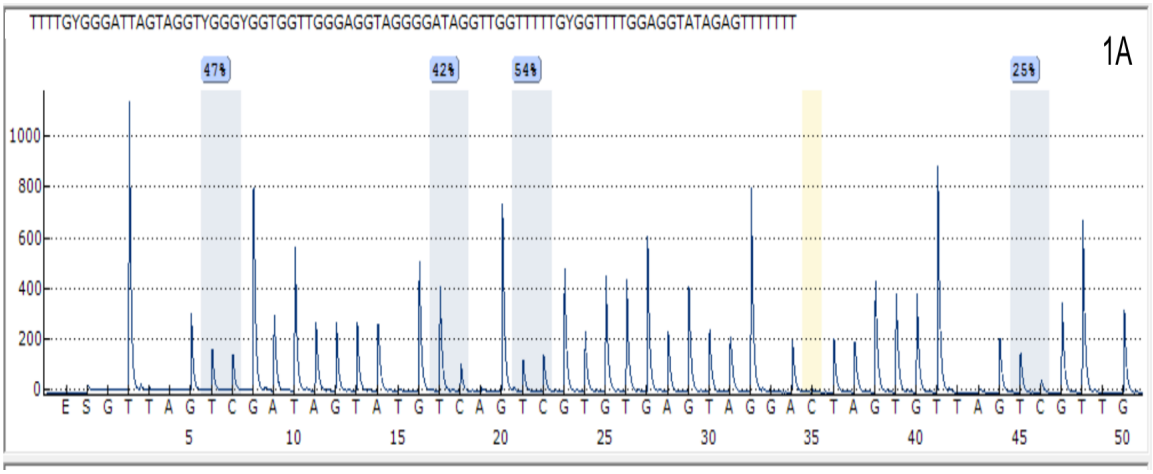
2q37.1	Set 6	Forward	ATGGTTTAGGGGGGTTAAAGT
		Reverse*	CAACCCCTCCCCCTTCCCAT
		Sequencing	AGTAGAGTTAGGTTTAGGA
6p21.33	Set 7	Forward	TTGGAGAATTTGATGGAGATTGAAGTTAA
		Reverse*	ACTATCCCTCCCAACCTTA
		Sequencing	TTTTTTTGAAATTTTATGATTTAGT
GFI1	Set 8	Forward	TTTAGTTTAGGTTGGTTATTTTAGTGAG
		Reverse*	CACCCCTCCCACAATCAATAAATTAACTT
		Sequencing	ATTTTAGTGAGAGGTTGTAT
F2RL3	Set 9	Forward	GTTTTTGGGTTGGGTGTTTATTAG
		Reverse*	CCAACAACAACACTAAACCATACATAT
		Sequencing	GTTTTGGTGGTGGGG
MYO1G	Set 10	Forward	TTTAGGGGTTTTGTTGATAGGGGGAAG
		Reverse*	ACCTCTAAATCTCCCACAATTTCA
		Sequencing	ATAGGGGGAAGTTTG
AHRR	Set 11	Forward	TGAAGAAATAGAGGGTTTTTAGTAGGA
		Reverse*	TTCACTACAACCAAAAAAAAACTCATTTA
		Sequencing	TTTGTTGTGGGTATAGG
AHRR	Set 12	Forward	AATGAGTTTTTTTTTGGTTGTAGTGAAT
		Reverse*	CCCCTATATCCTACCAACACT
		Sequencing	TTTTTTGGTTGTAGTGAATT
AHRR	Set 13	Forward	GTGGGGATTGTTTATTTTTGAGAG
		Reverse*	AACCTATCCCCTACCTCC
		Sequencing	ATTGTTTATTTTTGAGAGGGTA
AHRR	Set 14	Forward	GTTTTGGGAGTGGTTTTGGTAG
		Reverse*	CCCAACCACCCCAATTACCCATAATAAA

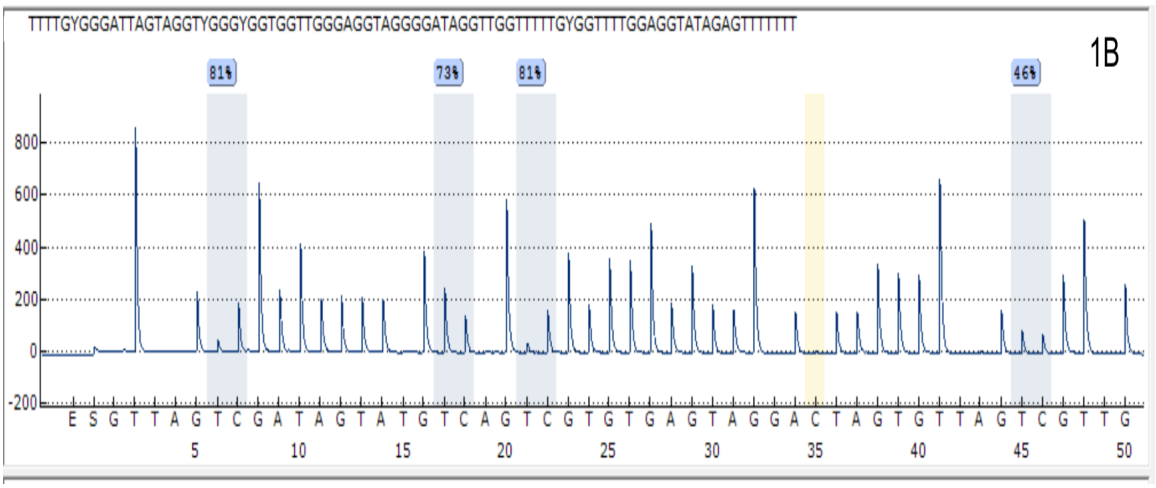
		Sequencing	GTAGGGTTTTTTTTTGTAGA
AHRR	Set 15	Forward	TTGTTGGTTTAGTTGTGTTTTGTAAG
		Reverse*	CCCAACCACCCCAATTACCCATAATAAA
		Sequencing	GTTGTGTTTTGTAAGGG
AHRR	Set 16	Forward	TGTTGGTTTAGTTGTGTTTTGTAAG
		Reverse*	CAACCCCAATCTCCTCCT
		Sequencing	TTTATTTATTTATTATGGGTAATTG
AHRR	Set 17	Forward	AGGAGATTGGGGTTGGAGA
		Reverse*	ACCACCACCTCCCCAAATCC
		Sequencing	GGATGGGGGTTTTTT
AHRR	Set 18	Forward*	AGATTGGGGTTGGAGAGG
		Reverse	CCCACTTCCCCCCTACTT
		Sequencing	AAACCAACCAAACCT
AHRR	Set 19	Forward	GGAGATTGGGGTTGGAGA
		Reverse*	CCCACTTCCCCCCTACTT
		Sequencing	AGGTTTGGTTGGTTTTA
AHRR	Set 20	Forward	TGGTTTGATGGGGAGTAGGTTAA
		Reverse*	TCCACCTACTAATCAAAATAATTCACACTT
		Sequencing	TGGGGAGTAGGTTAAT
AHRR	Set 21	Forward	GGTTTAGTGGGGAGTGAGG
		Reverse*	TATCCCCTCAAACAAATACTACCTTACCAC
		Sequencing	AGGGTTGTGTTTTTTTAAT

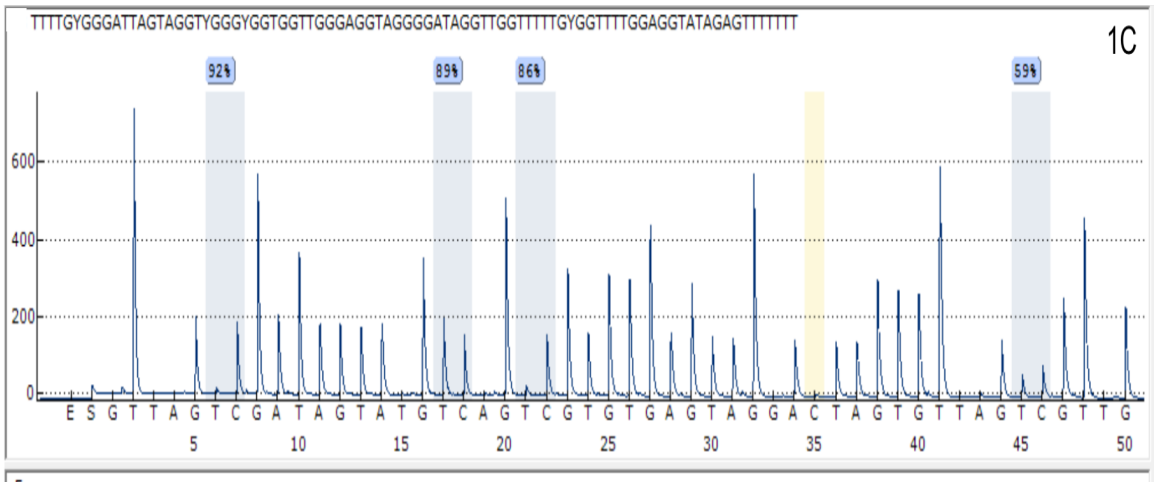
Supplementary Table 3.S4: Assay design and primer sequences targeting 4 CpG sites in AHRR for tobacco smoking. Chr.: chromosome *: biotinylated primer Amp.: Amplicon

Locus		Sequence	Chr./ Gene ID	Amp. size	CpG sites analyzed (**bold**) (sequence to analyze)
AHRR	Forward	GGGAGTGGTTTTGGTAGG	5/57491	169	TTG**Y**GGGATTAGTAGGT**Y**GGG**Y**GGTGGTTGGGAGGTAG GGGATAGGTTGGTTTTTG**Y**GGTTTTGGAGGTATAGAGTTTTTTT
	Reverse*	ACCCTTACAAAACACAACTAAAC			
	Sequencing	AGGGTTTTTTTTTGTAGATT			

Supplementary Figures 3.S (1A-1C): Pyrogram results generated from Q24 instrument shows the methylation levels for 1A: current, 1B: former and 1C: never smokers







7. References

[1] Bird, A. Genes Dev. 2002, 16, 6-21.

[2] Fraga, M. F., Ballestar, E., Paz, M. F., Ropero, S. Setian, F.,Ballestar, M. L., Heine-Suner, D., Cigudosa, J. C., Urioste,M., Benitez, J., Boix-Chornet, M., Proc. Natl. Acad. Sci. U. S. A. 2005, 102, 10604-10609.

[3] Martin, G. M. Proc. Natl. Acad. Sci. U. S. A. 2005, 102, 10413-10414.

[4] Breitling, L. P., Yang, R., Korn, B., Burwinkel, B., Brenner, H., Am. J. of Hum. Genet. 2011, 88, 450-457.

[5] Lee, K. W., Pausova, Z. Front. Genet. 2013, 4, 132.

[6] Cuozzo, C., Porcellini, A., Angrisano, T., Morano, A. Lee, B., Di Pardo, A., Messina, S., Iuliano, R., Fusco, A., Santillo, M.R., Muller, M. T., PLoS Genetics 2007, 3, e110.

[7] Lee, E. W., D'Alonzo, G. E. Arch. Intern. Med. 1993, 153, 34-48.

[8] Satta, R., Maloku, E., Zhubi, A., Pibiri, F. Hajos, M., Costa, E., Guidotti, A., Proc. Natl. Acad. Sci. U. S. A. 2008, 105, 16356-16361.

[9] Kadonaga, J. T., Carner, K. R., Masiarz, F. R., Tjian, R. Cell 1987, 51, 1079-1090.

[10] Han, L., Lin, I. G., Hsieh, C. L. Mol. Cell. Biol. 2001, 21, 3416-3424.

[11] Liu, Q., Liu, L., Zhao, Y., Zhang, J. Wang, D., Chen, J., He, Y., Wu, J., Zhang, Z., Mol. Cancer. Ther. 2011, 10, 1113-1123.

[12] Monick, M. M., Beach, S. R., Plume, J., Sears, R. Gerrard, M., Brody, G. H., Philibert, R. A., Am. J. Med.l Genet. B Neuropsychiatr. Genet. 2012, 159, 141-151.

[13] Zeilinger, S., Kühnel, B., Klopp, N., Baurecht, H. Kleinschmidt, A., Gieger, C., Weidinger, S., Lattka, E., Adamski, J., Peters, A., Strauch, K., PloS one 2013, 8, e63812.

[14] Shenker, N. S., Polidoro, S., van Veldhoven, K., Sacerdote, C. Ricceri, F., Birrell, M. A., Belvisi, M. G., Brown, R.,Vineis, P., Flanaga J. M., Hum. Mol. Genet. 2012, 22, 843-851.

[15] Joubert, B. R., Haberg, S. E., Nilsen, R. M., Wang, X. Vollset, S. E., Murphy, S. K., Huang, Z., Hoyo, C., Midttun Ø., Cupul-Uicab, L. A., Ueland, P. M., Environ. Health Perspect. 2012, 120, 1425-1431.

[16] Dogan, M. V., Shields, B., Cutrona, C., Gao, L. Gibbons, F. X., Simons, R., Monick, M., Brody, G. H., Tan, K., Beach, S. R., Philibert, R. A., BMC Genomics 2014, 15, 151.

[17] Bauer, M., Fink, B., Thürmann, L., Eszlinger, M. Herberth, G., Lehmann, I., Clin. epigenetics 2016, 8, 83.

[18] Zhang, Y., Elgizouli, M., Schöttker, B., Holleczek, B. Nieters, A., Brenner, H., Clin. epigenetics 2016, 8, 127.

[19] Nyrén, P., Pettersson, B., Uhlén, M. Anal. Biochem. 1993, 208, 171-175.

[20] Comey, C. T., Koons, B. W., Presley, K. W., Smerick, J. B. Sobieralski, C. A., Stanley, D. M., Baechtel, F. S., J. Forensic Sci. 1994, 39, 1254-1269.

[21] Nicklas, J. A., Buel, E., J. Forensic Sci. 2003, 48, 936-944.

[22] Alghanim, H., Antunes, J., Silva, Deborah Soares Bispo Santos, Alho, C. S. Balamurugan, K., McCord, B., Forensic Sci. Int.: Genet. 2017, 31, 81-88.

[23] Benjamini, Y., Hochberg, Y., J. R. Stat. Soc. Series B Stat. Methodol. 1995, 289-300.

[24] Elliott, H. R., Tillin, T., McArdle, W. L., Ho, K., Duggirala, A., Frayling, T. M., Smith, G. D., Hughes, A. D., Chaturvedi, N., Relton, C. L., Clin. Epigenetics 2014, 6,4.

[25] Andersen, A. M., Dogan, M. V., Beach, S. R., Philibert, R. A., Genes 2015, 6, 991-1022.

[26] Evans, B. R., Karchner, S. I., Allan, L. L., Pollenz, R. S. Tanguay, R. L., Jenny, M. J., Sherr, D. H., Hahn, M. E., Mol. Pharmacol. 2008, 73, 387-398.

[27] Madi, T., Balamurugan, K., Bombardi, R., Duncan, G., McCord, B., Electrophoresis 2012, 33, 1736-1745.

[28] Soares Bispo Santos Silva, Deborah, Antunes, J., Balamurugan, K., Duncan, G. McCord, B., Electrophoresis 2015, 36, 1775-1780.

[29] Wan, E. S., Qiu, W., Baccarelli, A., Carey, V. J. Bacherman, H., Rennard, S. I., Agusti, A., Anderson, W., Lomas, D. A., DeMeo, D. L., Hum. Mol. Genet. 2012, 21, 3073-3082.

[30] Philibert, R., Hollenbeck, N., Andersen, E., Osborn, T. Gerrard, M., Gibbons, F. X., Wang, K., Front. Psychol. 2015, 6, 656.

[31] Shenker, N. S., Ueland, P. M., Polidoro, S., van Veldhoven, K. Ricceri, F., Brown, R., Flanagan, J. M., Vineis, P., Epidemiology 2013, 24, 712-716.

[32] Tsaprouni, L. G., Yang, T., Bell, J., Dick, K. J. Kanoni, S., Nisbet, J., Vinuela, A., Grundberg, E., Nelson, C. P., Meduri, E., Buil, A., Epigenetics 2014, 9, 1382-1396.

[33] Endo, K., Li, J., Nakanishi, M., Asada, T. Ikesue, M., Goto, Y., Fukushima, Y., Iwai, N., BioMed Res. Int. 2015, 2015, 10.

Chapter VI: Evaluation of DNA methylation markers for sperm and blood identification through pyrosequencing and high-resolution melt analysis

1. Abstract

Discrimination of body fluids can bring useful information about the crime scenes. Recent advances in epigenome research demonstrate that tissue specific differentially methylated regions (tDMRs) provide different DNA methylation patterns indicative for a certain type of tissue making it possible to identify body fluid. In particular, two DNA methylation based methods are gaining popularity and can be well suited in the forensic casework: pyrosequencing and high resolution melt (HRM) analysis. The goal of this report is to identify new sets of tDMRs and based on those loci develop assays that could be utilized for forensic discrimination of body fluids. Two markers NMUR2 and UBE2U were found to be specific for sperm in which the mean DNA methylation level for sperm samples was significantly different from the means DNA methylation of the other body fluids tested. The assays developed using NMUR2 and UBE2U can be employed using both pyrosequencing or HRM analysis. The procedure for HRM is a quick and cost effective technique to identify seminal stains recovered from the crime scene using instrumentation that is familiar to the forensic analyst. Finally, markers in the AHRR gene were found be effective to distinguish blood from other tissues. The AHRR assay was designed to include four CpG sites and was found to be only effective through pyrosequencing. Using this assay, blood stains showed significantly different methylation levels when compared to other body fluids.

2. Introduction

Body fluids are biological materials that are either secreted or excreted from the body. Human body fluids such as blood, saliva, semen or sweat are commonly encountered at crime scenes. Determination of the type and origin of body fluids can provide valuable insights regarding the circumstances leading to the deposition of the DNA evidence. The type of sample from which DNA evidence originate can help to distinguish between a suspect or a victim and identify the type of crime being committed (i.e., physical vs. sexual assault). Body fluids are also very useful in crime scene reconstruction as they can help recreate the events of the crime by the indicating the presence or absence of specific tissue types (Lee et al., 2012). Currently, body fluid identification relies on conventional serological or immunological testing methods that are based on colorimetric detection of protein based markers. Because of the unspecific presence of many of these markers, in different body fluids, several of the current methods in use are presumptive in nature and thus can't confirm the presence of all body fluids (Butler et al., 2010). In addition, some of the methods may require the use of a large sample size and considered destructive, which is impractical for forensic specimens (Virkler & Lendnev et al., 2009). Thus, developing a confirmatory test for discriminating body fluids would be of great benefit to the forensic community. Recent advances in forensic genetics have demonstrated that DNA methylation is a promising new tool that can overcome many limitations of traditional serological methods. In mammals, DNA methylation occurs at the 5' carbon of cytosine residues in certain CpG dinucleotides (Mirnada et al., 2007). In humans, the

unmethylated CpGs are grouped in base sequences known as 'CpG islands' located at the 5' ends of many genes (Bird et al., 2002, Voet et al., 2011, Frumkin et al., 2011). In general, DNA methylation is an epigenetic modification that is believed to inhibit the expression of genes by affecting the chromatin structure (Hamshimary et al., 2003, Frumkin et al., 2011). DNA methylation plays a vital role in tissue development and cellular differentiation (Ohgane et al., 2008, Lee et al., 2012). After DNA methylation patterns are established during gestation, unique cell- and tissue-specific DNA methylation signatures appear (Kitamura et al.,2007, Igarashi et al., 2008, Song et al., 2009). Various epigenetic whole genome studies of DNA methylation have been used to detect ` the presence of these tissue-specific differentially methylated regions (tDMRs) (Song et al., 2005, Shen et al., 2007). In forensics, Frumkin et al. 2011 was the first to investigate the possibility of utilizing DNA methylation markers for forensic tissue identification. The authors tested several forensic tissues including blood, saliva, semen, skin, urine, mensural blood and vaginal fluids and could identify methylation regions that showed differential methylation patterns. A total of 205 CpG islands were screened from which a panel of 15 loci were selected that show significant differential methylation levels between forensically relevant tissues forming the foundation for DNA-based technique for body fluid identification (Frumkin et al., 2011).

Technologies for DNA methylation analysis include DNA digestion by methylation sensitive restriction enzymes, immunological detection of 5-methylcytosine, and bisulfite conversion of the unmethylated cytosine residues (Madi et al., 2012,

Vidaki et al., 2013). Among these methods, bisulfite treatment is the most common. In this procedure unmethylated cytosines are chemically converted to uracil base whereas the methylated cytosines remain unchanged. Next, the modified DNA is PCR amplified using specially designed primers. The resulting PCR products can be detected via pyrosequencing, High Resolution Melt (HRM) analysis and MALDI-TOF mass spectrometry (Gut et al., 2004, Witter et al., 2009, Madi et al., 2012). Pyrosequencing is a sequencing-by-synthesis technique that monitor the nucleotide addition and sequence extension in real time. After bisulfite treatment, the target sequence is PCR amplified using a primer pair from which one of the pair is biotin-labeled at the 5' end (either forward or reverse). Next, a sequencing primer is added which uses the biotinylated amplicon as a template to undergo the four enzymatic reactions of pyrosequencing (Madi et al., 2012).

A method for methylation detection utilizing real time PCR instrument with HRM capability can also be used. HRM analysis can differentiate between methylated and unmethylated templates, that differ in GC content, based on their melting temperature (Tm) (Witter et al., 2009, Antunes et al., 2016). The method involves the amplification of the bisulfite-modified DNA template using real-time PCR in the presence of a double-stranded DNA intercalating dye such as SYBR Green. As the template concentration increases in the reaction mixture, the fluorescence intensity exhibited by the dsDNA increases. The melt analysis is a post PCR process that immediately follows the end of the PCR and is used to detect the difference in Tm between the methylated and unmethylated DNA templates.

Several studies have explored the application of pyrosequencing to detect tDMR for body fluid identification (Madi et al., 2012, Park et al., 2014, Alghanim et al., 2017). Other studies rely on HRM analysis as a simple and cost-effective technique that can detect differentially methylated DNA templates generated by tDMR for body fluids ID (Hanson et al., 2013, Antunes et al., 2016). The results indicate that a panel of body-specific CpG Sites could become useful DNA markers for body fluid discrimination in forensics. The aim of this study is to develop a new set of CpG sites that can be used to distinguish some forensically relevant tissues types particularly sperm and blood which can contribute to a universal panel needed for body fluid identification of all type of cells. In this study, both pyrosequencing and HRM analysis are utilized to detect methylation profiles of the biomarker developed in order to provide a way for comparison of the two methods. The results permit a comparison of these methods for use as tools in the forensic identification of body fluids.

3. Material and methods

3.1 Sample collection

Forensically relevant body fluid samples were collected from unrelated volunteers living in Dubai and Miami. In total, 23 venous blood, 24 saliva (buccal swab), 22 vaginal secretion and 20 sperm samples were collected. Freshly ejaculated semen fluid containing sperms was collected in a plastic cup and then put onto sterile cotton swabs to dry. Venous blood, saliva and vaginal secretions were collected directly on sterile cotton swabs and allowed to dry at room temperature. All biological samples were acquired from volunteers using procedure's approved

by the Institutional Review Board at Florida International University under IRB-16-0341 and General Headquarters of Dubai Police (approval letters #410126/11/33/3583). All participants signed and gave their informed consent forms in writing after the aims and protocols of the study were explained.

3.2 Screening strategies

In the discovery step, blood, saliva, sperm and vaginal secretion (n=3 per sample type) were examined to screen candidates of CpG sites located in 11 genomic loci in order to identify their tDMRs. In this step, candidates of CpG sites were examined by pyrosequencing. The CpG sites that showed differences in their methylated profiles between various body fluids were selected for further DNA methylation analysis by pyrosequencing and HRM analysis. In this study, only three tissue-specific differentially methylated regions (tDMRs) were discovered including locations at NMUR2, UBE2U, and AHRR.

3.3 DNA extraction and bisulfite conversion

Dried swabs containing various body fluid samples were DNA extracted by either the EZ1® DNA Investigator Kit on the BioRobot® EZ1 automated purification workstation (Qiagen) or by standard organic extraction using phenol-chloroform-isoamyl alcohol (Thermo Fisher Scientific) (Comey et al, 1994). The extracted DNA was bisulfite-modified using the EpiTect® Fast DNA Bisulfite Kit (Qiagen), which can modify 1 ng–2 µg of DNA, to convert the unmethylated cytosine to uracil.

3.4 Pyrosequencing

To determine the methylation status of potential markers in various body fluid samples, sequencing was carried out by pyrosequencing. First, specific PCR primers and sequencing primer were designed using PyroMark Assay Design 2.0 software (Qiagen Inc. CA) to amplify the bisulfite modified target regions. After screening more than 100 CpG sites across 11 genetic loci, three loci were found to be tissue specific. The three assays located at NMUR2, UBE2U, and AHRR genes were designed to target different numbers of CpG sites (Table 4.1). The specific locations were next amplified with one member of each PCR primer pair labeled with biotin to produce biotinylated PCR amplicons needed for the downstream pyrosequencing reaction. The target regions were amplified in a singleplex fashion by utilizing the PyroMark® PCR kit (Qiagen) on the GeneAmp® PCR system 9700 (Applied Biosystems, Foster City, CA). The PCR reaction was modified to utilize 15μl reaction volumes based on the total volume specified by the manufacturer's protocol [22]. The PCR products were pyrosequenced using a Pyromark Q24 pyrosequencer (Qiagen) as per the manufacturer's instructions. Pyromark® Q24 software (Qiagen) was used to calculate the percent methylation for each CpG site. The PCR products were pyrosequenced using a Pyromark Q24 pyrosequencer (Qiagen) as per the manufacturer's instructions. Pyromark® Q24 software (Qiagen) was used to calculate the percent methylation for each CpG site. The results were displayed as a pyrogram with the methylation percentage.

Locus		Sequence	Chr.	CpG no./ Ampl. size	CpG sites analyzed (**bold**)
NMUR2	Forward	GTGTTGGGTAGGGAGAAGAGTA	5	7/277	G**CGCG**GAA**CG**GGTGTAGGATGGTTA**CG**TAGC**CG**TTTTA**CG**TTGA**CG**GTGGTGATGTTGAGGATGGAGG
	Reverse*	CTAACCTCCTAATCCTACTCCTTAAA			
	Sequencing	GGGTGTTTTGTAGTTTG			
UBE2U	Forward*	GTTTTGAGATTGGGTTGTGAG	1	3/207	**CG**GTATTGTAGTGAAA**CG**T**CG**TAGATGAGGAAGTGTTTAA GTTTT
	Reverse	CACTTTCCCACACTTAATAAACTAATA			
	Sequencing	GATTGGGTTGTGAGT			
AHRR	Forward	TGGGGTTTTAAGGTTAGGGTG	5	4/233	**CG**AG**CG**TGTGATTTTGGTGAT**CG**TAGAGTTTTTTTGAGGTTTT**CG**GGTTTTGTGATTTTAGAAAGTGGT
	Reverse*	AATTTCACACTTCCTCACAATACA			
	Sequencing	GGTGTGTTTTTTTTGTAGGA			

Table 4.1: Assays designed to evaluate CpG sites in three different genetic loci. Chr. : chromosome *: biotinylated primer no: numbers Ampl.: amplicon

3.5 HRM Analysis

HRM is a real-time PCR method that utilizes an unlabeled primer pair for amplification and includes an intercalating dye for amplification detection and melt analysis. The samples were PCR amplified using Rotor-gene SYBER Green kit (Qiagen) on a Rotor Gene 6000 real-time instrument (Qiagen). The kit composed of a buffer that consists of SYBER Green I, HotStarTaq plus, and dNTP mix. The amplification reaction was adjusted to 20 μl based on the total volume specified by the manufacturer. PCR amplifications were performed by adding 1μL of bisulfite modified DNA to a master mix that consisted of 2X Rotor-Gene SYBER Green

PCR master mix and 1uM of each unlabeled forward and reverse primers. The amplifying HRM primers utilized were unlabeled and had the same sequence as that used for pyrosequencing analysis. Custom designed primers were obtained from Integrated DNA Technologies (IDT), Inc. (Coralville, IA). PCR cycling was conducted on the GeneAmp® PCR system 9700 (Applied Biosystems, Foster City, CA) under the following conditions: 95°C for 5 min; 45 cycles of 95°C for 10 s, 59°C for 16 s, and 72 °C for 10 s. Immediately afterward, melt analysis was performed by increasing the temperature from 65 to 95 °C in 0.3 °C increments and detecting fluorescence in the HRM channel. Melt curve analysis was generated and melting temperatures were determined using the Rotor-Gene 6000 series software (version 1.7).

3.6 Statistical analysis

Statistical analysis was performed on the methylation profiles generated as percent methylation values from pyrosequencing or as melting points from HRM analysis. The average percent methylation value for each CpG site identified was calculated along with the standard deviations for each cell type. A one-way ANOVA with Tukey's test and Wetch test were carried out to determine if there were statistical difference in percent methylation level between the four types of body fluids tested. Methylation differences were considered statistically significant if *p*-values were 0.05 or less. All the analyses were performed using SPSS statistics software ver. 23.

4. Results

In the discovery step, a set of 12 body fluid samples including blood, saliva, sperm and vaginal secretions were used for preliminary evaluation of body fluid identification using 20 different probe methylation sites. Pyrosequencing was used to perform the methylation analysis for the preliminary evaluation at the 20 probe sites located at 11 genetic loci. Three genomic locations were found that contained differentially methylated regions that were tissue specific of which two were found to be markers for sperm and one for blood. The three body fluid markers were further screened using pyrosequencing and HRM analysis.

4.1 Pyrosequencing data

A total of 89 samples were used to further examine the three differentially methylated regions identified including 23 blood, 24 saliva, 20 semen fluid containing sperms, and 22 vaginal secretions. The methylation profiles were successfully analyzed using bisulfite conversion and pyrosequencing. For loci NMUR2 and UBE2U, semen containing sperms presented low percent methylation (less than 20%) whereas the other body fluids had high levels of methylation (more than 80%) (Figures 4.1 and 4.2). Seven CpG sites in NMUR2 showed statistically significant differences between the body fluids and could distinguish sperm from other body fluids (Table 4.2). The assay located at UBE2U consisted of 3 CpG sites showing sperm-specific methylation profiles (Figure 4.2) and showed significant difference in methylation level between sperm and the other body fluids (Table 4.2). Three seminal samples without sperms produced hypermethylation patterns when tested using NMUR2 and UBE2U. These results confirm that the

two assay are sperm specific and can be very effective tools for forensic identification of sperm in crime scene.

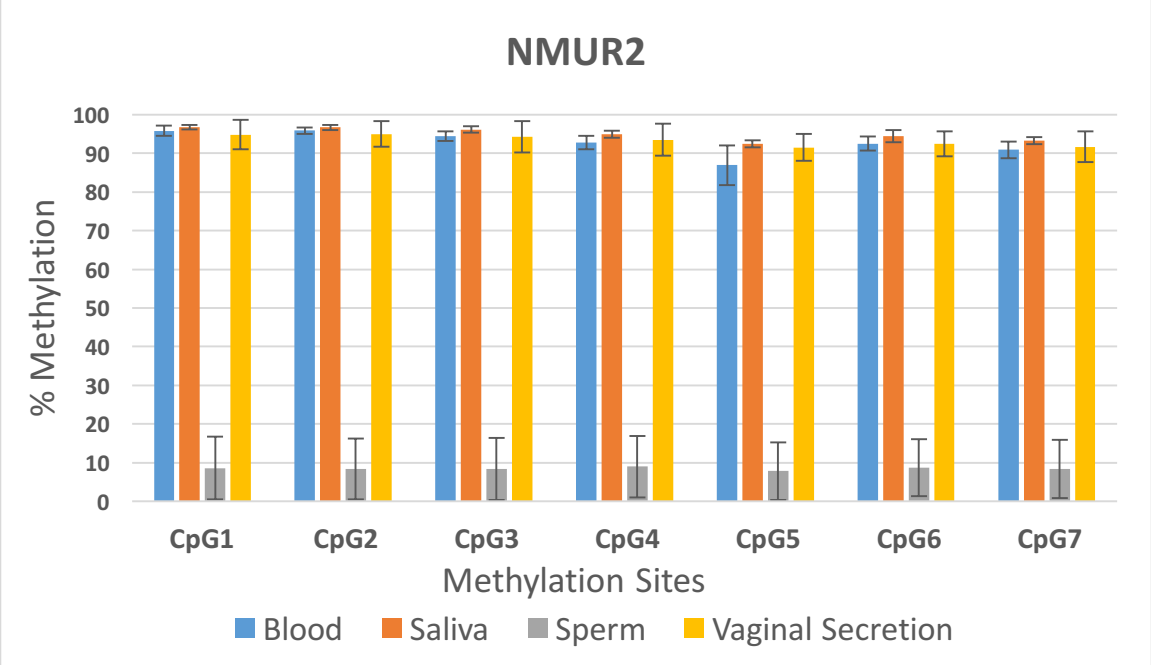


Figure 4.1: chart showing the mean percent of methylation on locus NMUR2 determined by pyrosequencing for samples of blood (n=23), saliva (n=24), sperm (n=20), and vaginal secretion (n=22), +/- standard deviation of the mean.

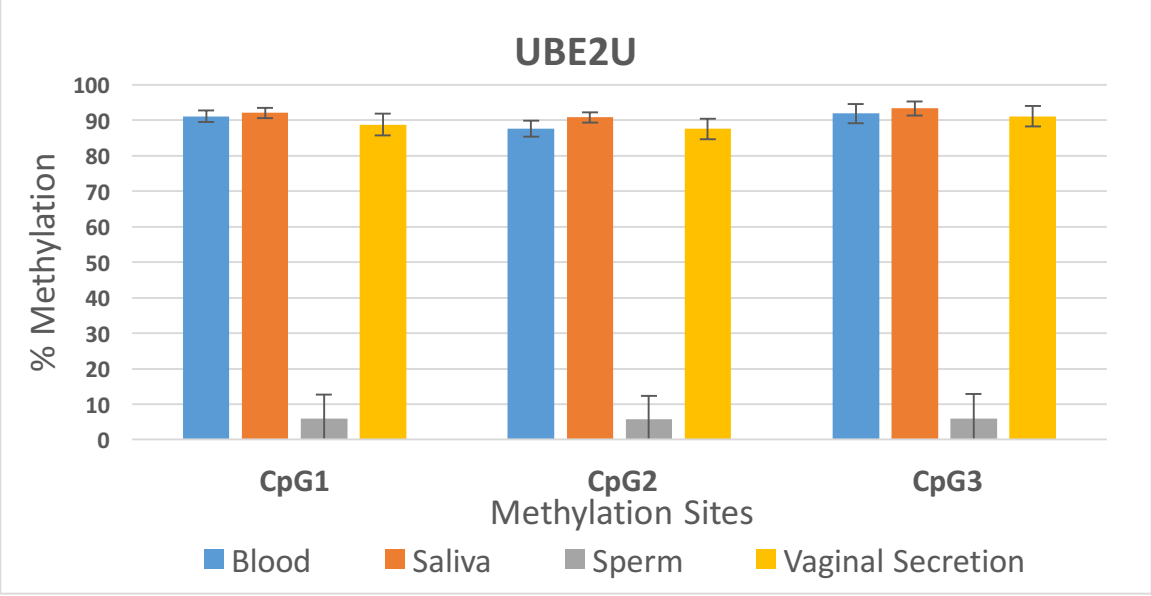


Figure 4.2: chart showing the mean percent of methylation on locus UBE2U determined by pyrosequencing for samples of blood (n=23), saliva (n=24), sperm (n=20), and vaginal secretion (n=22), +/- standard deviation of the mean.

Marker	Body Fluid	CpG (% mean methylation ± standard deviation)							HRM Tm (°C)
NMUR2 Sperm Specific		CpG1	CpG2	CpG3	CpG4	CpG5	CpG6	CpG7	All
	Sperm	8.6 ±8.1	8.4 ±7.8	8.4 ±8.1	9.0 ±8.0	7.8 ±7.5	8.6 ±7.3	8.4 ±7.6	80.9 ±0.1
	Vaginal Secretion	94.8 ±3.8	95.0 ±3.2	94.3 ±4.0	93.5 ±4.1	91.5 ±3.5	92.5 ±3.2	91.6 ±4.0	84.6 ±0.1
	Saliva	96.7 ±0.5	96.7 ±0.6	96.1 ±0.9	95.0 ±1.0	92.4 ±0.9	94.4 ±1.5	93.2 ±0.9	84.5 ±0.2
	Blood	95.8 ±1.3	95.9 ±1.3	94.4 ±0.9	92.8 ±1.3	87.0 ±1.7	92.5 ±5.1	90.9 ±2.2	84.5 ±0.2
	p-value	1.6×10^{-79}	5.5×10^{-82}	5.6×10^{-79}	1.0×10^{-77}	3.2×10^{-74}	3.2×10^{-81}	6.1×10^{-75}	2.0×10^{-83}
	p-value	1.1×10^{-33}	2.7×10^{-36}	2.1×10^{-35}	1.2×10^{-34}	1.6×10^{-33}	5.4×10^{-38}	5.6×10^{-33}	3.2×10^{-53}
UBE2U Sperm Specific		CpG1		CpG2		CpG3			All
	Sperm	5.8 ±6.8		5.8 ±6.6		6.0 ±7.0			77.1 ±0.1
	Vaginal Secretion	88.8 ±3.1		87.5 ±2.9		91.1 ±2.9			78.8 ±0.1
	Saliva	92.1 ±1.5		90.8 ±1.5		93.3 ±2.0			78.8 ±0.1
	Blood	91.1 ±1.3		87.6 ±2.2		91.9 ±2.7			78.7 ±0.1
	p-value	9.2×10^{-84}		2.1×10^{-83}		4.6×10^{-82}			1.6×10^{-77}
	p-value	1.6×10^{-39}		2.2×10^{-39}		7.8×10^{-40}			4.3×10^{-47}
AHRR Blood Specific		CpG1		CpG2		CpG3		CpG4	
	Sperm	84.7 ±11.8		88.9 ±9.8		93.1 ±9.8		92.7 ±12.4	
	Vaginal Secretion	48.2 ±20.6		57.3 ±20.1		62.5 ±18.8		42.9 ±17.1	
	Saliva	63.0 ±13.9		69.0 ±13.6		73.7 ±14.2		55.0 ±12.9	
	Blood	6.5 ±1.9		14.8 ±4.2		18.1 ±4.0		4.7 ±1.6	
	p-value	2.2×10^{-31}		5.5×10^{-31}		3.1×10^{-32}		4.4×10^{-37}	
	p-value	1.2×10^{-26}		1.4×10^{-28}		4.3×10^{-29}		3.0×10^{-27}	

Table 4.2. Mean % methylation for the pyrosequencing based assays and mean Tm for the HRM analysis with the significance values based on ANOVA (p-value) and Wetch test (*p-value*).

On the other hand, the assay of four CpGs sites at AHHR could also be used as a biomarker for blood identification. At the AHRR gene, blood is hypomethylated compared to other body fluid types which have low levels of methylation (Figure 4.3).

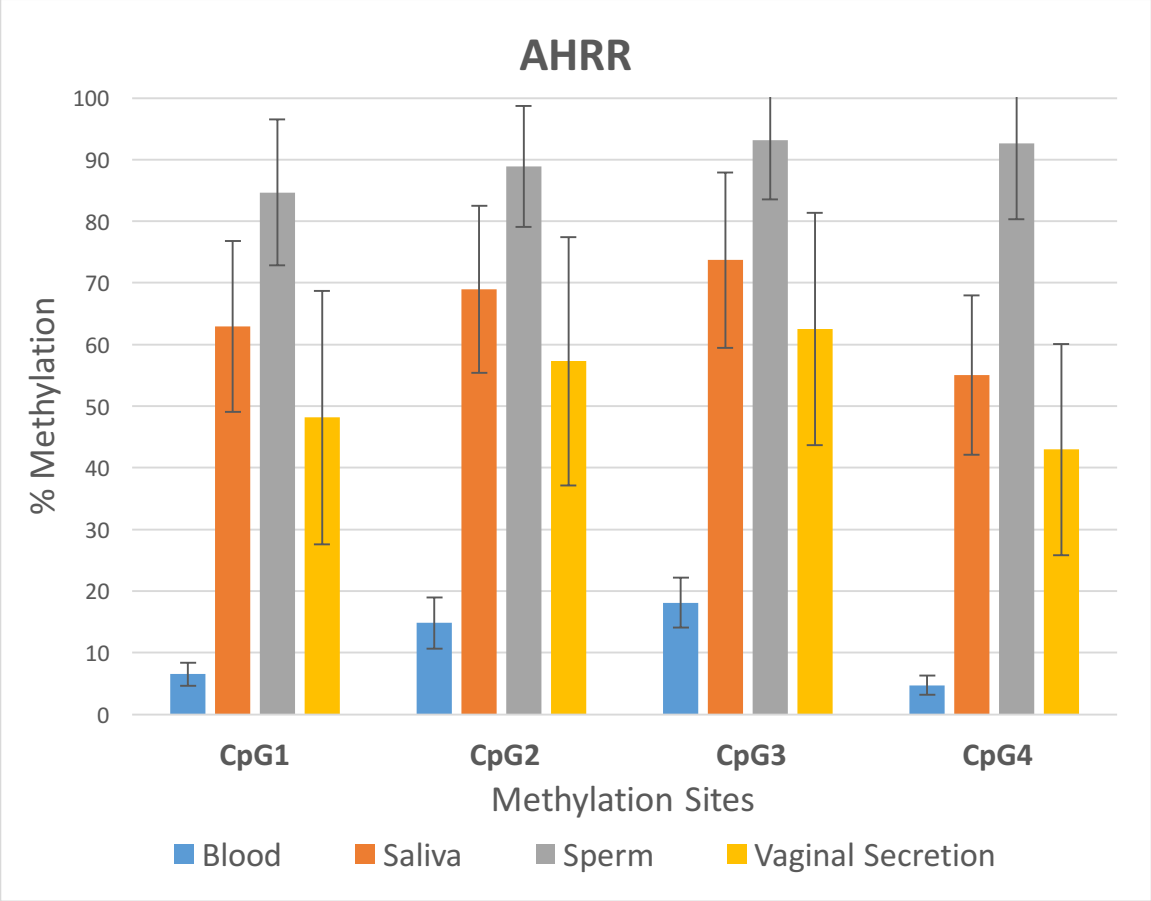


Figure 4.3: chart showing the mean percent of methylation on locus AHRR determined by pyrosequencing for samples of blood (n=23), saliva (n=24), sperm (n=20), and vaginal secretion (n=22), +/- standard deviation of the mean.

4.2 HRM data

The same set of samples consisting of 80 different body fluid types were also used to determine whether HRM analysis would also be suited to discriminate the methylation profiles of sperm based on CpG markers at NMUR2 and UBE2U. The melt curve representing the derivative slope of fluoresce (-df/dT) over temperature for the NMUR2 marker showed a distinct pattern for sperm samples when compared to other tissues. The melt curve showing the data for differing tissue types at the NMUR2 marker is shown in Figures 4.4 and 4.5. At the NMUR2, the sperm samples (n= 22) had a lower melting temperature averaging (80.9 °C) than those detected for the other tissue types including blood (84.5 °C), saliva (84.5 °C) and vaginal sections (Tm of 84.6 °C) (Figures 4.4 and 4.5).

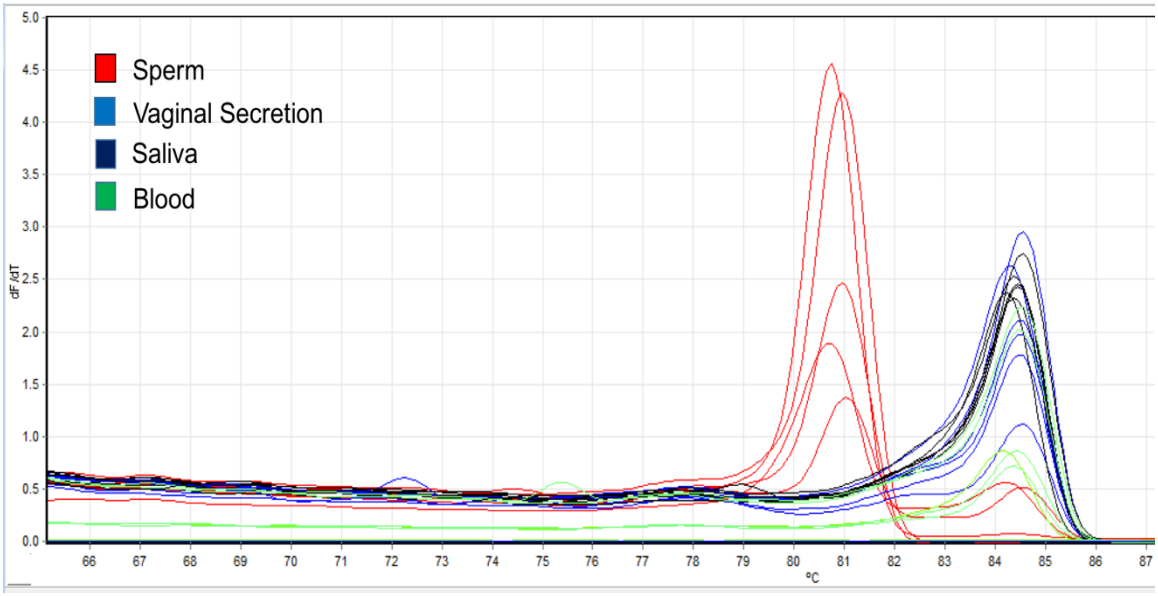


Figure 4.4: Melt curves from samples amplified and analyzed with NMUR2 marker showing melting temperatures for sperm (n=5) is lower than those of other body fluids (vaginal secretion n=6, saliva n=6, and blood n=4).

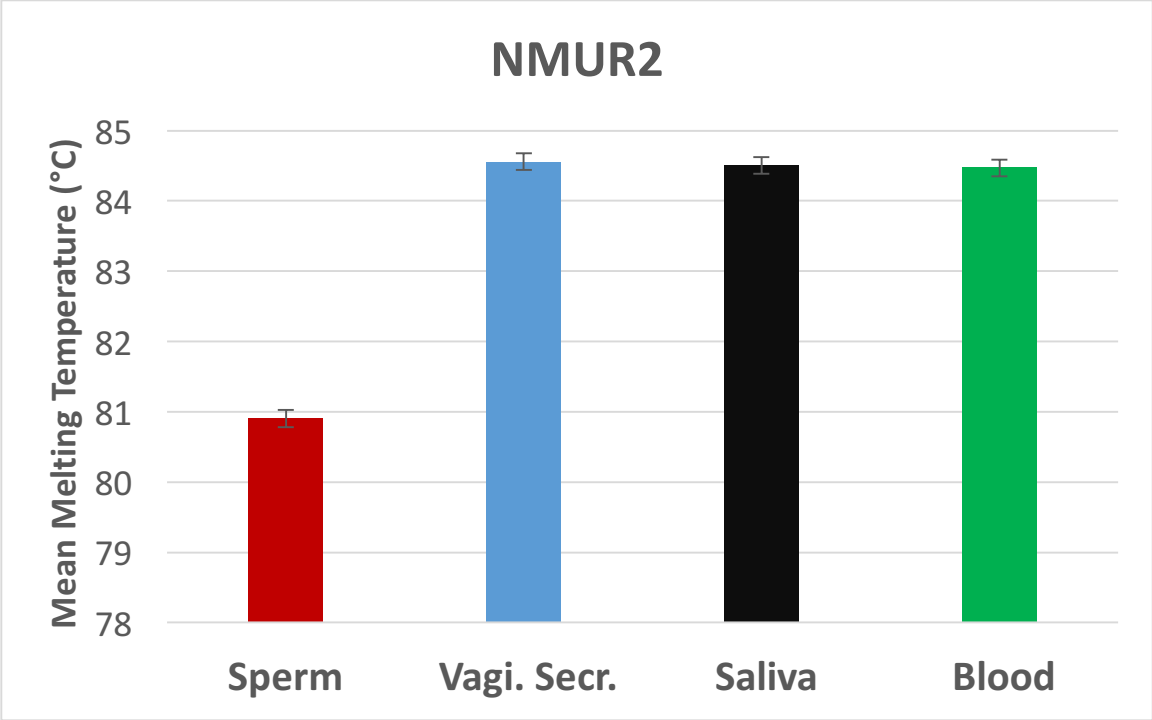


Figure 4.5: Graph showing the mean values for melting temperatures (°C) for NMUR2 marker obtained by HRM analysis for samples of sperm (Tm= 80.9 °C), vaginal secretion (84.6 °C), saliva (84.5°C) and blood (84.5°C), +/- standard deviation.

In the same way, the melting curve for the UBE2U marker also showed distinct pattern of sperm samples compared to other body fluids as illustrated in Figures 4.6 and 4.7. The melting temperature of sperm samples is lower than the melting temperatures of the other three body fluids. For the UBE2U, the sperm samples (n=22) Tm averaged is 77.1 °C while the other tissue types had higher average melting temperature for blood (Tm=78.7 °C, n=20), for saliva (Tm=78.8 °C, n=21) and for vaginal secretion (Tm=78.7 °C, n=20) (Figure 4.7). However, all four body fluids showed very similar and overlapping melting curves and no distinct pattern can be identified when using AHRR marker.

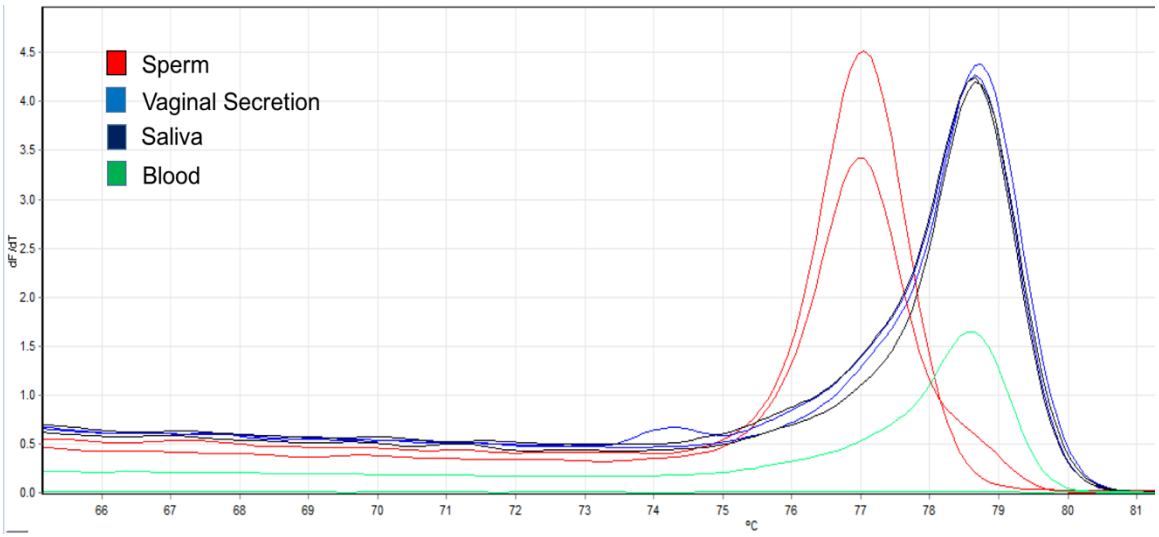


Figure 4.6: Melt curves from samples amplified and analyzed with UBE2U marker showing melting temperatures for sperm (n=2) is lower than those of other body fluids (vaginal secretion n=2, saliva n=2, and blood n=1).

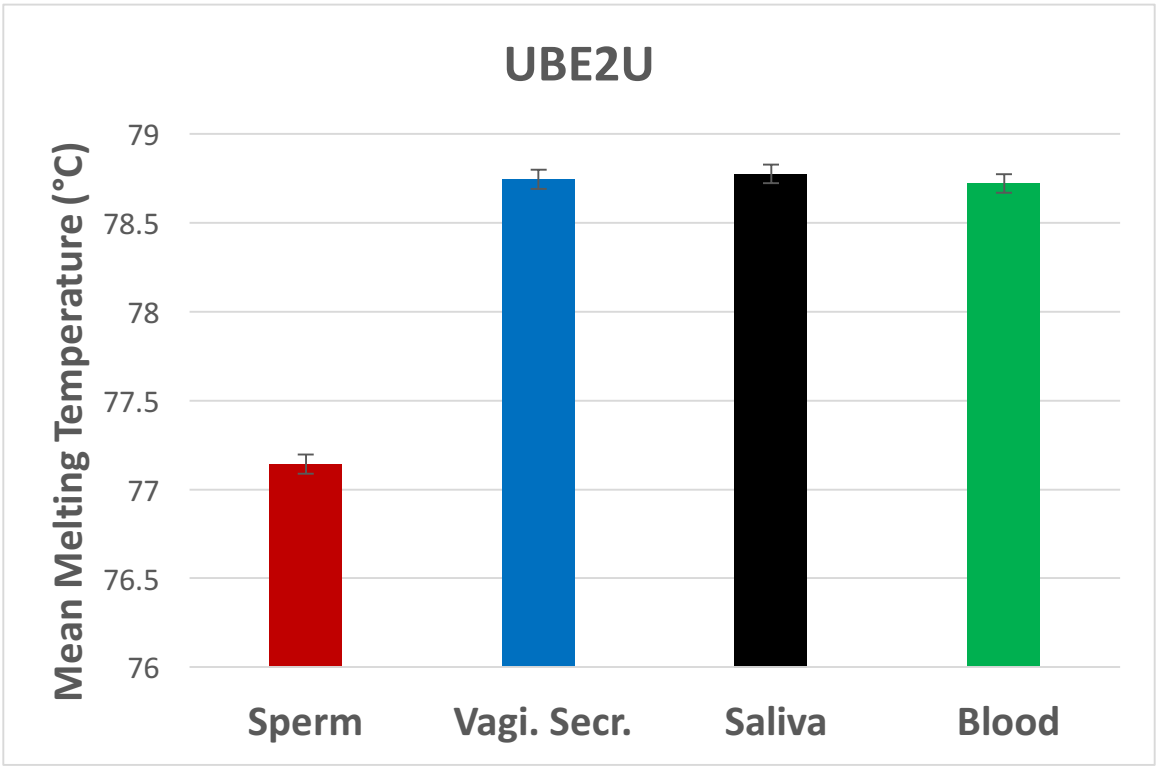


Figure 4.7: Graph showing the mean values for melting temperatures (°C) for UBE2U marker obtained by HRM analysis for samples of sperm (n=22, Tm=77.1 °C), vaginal secretion (n=20, Tm=78.7 °C), saliva (n=21, Tm=78.8 °C) and blood (n=20, Tm=78.7 °C), +/- standard deviation.

5. Discussion

The determination of the tissue source of a DNA sample is important information that can aid in crime scene reconstruction, a vital aspect of forensic science casework. For example, DNA derived from blood may indicate a case of physical assault and DNA originating from sperm sample may derived from case of a sexual contact. Current practice in forensic science for the detection and identification of various body fluids relies on conventional immunological or biochemical methods. These methods are generally quick and easy but they are generally destructive to the sample. These tests also require large amounts of high quality evidence and many of these tests are presumptive with the potential to produce false positive or false negative results (An et al., 2013). Blood is the most common body fluid found in crime scenes. Most enzymatic and chemical presumptive tests for blood such as Kastle-Mayer (KM) and the Leucomalachite Green (LGM) depend on the ability of the hemoglobin found in red blood cells to catalyze the oxidation of certain reagents in the presence of oxidizing agents causing the reagents to change color (Webb et al., 2006, Virkler & Lendnev et al., 2009). For semen, the most common and widely utilized presumptive test is the acid phosphatase test based on the enzymatic ability to catalyze the hydrolysis of phosphates resulting in change in solution (Virkler and Lendnev et al., 2009). The microscopic identification of sperm cells is a confirmatory test for the presence of semen although an azospermic person would produce to false negative. Although these serological technologies for tissue identification are still currently used in the crime labs, newer technologies

are on the rise to overcome some of disadvantages. One of the most investigated techniques for tissue identification is based on DNA methylation.

Methylation is an epigenetic modification that has a role in gene expression. Recently, several studies have reported the use of DNA methylation markers that are tissue specific (Madi et al. 2012, Park et al., 2014, Antunes et al., 2016). These reports identified certain sites (tDMR) that are differentially methylated between different tissue types such that they are hypermethylated in a specific tissue and hypomethylated on the others or vice versa. The tDMR have been proven to be a very useful tool for body fluid identification and both pyrosequencing and HRM analysis have been used in its analysis. In this study, the goal was to identify new tDMRs that can be used as biomarkers for forensic body fluid discrimination. Using the discovered tDMR, first we developed DNA methylation assays based on pyrosequencing. HRM was also utilized as another DNA methylation based technology for body fluid identification.

From the discovery step, two new epigenetic loci NMUR2 and UBE2U were found to differentiate sperm from other body fluids (blood, saliva and vaginal secretion) while AHRR was able to distinguished blood from other three tissues using the analysis of methylation signatures of each tissue. Neuromedin U receptor 2 (NMUR2) is a gene that encodes a protein from G-protein coupled receptor 1 family (Shan et al., 2000). Such protein serve as a receptor for neuromedin U which is a neuropeptide that is extensively distributed in the gut and central nervous system (Shan et al., 2000). This receptor has an important role in regulating food intake and body weight. UBE2U is a protein coding gene for

Ubiquitin Conjugating Enzyme E2 U (Guo et al., 2017). This protein can catalyze the covalent attachment of ubiquitin to other proteins. The proteins encode for NMUR2 and UBE2U seem to be very important for sperm fluid and its components and not so for the other body fluids tested. AHRR is a gene that encode for Aryl-Hydrocarbon Receptor Repressor. This protein involves in mediating detoxification of harmful substances such as the toxin involve in tobacco smoking. Previously, we identified some CpG sites at this gene that can be a great indicator for the smoking habits (Alghanim et al., 2018). In addition, AHRR is associated with regulation of cell growth and differentiation (Haarmann-Stemmann et al., 2007) the function that correlate with our new finding of methylation sites that can serve as a blood biomarker.

The results of testing three seminal fluids from victimized male indicate that NMUR2 and the UBE2U are sperm specific. The two sperm markers can effectively discriminate sperm markers from other body fluids using pyrosequencing (Figures 4.1 and 4.2). The NMUR2 assay consists of 7 CpG sites all of which showed hypomethylated levels for sperm while being hypermethylated in other tissues (Figure 4.1). The UBE2U assay contains 3 CpG sites that show low levels of methylation for sperm compared to the other three tissues (Figure 4.2). The AHRR blood marker consists of 4 CpG sites which are hypomethylated for blood when compared to other body fluids (Figure 4.3). In general, all CpGs within each of the identified markers present clearly distinguishable methylation levels between the target body fluid and all other body fluids examined, and the differences are statistically significant ($p > 0.05$) (Table 4.2). Therefore,

pyrosequencing provides quantitative results for each individual CpG and permits these markers to be utilized for forensic identification of sperm and blood samples. However, if specific methylation values are not required, another quick, simple and inexpensive method to utilize is HRM analysis. HRM analysis requires only a pair of unlabeled primers and the analysis can be completed in one step. The sperm assay based on NMUR2 produces a melt curve with lower melting temperature for sperm when compared to blood, saliva and vaginal secretions. Figures 4.4 and 4.5 illustrate that DNA from sperm samples presents a melting temperature (Tm) of 80.9 °C with a standard deviation of 0.1 °C which is lower than other body fluids (84.5 °C +/- 0.2 for blood, 84.5 °C +/- 0.2 for saliva, 84.5 °C +/- 0.1 for vaginal secretions). This indicates that the melting temperatures for sperm samples were approximately 3.6-degrees lower than the melting temperatures of the other tissue types. For the UBE2U assay, sperm also gave a melt curve that had lower melting than the other body fluids. Sperm DNA demonstrated melting temperature averages of 77.1 °C with a standard deviation of 0.1 °C that were lower than other body fluids (78.7 °C +/- 0.1 for blood, 78.8 °C +/- 0.1 for saliva and 78.7 °C +/- 0.1 for vaginal secretions) as shown in Figures 4.6 and 4.7. This locus showed an approximately 1.6 °C difference in melting temperature in sperm samples when compared to the Tm averages of the other tissues. The fact that these two sperm markers are almost entirely hypomethylated means that most of the cytosines were converted into thymines resulting in amplicons with low GC content and low Tms. However, HRM analysis for the AHRR blood marker did not provide a clear separation in Tm between blood and the other three body fluids. The

pyrosequencing data for AHRR marker showed that methylation levels for the blood samples were between 18% or below whereas the other body fluids had methylation levels of 43% or above (Figure 4.3). This difference was not sufficient to clearly differentiate the blood samples from the other body fluids.

6. Conclusion

In this study, the main goal was to identify and evaluate sets of CpG sites to distinguish the source of biological material normally left behind at crime scenes. NMUR2 and UBE2U assays were developed and shown to be very effective to identify sperm samples. These assays can be employed by pyrosequencing or high resolution melt analysis using the same primer set (i.e. the primer sequences required for each assay are the same for both methods). The AHRR marker was utilized for the identification of blood using a pyrosequencing based method, however the difference in methylation between blood and other body fluids was not sufficient to permit HRM analysis. Overall these new markers show great promise for utilization in methods for body fluid analysis and expand the potential of epigenetic tools in forensic analysis.

7. Reference

Alghanim, H., Wu, W., McCord, B. (2018). DNA methylation assay based on pyrosequencing for determination of smoking status. Electrophoresis, 39(21), 2806-2814.

An, J. H., Choi, A., Shin, K. J., Yang, W. I., Lee, H. Y. (2013). DNA methylation-specific multiplex assays for body fluid identification. International journal of legal medicine, 127(1), 35-43.

Antunes, J., Silva, D. S., Balamurugan, K., Duncan, G., Alho, C. S., McCord, B. (2016). Forensic discrimination of vaginal epithelia by DNA methylation analysis through pyrosequencing. Electrophoresis, 37(21), 2751-2758.

Bird, A. (2002). DNA methylation patterns and epigenetic memory. Genes & development, 16(1), 6-21.

Butler J. (2010). Sample collection, storage and characterization. In: Fundamentals of Forensic DNA Typing. 1st ed. USA: Elsevier, Inc.; 2010. p. 92-3.

Comey, C. T., Koons, B. W., Presley, K. W., Smerick, J. B., Sobieralski, C. A., Stanley, D. M., Baechtel, F. S. (1994). DNA extraction strategies for amplified fragment length polymorphism analysis. Journal of Forensic Science, 39(5), 1254-1269.

Frumkin, D., Wasserstrom, A., Budowle, B., & Davidson, A. (2011). DNA methylation-based forensic tissue identification. Forensic Science International: Genetics, 5(5), 517-524.

Guo, Y., An, L., Ng, H. M., Sy, S. M., & Huen, M. S. (2017). An E2-guided E3 Screen Identifies the RNF17-UBE2U Pair as Regulator of the Radiosensitivity, Immunodeficiency, Dysmorphic Features, and Learning Difficulties (RIDDLE) Syndrome Protein RNF168. Journal of Biological Chemistry, 292(3), 967-978.

Haarmann-Stemmann, T., Bothe, H., Kohli, A., Sydlik, U., Abel, J., Fritsche, E. (2007). Analysis of the transcriptional regulation and molecular function of the aryl hydrocarbon receptor repressor in human cell lines. Drug Metabolism and Disposition, 35(12), 2262-2269.

Hanson, E. K., & Ballantyne, J. (2013). Rapid and inexpensive body fluid identification by RNA profiling-based multiplex High Resolution Melt (HRM) analysis. F1000Research, 2.

Hashimshony, T., Zhang, J., Keshet, I., Bustin, M., & Cedar, H. (2003). The role of DNA methylation in setting up chromatin structure during development. Nature genetics, 34(2), 187.

Igarashi, J., Muroi, S., Kawashima, H., Wang, X., Shinojima, Y., Kitamura, E., et al. (2008). Quantitative analysis of human tissue-specific differences in methylation. Biochemical and biophysical research communications, 376(4), 658-664.

Kitamura, E., Igarashi, J., Morohashi, A., Hida, N., Oinuma, T., Nemoto, N., et al. (2007). Analysis of tissue-specific differentially methylated regions (TDMs) in humans. Genomics, 89(3), 326-337.

Lee, H. Y., Park, M. J., Choi, A., An, J. H., Yang, W. I., & Shin, K. J. (2012). Potential forensic application of DNA methylation profiling to body fluid identification. International journal of legal medicine, 126(1), 55-62.

Madi T, Balamurugan K, Bombardi R, Duncan G, McCord B. The determination of tissue-specific DNA methylation patterns in forensic biofluids using bisulfite modification and pyrosequencing. Electrophoresis. 2012;33(12):1736-45.

Miranda, T. B., & Jones, P. A. (2007). DNA methylation: the nuts and bolts of repression. Journal of cellular physiology, 213(2), 384-390.

Ohgane J, Yagi S, Shiota K (2008) Epigenetics: the DNA methylation profile of tissue-dependent and differentially methylated regions in cells. Placenta 29:S29–S35.

Park JL, Kwon OH, Kim JH, Yoo HS, Lee HC, Woo KM, Kim SY, Lee SH, Kim YS (2014) Identification of body fluid-specific DNA methylation markers for use in forensic science. Forensic Science International: Genetics, 13:147–153

Shan, L., Qiao, X., Crona, J. H., Behan, J., Wang, S., Laz, T., et al. (2000). Identification of a novel neuromedin U receptor subtype expressed in the central nervous system. Journal of Biological Chemistry, 275(50), 39482-39486.

Shen, L., Kondo, Y., Guo, Y. I., Zhang, J., Zhang, L. I., Ahmed, S., et al.. (2007). Genome-wide profiling of DNA methylation reveals a class of normally methylated CpG island promoters. PLoS genetics, 3(10), e181.

Song, F., Mahmood, S., Ghosh, S., Liang, P., Smiraglia, D. J., Nagase, H., & Held, W. A. (2009). Tissue specific differentially methylated regions (TDMR): Changes in DNA methylation during development. Genomics, 93(2), 130-139.

Song, F., Smith, J. F., Kimura, M. T., Morrow, A. D., Matsuyama, T., Nagase, H., Held, W. A. (2005). Association of tissue-specific differentially methylated regions (TDMs) with differential gene expression. Proceedings of the National Academy of Sciences, 102(9), 3336-3341.

Vidaki, A., Daniel, B., & Court, D. S. Forensic DNA methylation profiling—potential opportunities and challenges. Forensic Science International: Genetics,2013;7(5), 499-507.

Virkler, K., & Lednev, I. K. (2009). Analysis of body fluids for forensic purposes: from laboratory testing to non-destructive rapid confirmatory identification at a crime scene. Forensic science international, 188(1-3), 1-17.

Voet, D., Voet J. G. Biochemistry. (2011). John Wiley & Sons Inc. Fourth Edition. Pg. 1246- 1247.

Webb, J. L., Creamer, J. I., & Quickenden, T. I. (2006). A comparison of the presumptive luminol test for blood with four non-chemiluminescent forensic techniques. Luminescence: The journal of biological and chemical luminescence, 21(4), 214-220.

Wittwer, C. T. (2009). High-resolution DNA melting analysis: advancements and limitations. Human Mutation, 30(6), 857-859.

VII: Concluding Remarks

Advances in DNA technology in the 1970s paved the way to detect variation in certain region of the human genome and shifted the focus of human variation research from the protein to DNA. By mid 1980s, the DNA typing was introduced to the forensic community as a method to compare suspects' profiles to DNA evidence in order to determine their potential presence at a crime scene. Today, DNA typing plays an ever increasing role in the criminal justice system. However, if the DNA evidence reported produces no match to either the suspect(s) or the DNA databases, then the case may remain unsolved. When the suspect is missing in a challenging case, law enforcement needs a tool to provide investigative information with respect to the specimens found at the crime scene. For this reason, forensic biologists have utilized a variety of genomic data to develop biomarkers in order to provide phenotypic and ancestral information on suspects. These markers have included single nucleotide polymorphism markers (SNPs), insertion/deletions (InDels) and microhaplotypes. These genetic markers provide a valuable source of intelligence information to investigators in situations in which the DNA is not present in a database.

In this thesis, a new type of biomarker has been developed involving the use of epigenetics. Investigative leads generated by epigenetic markers recovered from DNA evidence have great potential in assisting investigators to develop a probable suspect list and serve as "DNA Witness.

The need for such work is clear when the issues are considered. In the United States for example, according to FBI CODIS and NDIS fact sheet, there are 840

thousand forensic DNA profiles as of March 2018 that are still not associated with any known suspect. Applying forensic DNA phenotyping makers to this evidence can help to reveal the identity of those unidentified profiles in FBI databases as well as in unknown DNA evidence worldwide. In addition, forensic DNA phenotyping could help with missing person identification cases in which reference DNA profiles from ante-mortem samples or from relatives are unavailable. Currently, there are 16,138 open cases in the missing person database in United States that need to be closed according to NIJ-Missing person website. Forensic DNA phenotyping can also be useful in case of unidentified remains, and for disaster victim identification by association with available relevant ante-mortem samples and relatives. The method has been useful in many parentage disputes and genealogical and medical research.

By studying the relationship between various DNA methylation sites and the dynamic environmental effects, additional information about an individual's lifestyle can be traced such as diet, body size and shape, and usage of illicit drugs. Thus, DNA methylation biomarkers can be developed for a wide range of applications which can aid in the new field of forensic DNA phenotyping (FDP) to infer externally visible characteristics and to trace person's lifestyle. In this project, DNA methylation assays were developed to provide three very value tips about the DNA evidence. First, DNA methylation assays were designed so that to identify the type of the body fluid left behind on the scene of the crime which can help in crime scene reconstruction. Second, models were build based on sets of CpG sites capable of estimating the age of the person who to the unidentified DNA profile.

Finally, 4-CpG assay was constructed to predict the smoking status of the person associated with a crime. Successful and continuous research is an important key to the future for scientific endeavor. From those research, new technologies are regularly introduced to expand the capabilities of generating more information from the DNA material. Using the information brought about in this research project together with the other studies being published, it is possible to envision a next generation sequencing based technique that could extract a wide variety of investigative tips about a suspect of a crime using a single multiplex reaction. This important information can provide investigative leads for use in tracing unknown perpetrators, who are unidentifiable using the conventional autosomal STR based DNA typing.

VITA

HUSSAIN JAFFAR HUSSAIN ALGHANIM

Born, Dubai, United Arab Emirates

1996-2000	B.S., Forensic Science + B.S., Chemistry University of New Haven New Haven, CT. USA
2001-2003	M.S., Forensic Science Florida International University Miami, FL. USA
2002-2003	Researcher at USDA Miami, FL. USA
2003-2014	Forensic DNA Analyst Dubai Crime Lab, Dubai, UAE
2015-2019	M.S., Chemistry Florida International University Miami, FL. USA
2015-2019	Doctoral Candidate Florida International University Miami, FL. USA

PUBLICATIONS

Alghanim, H. J., & Almirall, J. R. (2003). Development of microsatellite markers in Cannabis sativa for DNA typing and genetic relatedness analyses. Analytical and Bioanalytical Chemistry, 376(8), 1225-1233.

Alghanim, H., Antunes, J., Silva, D. S. B. S., Alho, C. S., Balamurugan, K., & McCord, B. (2017). Detection and evaluation of DNA methylation markers found at SCGN and KLF14 loci to estimate human age. Forensic Science International: Genetics, 31, 81-88.

Alghanim, H., Wu, W., & McCord, B. (2018). DNA methylation assay based on pyrosequencing for determination of smoking status. *Electrophoresis*, *39*(21), 2806-2814.

McCord, B., Gauthier, Q., Alghanim, H., Antunes, J., Tejero, N. F., Duncan, G., & Balamurugan, K. (2019). Applications of epigenetic methylation in body fluid identification, age determination and phenotyping. Forensic Science International: Genetics Supplement Series.