11-21-2017

The Study of Tissue-Specific DNA Methylation as a Method for the Epigenetic Discrimination of Forensic Samples

Joana AP Antunes

jantu008@fiu.edu

FLORIDA INTERNATIONAL UNIVERSITY

Miami, Florida

THE STUDY OF TISSUE-SPECIFIC DNA METHYLATION AS A METHOD FOR

THE EPIGENETIC DISCRIMINATION OF FORENSIC SAMPLES

A dissertation submitted in partial fulfillment of

the requirements for the degree of

DOCTOR OF PHILOSOPHY

in

BIOCHEMISTRY

by

Joana Alexandra Pereira Antunes

2018

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

This dissertation, written by Joana Alexandra Pereira Antunes, and entitled The Study of Tissue-specific DNA Methylation as a Method for the Epigenetic Discrimination of Forensic Samples, having been approved in respect to style and intellectual content, is referred to you for judgment.

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

_____
Yuan Liu

_____
DeEtta Mills

_____
Kuppareddi Balamurugan

_____
José Almirall

_____
George Duncan

_____
Bruce McCord, Major Professor

Date of Defense: November 21, 2017

The dissertation of Joana Alexandra Pereira Antunes 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, 2018

DEDICATION

I dedicate this work to you.

What brought me here were the struggles, laugh, happiness and sadness from me, my

ancestors' and their ancestors and all the elements that combined to make us all.

We all descend from that one singularity.

ACKNOWLEDGMENTS

I would like to thank all of those who made this work possible. In my daily life, I get to interact with many people that affect my perception and challenge me to grow. Moving to Miami was one of the most difficult things I've endured in my life, which means it brought one of the most valuable opportunities to grow, and I did. This work represented the most stable form of that growth. I was able to see my work capacities, including my limitations and my strengths as a researcher, and evolve from them. Everyone I've encountered either further challenged me or held my hand to help me move. Each one showed up exactly when they were supposed to.

First I would like to thank Dr. McCord. His optimism and belief in me happened even before I arrived in Miami. Following my arrival and during my research efforts, he showed great support and understanding of my goals, not only with my work, but also in my personal life. Words cannot describe the profound admiration and appreciation I have for him as a scientist, a leader, and a person. Thank you!

Another individual with whom I contacted regularly regarding this project is Dr. Duncan. His dedication and profound interest in science, being informed of recent developments, all novelty and technology is inspiring. From the many explorations of this topic to the many other topics we've talked about along the years, I sincerely hope to keep contact with you in the future and hopefully develop great ideas together. Thank you!

To my dissertation committee, thank you for your support and challenging ideas you have developed with me throughout these years. Thank you for believing in my

abilities to succeed in this degree and for getting through the ups and downs that existed along the way.

Dr. Bala, thank you for all your help in troubleshooting methods, experiments, and ideas on this project. Having you as a collaborator, and also my committee member allowed me to consult with you more often than only in committee meetings. Your contribution to this project is amazing and continues to provide more pieces of this puzzle so that one day all of this can be routine use in forensic laboratories.

Dr. Liu, ever since the first time I talked to you I have received the upmost respect and also felt great admiration for your work. Thank you for being a great teacher of biochemistry and to always bring up the more fundamental aspect of my research, which is by itself a separate work from its forensic application.

Dr. Mills, thank you for showing me the side of forensics that deals with non-human samples. Whether it is for species protection or on the use of other species cues to help human cases, the diversity of fields your research includes is truly admirable. It is very inspiring for me to have you as a model representing the professional woman in the forensic field.

Dr. Almirall, thank you for your continuous support and belief in this work and my ability to succeed. Even though the topic of this dissertation is somewhat different from your day-to-day work, I am extremely grateful for your dedication to this project demonstrated so many times during my annual evaluation presentations.

I also would like to thank all my lab mates throughout the years that helped me to grow as a researcher.

ABSTRACT OF THE DISSERTATION

THE STUDY OF TISSUE-SPECIFIC DNA METHYLATION AS A METHOD FOR

THE EPIGENETIC DISCRIMINATION OF FORENSIC SAMPLES

by

Joana Alexandra Pereira Antunes

Florida International University, 2018

Miami, Florida

Professor Bruce McCord, Major Professor

In forensic sciences, the serological methods used to determine which body fluid was collected from the crime scene are merely presumptive or labor intensive since they rely on protein detection or on microscopic identification of cells. Given that certain forensic cases may need the precise identification of a body fluid to determine criminal contact, such is the example of a suspected sexual assault of a minor; certainty in the body fluid of origin may depict a precise picture of the events. The identification of loci that show differences in methylation according to the tissue of origin can aid forensic analysts in determining the origin of a DNA sample. The process of DNA methylation occurs naturally in the genome of living organisms and consists in the presence of a methyl group on the carbon 5 of a cytosine, which is typically followed by a guanine (CpG). Analyzing patterns of DNA methylation in body fluids collected from a crime scene is preferential to the analysis of proteins or mRNA since the same extracted DNA used for STR typing can be used for DNA methylation analysis. We have validated and identified loci able to discriminate blood, saliva, semen and vaginal epithelia. In the current study, we have also established the minimum amount of DNA able to provide

reliable results using methodologies such as pyrosequencing and high-resolution melt (HRM) analysis for the different markers identified. Lastly, we performed an alternative bioinformatic analysis of data collected using an array that studied methylation in over 450,000 individual cytosines on the human genome. We were able to sort the locations that showed potentially higher methylation differences between body fluids and investigated over 100 of them using HRM analysis. The results of that study, allowed the identification of three new loci able to distinguish blood and two new loci able to distinguish saliva and vaginal epithelia, respectively.

The use of DNA methylation patterns to aid forensic investigations started with a publication in 2010, therefore each small contribution such as this work may, similarly to what occured in the biochemistry field, result in the discovery of a method able to put the technology in the hands of forensic analysts.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABBREVIATIONS AND ACCRONYMS

3'-OH	3'-hydroxyl group
5-caC	5-carboxylcytosine
5-fC	5formylcytosine
5-hmC	5-hydrozymethylcytosines
5-mC	5-methylcytosines
A	Adenine
ANOVA	Analysis of variance
ATP	Adenosine triphosphate
BER	Base excision repair
bp	Base pairs
BSA	Bovine serum albumin
C	Cytosine
CpGi	CpG islands
DEPC	diethyl pyrocarbonate
DMSO	Dimethylsulfoxide
DNA	Deoxyribonucleic acid
DNMTs	DNA methyltransferases
dNTP	deoxynucleoside triphosphate
dNTPs	Deoxyribonucleotides
ESC	Pluripotentent stem cells

G	Guanine
HRM	High-resolution melt
kb	kilobases
KMTs	Histone lysine methyltransferases
Mb	Megabases
miRNA	Micro RNA
mRNA	Messenger RNA
ms-SNuPE	Methylation-sensitive single nucleotide primer extension
ORF	Open reading frame
PCR	Polymerase chain reaction
RFLP	Restriction fragment length polymorphism
RFU	Relative fluorescence units
RNA	Ribonucleic acid
SBE	Single base extension
ssDNA	Single-stranded DNA
STRs	Short Tandem repeats
T	Thymine
Taq	*Thermus acquaticus*
tDMRs	tissue-specific DNA methylation regions
TET	Ten-eleven translocation
TGS	Transcriptional gene silencing
T_M	Melting temperature

tRNA   Transfer RNA

TSS   Transcription start site

UHRF1   (ubiquitin-like, containing PHD and RING finger domains 1

UTRs   Untranslated regions

CHAPTER I - INTRODUCTION AND LITERATURE REVIEW

A. The Molecular Structure and Biochemistry of DNA

The molecular structure of deoxyribonucleic acid (DNA) was established in 1953 as a result of several years of research. Maurice Wilkins and Rosalind Franklin were able to capture high quality X-ray diffraction photographs of the molecular structure of DNA (1). In 1952, Alexander Todd and colleagues (2) established the existence of 3'-5' phosphodiester bonds in DNA which laid the ground work for the final structure of the double helix explained by James D. Watson and Francis Crick in 1953 (3). The DNA structure includes a double helix composed of two linear strands, similar in structure, that bind to one another through hydrogen bonds. Each linear strand is composed of building blocks named adenine (A), cytosine (C), guanine (G) and thymine (T) that compose the interior part of the double helix. These nitrogen bases are linked to a sugar-phosphate backbone that forms the linear structure. The two linear strands of DNA form the double helix by complementary hydrogen bonds between the nucleotides or nitrogen bases (Figure 1.1). Each base-pair on the double helix was named Watson-Crick base pairing. The double helix coils around a central axis and the stacking of the bases on its interior has the distance of 3.4 Å. Each base has a rotation of 36 degrees from the adjacent ones and therefore each turn (360 degrees) is composed of 10 bases (4). The two DNA strands within the double helix are also organized in an anti-parallel fashion. The connection between adjacent nucleotides is made between the previous nucleotide and the 5'carbon and the hydroxyl group on the 3'carbon of the deoxyribose to the next nucleotide.
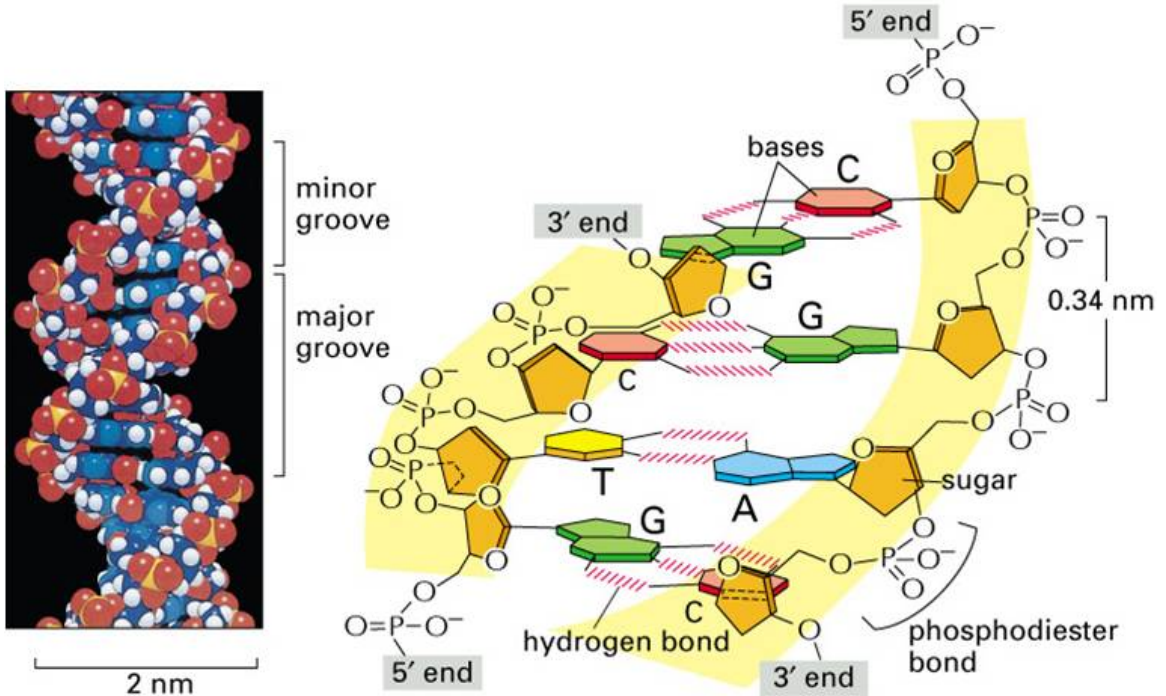
Figure 1.1 − The structure of the DNA double helix showing hydrogen bonds between two antiparallel strands. The nucleotides on the same strand are linked together by phosphodiester bonds through the 3'-OH of one group to the 5'-phosphate of the next. Adapted from (4).

The formation of the double helix depends on different chemical forces that act on the structure's constituents. The covalent binding of atoms in the nucleotides' structures is the strongest chemical force present in the DNA structure. For that reason, the building blocks are stable structures and rarely suffer modifications on their structure. Several forces are responsible for the high stability of double-stranded DNA when compared to single-stranded structure. Hydrogen bonds are formed between a hydrogen donor and an acceptor. The hydrogen atom itself is shared between two electronegative atoms such as nitrogen and oxygen, typically in complementary nucleotides of the double-strand structure. In this fashion, adenine base pairs with thymine in a double hydrogen bond and cytosine base-pairs with guanine with a triple hydrogen bond (5). When the two DNA

strands exist apart from one another, the nucleotides are exposed to water forming hydrogen bonds with the molecules of water. As the correct complementary nucleotide gets closer, the hydrogen bonds that form between the complementary nucleotides replace those formed with water molecules. Because the energy required to make the new bonds and to disrupt those with water is similar, the total energy of the molecule and its environment is constant and the First Law of Thermodynamics is observed (5). However, if the nucleotides are not complementary, some of the hydrogen bonds cannot form which makes the reaction greatly disfavored. Additionally, the formation of a non-polar moiety within the double-strand contributes to hydrophobic interactions that force water to the outside of the structure. These forces help explain why there is a release of energy in the formation of the double strand, therefore increasing the entropy of the surroundings and favoring the double-helix formation. Van der waals forces between adjacent nucleotide bases in the double helix are also optimal which further contributes to the stability of this structure when compared to single stranded DNA (4). However, some unfavorable electrostatic interactions between the negative phosphate groups of the backbone also exist. These unfavorable interactions occur mainly on adjacent phosphate groups present when the double strand circles around the axis. These forces are reduced because of the presence of positively charged ionic species like sodium and magnesium and the dielectric constant of water that stabilizes the negative charge of phosphates.

Even before Watson and Crick proposed this model for the double-helix, there was already the notion that DNA was responsible for storing genetic information able to code for effector proteins that carry specific tasks within a cell (5).

B. The Location of Genomic DNA and The Central Dogma

Deoxyribonucleic acid itself is not the effector biomolecule that allows all cellular processes to exist. Genomic or nuclear DNA is located in the nucleus of a cell and therefore cannot process cellular mechanisms that operate on the cytoplasm and other organelles of the cell. Instead, DNA contains the information necessary to synthesize proteins, which are the effector molecules within cells (5).

To explain how the information stored in the DNA can be transformed into effector proteins, Francis Crick proposed the central dogma pathway in 1958 (6). According to the central dogma model, genetic information stored in the DNA molecule is transcribed to ribonucleic acid (RNA), which carries it to the cytoplasm. Once in the cytoplasm of the cell, RNA is then translated to proteins. Specific characteristics make this model unique (6, 7). The first characteristic is that the information stored in the DNA code can be duplicated to form two identical copies of the same template of DNA. The duplication of DNA happens before cell division in order for the new cell to obtain one DNA molecule with the same genetic information as the template. The second characteristic of the central dogma is that the flux of information happens in a single direction from DNA, to RNA and then proteins. Third, the intermediate player – RNA – is transcribed from the template DNA and travels to the cytoplasm transferring the information to be translated into proteins.

The duplication (also known as replication) of DNA depends on the concerted action of enzymes such as helicases, DNA polymerases and DNA ligases. The initial step necessary to DNA replication is the unwinding of the double strand at specific regions called origins of replication by the initiator proteins. Each human chromosome has an

origin of replication for each 30-40 kilobases (kb). The DNA helicase then binds to the

single-stranded DNA (ssDNA) to further unwind it and allow the primase to synthesize

RNA primers necessary for DNA duplication (8). The primase is a specific RNA

polymerase that specifically synthesizes 5-10 RNA nucleotides complementary to the

ssDNA at the origin of replication. The primer:template junction acts as a substrate for

DNA polymerase and signals it to initiate copying the DNA molecule. The tertiary

structure of human DNA polymerase enzymes resembles a hand that grips the DNA

molecule. The catalysis occurs in the palm whereas the fingers and thumb provide the

stabilization residues necessary to the correct nucleotide to be added to the newly formed

strand(9). In the moiety of the DNA polymerase, the free 3'-hydroxyl (3'-OH) group of

the primer:template junction attacks the alpha-phosphoryl group available on the next

deoxynucleoside triphosphate (dNTP) brought close by the polymerase.

Also associated with the DNA polymerase are sliding clamps that maintain its association

with DNA even after the end of new strand by the polymerase. The sliding clamp allows

the action of other enzymes necessary to the correct duplication of DNA and provides

easier association of different DNA polymerases in case corrections are needed in the

new copy of DNA. The process continues through DNA unwinding by helicases ahead of

the association between the DNA polymerase that copies the leading strand with a direct

addition of nucleotides to the 3' –end of the new strand and the template DNA, which

forms the replication fork (9). Since the replication can only occur by addition of

nucleotides on a free 3'-OH and each ssDNA runs anti-parallel to one another on the

double helix, only one template strand (called the leading strand) can be directly copied

as the replication fork forms. The complementary strand that runs in the opposite

direction of the leading strand cannot be immediately copied and is therefore called the lagging strand. Only when there is enough lagging strand template present on the replication fork, can a second sliding clamp associate to the lagging strand, allowing the primase to synthesize a new primer:template junction. The replication of the lagging strand is therefore achieved by the synthesis of small DNA fragments, called Okazaki fragments, instead of a single long strand (8). Each fragment is formed as a result of the progression of the replication fork and recruitment of its own set of replication machinery. At the same time that the DNA polymerase copies the lagging strand on the second Okazaki fragment, it also uses its 5'-3' exonuclease activity to remove the portion of RNA primer from the first fragment. The existing nick between the two Okazaki fragments is sealed by another enzyme present in the replication machinery called DNA ligase. The process of copying one entire DNA molecule allows the duplication of information prior to cell division and thus the propagation of the genetic code to the subsequent generations.

The second process described by the central dogma is called transcription. Transcription allows the code stored in the DNA to be transcribed to RNA. Because of the similarity between the two molecules, the process of transcription presents some common steps to replication. The main differences reside in the specific enzymes that catalyze the reaction and in the product formed.

RNA polymerases are enzymes capable of reading deoxyribonucleotides from DNA and convert them to ribonucleotides in the RNA. The pairing of complementary nucleotides is similar to DNA, except to the adenine that is base-paired with uridine, since RNA does not contain thymines. Another difference is that the RNA molecule exists as a single

strand, the only time it exists as double-paired is during transcription, when it is paired with DNA.

The initiation of transcription happens when RNA polymerase binds to the DNA at promoter regions. The DNA nucleotide that is first transcribed is called the transcription start site (TSS) and is referred to as the '+1' position. Nucleotides located before the TSS are called upstream and their relative distance to the TSS is established by negative numbers, for example, -10 corresponds to the nucleotide located 10 base-pairs upstream of the TSS. Similarly, nucleotides located after the TSS are called downstream and carry positive numbers (10). Before the initiation of transcription, proteins capable of recruiting the transcription machinery, which includes the RNA polymerase, recognize specific DNA sequences. Those DNA sequences, also called elements, are present near the TSS in the core promoter regions or in other regulatory sequences further away upstream or downstream of the TSS (10). Each core promoter region can be 40-60 nucleotides long and contain different combinations of the possible elements. Transcription occurs from the 5' to the 3'end similarly to replication and it does not require primers. As elongation occurs, the polymerase encounters specific sequences that signal the end of the gene thus promoting termination. Transcription of those specific sequences triggers the transfer of polyadenylation enzymes to the newly formed RNA. The RNA copied from a single strand DNA needs to undergo post-transcription modifications such as addition of a poly-A tail, addition of a 5'-cap and splicing (11). Polyadenylation enzymes catalyze the addition of about 200 adenines (poly-A tail) to the 3'-end of the newly formed RNA (pre-mRNA) and also allow the complete detachment of RNA polymerase and the messenger RNA (mRNA) from the DNA molecule. The

5'cap consists in the addition of a modified guanine to the 5'-end of the newly formed RNA. The 5' cap has functions in the transport of mRNA to cytoplasm as well as important functions in translation to proteins and protection against exonucleases. Before exiting to the cytoplasm, the pre-mRNA must also be spliced to remove any regions of RNA transcribed from DNA but that do not code for proteins (11).

Within each DNA region that codes for proteins, downstream of the promoter, the DNA sequence has coding regions or exons interspaced by non-coding regions called introns. The pre-mRNA contains a transcription of both exons and introns, which need to be removed. The spliceosome is composed of 150 proteins and 5 RNAs that catalyze transesterification reactions responsible for RNA splicing - the removal of introns and combination of exons to form a mature mRNA. The importance of this process is that it allows the generation of several spliced mRNAs that code for different polypeptides from one pre-mRNA molecule transcribed from the same DNA region (11). The process from splicing the RNA and formation of a mature mRNA with different exons is called alternative splicing and it accounts for the diversity of proteins that each gene can code. In humans, 75% of genes undergo alternative splicing, explaining the enormous diversity observed for the number of genes carried by each individual considering the size of the genome. Since exons from different genes can combine to form new proteins, alternative splicing is thought to have greatly contributed to evolution and to the immense increase in cell functions from simpler organisms to the complex biology in humans (12).

The last process described in the central dogma is translation. The translation of mRNA happens in the cell cytoplasm, therefore the mRNA must exit the nucleus through a pore protein called nuclear pore complex. As it reaches the cytoplasm, translation can

start at the 5' end of a string of codons called open reading frame (ORF, (13)). In eukaryotes, the first codon is called start codon and is the sequence of the ribonucleotides 5'-AUG-3'. Translation ends with one of three possible stop codons: 5'-UGA-3', 5'-UAA-3' or 5'- UAG-3'. The translation process occurs by the ribosome being recruited and assembled at each mRNA. The recruitment is initiated by the presence of the 5'cap on the mRNA which promotes two subunits of the ribosome (60S and 40S) to assemble at the 5-end of the mRNA. In the newly formed protein, the larger subunit is responsible to form the peptide bonds between amino acids whereas the smaller subunit decodes the 3 nucleotide codons to recruit the corresponding amino acid. In fact, even before assembly, the small subunit is already associated with a transfer RNA (tRNA) that codes for methionine and recognizes the start codon (13). After assembly, the ribosome scans the mRNA until it finds that start codon allowing the first methionine to initiate translation. As translation progresses, the ribosome translocates to the next codon allowing several ribosomes to simultaneously translate a single mRNA molecule. Each ribosome is capable of associating simultaneously with 3 tRNAs that recognize 3 successive codons. Thus, the ribosome is a specialized organelle that allows the concerted action of tRNAs, which specifically recognize codons providing the respective amino acid for protein synthesis. Upon encountering the stop codon, the ribosome disassembles, translation ends and the polypeptide formed dissociates from both subunits (13).

The genetic code followed by each tRNA corresponds a codon and an amino acid. Each tRNA possesses an anti-codon sequence which is complementary to the mRNA, and it also carries the respective amino acid according to the genetic code. Since each codon has 3 ribonuclotides, there are 64 possible codons. From those 64 codons, 3 are used as a

stop signal and the remaining 61 codons are translated to 20 amino acids(14). As a result, each amino acid is coded by more than one codon, which classifies the genetic code degenerate. The degeneracy of the genetic code promotes correct translation even in the presence of random single nucleotide errors that could accumulate during either replication or transcription (15).

C. The Structure of the Eukaryotic Cell

The evolution of species throughout time depended on the selective advantage for unicellular organisms to combine into multicellular colonies. With evolutionary time, certain individuals in each colony were able to differentiate and have specific functions. The colony was then able to form an independent tissue with complex signaling pathways between all individual cells. Therefore, the single, common structure to all living organisms is the cell. Cells can be prokaryotic or eukaryotic. The cells that compose the tissues in the human body are eukaryotic and share common characteristics with other organisms that are composed of the same type of cells (16).

In eukaryotic cells, the cellular membrane separates the cell from the outside environment, keeping a balanced level of salts and an adequate pH to cell functions. The membrane is composed of a bilayer of phospholipids associated with proteins. Some proteins can function as pores or channels that allow specific passage of extracellular components, such as nutrients, neurotransmitters, hormones and others. Inside the cell, a fluidic structure called the cytoplasm maintains several organelles that carry out pathways allowing for cell survival and response to external stimulus. Another structure present inside the cells is the nucleus. A nuclear membrane separates the genetic

information stored in the DNA from the cytoplasm where the organelles are located. The compartmentalization of organelles is a characteristic of eukaryotic cells which provides them with a higher complexity when compared to prokaryotic cells. Since each specific compartment can have its own environment, the compartmentalization allows the survival of organelle-specific enzymes which leads to more specialized organelles. At any certain point in time, numerous pathways take place in an eukaryotic cell due to the presence of proteins, ions, water, genetic material (DNA and RNA), organelles and several biochemical compounds that are either the substrate or the product for such pathways (16).

Genomic DNA is found in the nucleus of the cell condensed into chromosomes and mitochondrial DNA is found in the mitochondria in the form of a circular molecule. The genomic human DNA is extremely lengthy and therefore needs to be compacted into chromosomes in order to fit inside the nucleus of the cell. Human chromosomes are linear structures, contain telomeres which are located at the ends of the chromosome arms; and a central region – centromere – connecting the arms. In the process of cell division, the centromeres form a kinetochore which interacts with protein filaments that help pull the two copies of each chromosome to opposite sides of the dividing cell (17). The mitochondrial DNA is located in the mitochondria which is the main organelle responsible for the synthesis of adenosine triphosphate (ATP) - the main energy currency in the cell. The mitochondria are capable of oxidizing proteins, lipids and sugars to convert them to carbon dioxide and water with ATP production. The mitochondria also has its own RNA, ribosomes and a small number of proteins that allow for self-assembly. Similar characteristics between contemporary mitochondria and bacteria, illustrates how

11

evolution may have occurred. A bacterium capable of oxidizing nutrients fused with an archaic eukaryotic cell providing it with an enhanced system of acquiring ATP from more complex nutrients. The fusion of both cells proved to be advantageous and therefore such symbiosis relationship was selected (17).

D. The Cell Cycle

In order to form an organism, the initial cell is composed by the fusion of male and female gametes and must divide several times to form a population of cells. Once that occurs, some cells gain specialized functions and then further divide to maintain a constant replacement of cells that are either damaged or that have died resulting from loss of functions (apoptosis). Both pathways – expansion of a population of cells to form and organism, or cell division to maintain the organism – have similar characteristics. In both cases the cell cycle is composed of a series of events that allow for the duplication of all the constituents of a cell followed by the division of such cell giving origin to two identical daughter cells (18). When a new organism is being formed, the initial cell is undifferentiated and is commonly called a stem cell. Once it receives specific stimulus, the stem cell can divide and one of the daughter cells undergoes a cycle of rapid duplication and division to form a subpopulation of cells. Upon the cessation of cell divisions, the subpopulation of cells differentiates to acquire the characteristics of the tissue they form. In humans, the process from stem cell to differentiated cell occurs for example in the zygote cell formed upon egg fertilization and also in adult tissues that are constantly regenerated, such as skin. In both cases, the cell cycle is divided into

interphase in which the cell grows and duplicates all of its contents, and a second phase called M phase, where the cell divides into two identical daughter cells (18).

When differentiated cells divide, they do so in order to replace cells that die. In such case, the number of chromosomes must be maintained and the division is called mitosis. During the interphase, more specifically in the S phase, the chromosomes duplicate by forming an identical copy of each other through DNA duplication. Before and after the S phase, two gap phases G1 and G2 function as check points to both prepare for the next phase and to confirm that the previous actions were completed successfully. The M phase follows which is composed of a rearrangement of the chromosomes in the center of the cell to prepare it for division (18). The kinetochores and associated microtubules pull the chromosomes in separate directions and so when the cytoplasm of the cell divides, each new daughter cell has a diploid number of chromosomes, similar to the cell that gave them origin.

Somatic cells have '2n' or diploid number of copies of chromosomes whereas gametes have haploid or 'n' number of chromosomes. In humans the gametes are the sperm cell and the egg and when they fuse, they form the zygote with a diploid number of chromosomes. Both egg and sperm are formed from precursors of reproductive cells which are diploid. When these cells divide, they do so by a process called meiosis. In meiosis, after the initial duplication of chromosomes and cell components, there are two consecutive rounds of cell divisions without further duplication of chromosomes. The two cell divisions that occur after chromosome duplication leave each gamete with haploid number of chromosomes, allowing them to combine to another gamete and form a diploid cell (18).

E. DNA organization in Nucleosomes and Chromosomes

Human chromosomes are linear structures composed of DNA wrapped around proteins. The number of chromosomes is characteristic of each type of eukaryotic cell. In humans, all cells except the reproductive cells, are diploid and contain two copies of each of the 22 chromosomes plus a single copy of the sex chromosomes, X and Y for males and two copies of X for females.

The chromosomes are visible under the microscope and their structure varies according to each phase of the cell cycle. For example, chromosomes in the M phase are condensed and present a typical 'X' shape with visible arms and the centromere (18, 19). During interphase, the structure is less compact and it is called chromatin. The more compact version of chromatin is called the 30-nm fiber whereas the less compact is called the 10-nm fiber. Under the microscope, the 10-nm fiber has the appearance of beads on a string. The chromatin fiber is composed of nucleosomes which are constituted by a core of histone proteins and a DNA molecule that is wrapped around them. The shape of beads on a string visible under the microscope is therefore the nucleosomes (beads) with a string of approximately 20-60 base pairs (bp) of DNA (linker DNA) uniting them (19). The histones are proteins that are specific to nucleosomes. Each nucleosome contains two copies of H2A, H2B, H3 and H4 histones. Immediately after DNA replication, the nucleosomes assemble thanks to the specific action of histone chaperones which are negatively charged. The replicating DNA functions as a signal for the chaperones, which are then directed to form complexes with the tetramers or dimers of the H2A, H2B, H3 and H4. To initiate the formation of the nucleosome, a tetramer of two copies of H3 and H4 associates with DNA, promoting the recruitment of two H2A.H2B dimers. Each

14

newly formed nucleosome contains histones from the parental cell that were released upon relaxation of the chromatin fiber and new histones formed during the interphase. Since some of the existing histones are recycled to the new cell, the chaperone needs only to direct the assembly of the new ones. Additionally, the histones regenerated from the mother cell carry with them important modifications that can be copied to the daughter cells (19).

The DNA is wound 1.65 times around the core of histones for a total of 147 bp per nucleosome. Further condensation of the structure is achieved when a fifth histone – H1 – binds to the linker DNA, further condensing the chromatin fiber and originating the 30-nm fiber. The histones have amino-terminal tails that protrude from the nucleosome core and that can have modifications such as presence of methyl or acetyl groups that help recruit proteins and also stabilize the DNA. Depending on which modification is present in the histone tails at each time, the proteins recruited can help further condensing or relaxing the chromatin structure. A more relaxed structure is more suitable for DNA replication or transcription (19).

Overall the human genome is composed of 3,200 Mb (Mega-bases) of DNA from which only 48 Mb are genes that code for proteins. Introns, unstranslated regions (UTRs) and pseudogenes compose about 1,152 Mb and another 2,000 Mb are intergenic DNA. Regulatory regions such as those that code for miRNA allow the regulation of gene expression (19). About 3% (or 90 Mb) of DNA is composed of microsatellites, which are short-tandem repeated sequences of less than 13 nucleotides. The gene density defined as the number of genes per Mb of DNA, and not the total number of genes or length of DNA is what determines the level of complexity of a species. More complex species have

their gene-coding regions interspaced with non-coding DNA such as repeated sequences, intergenic sequences, introns, etc, thus having a lower gene density (19).

F. Epigenetics

Epigenetics is a term that was first defined by Doctor C.H. Waddington in 1942 to explain the interactions between "genes and their products, which bring the phenotype into being." (20). Nowadays the term 'epigenetics' is used to coin all the genome modifications that cause differential gene expression, independent from changes in the DNA sequence (21). Some of the mechanisms included in epigenetics cause regulation of gene expression through changes in the level of chromatin condensation, changes in the mRNA resulting from transcription or changes in the DNA itself - DNA methylation. Since the DNA contains the genetic information that codes for chromatin condensation and mRNA synthesis, DNA methylation can be linked to modifications in the chromatin and RNA therefore resulting in cell-specific gene expression (22).

The epigenetic mechanisms that allow changes in the degree of relaxation of chromatin are commonly known as chromatin remodeling and can directly influence DNA transcription. The nucleosomes are composed of 147 bp of DNA wrapped around histones. If the wrapping is tight, the DNA is less likely to be available for transcription (19). The tightening of the chromatin fiber can be promoted by the presence of different modifications on the histone tails, which are mostly composed of residues of arginine and lysine (5). The specific modifications on the histone tails occur post-translationally and can be methylation, acetylation, phosphorylation, propionylation, butyrylation, ubiquitylation, sumoylation and citrullination. Each set of modifications in the histone

16

tails is known as 'histone marks' and a combination of several histone marks leading to a gene regulation event is known as 'histone code' (23).

One of the most studied examples of histone modification is the methylation of lysine (K) residues in tails of histone H3 that can either cause gene repression when tri-methylation occurs on lysine 9 of histone 3 (H3K9me3) or gene activation for H3K4me3 (24). Another type of histone modifications intimately related to gene expression consists on the presence of histone variants. Some histones contain differences in few amino acid residues when compared to the canonical histones. For example, in mammals there is a known variant of the histone H2 called H2A.X. When H2A.X is phosphorylated, it marks a double strand DNA break. Another variant known as H2A.Z, which is widely present at transcription start sites, shows a negative correlation with the existence of DNA methylation. Even though information regarding the mechanism of action of these variants is limited, the existence of dedicated mechanisms responsible for the removal and addition of these histone variants into the nucleosome further reinforces the importance of these histones in gene expression (22).

Gene expression can also be regulated at posttranscriptional levels through RNA. Two main mechanisms are known to date: one is related to the RNA secondary structure and the other related to the binding of small regulatory RNAs to the mRNA. Regulation through the RNA secondary structure occurs when a specific factor binds to certain mRNA transcripts preventing them from unfolding thus preventing translation. For example, in eukaryotes several proteins such as ferritin mediate the harmful accumulation of iron. Ferritin is an iron-storage protein mainly found in the liver and kidneys. The ferritin mRNA can only be translated when the 5'-end stem-loop is free

from another protein called IRP. Since iron binds to IRP strongly, in high concentrations of iron, the IRP preferentially binds to iron releasing the ferritin mRNA and promoting translation for higher protein availability (5).

Another type of gene expression at the RNA level can occur when specific mRNA transcripts are deleted. The deletion of mRNA after it has been transcribed occurs when certain small RNA – called micro RNAs (miRNA) – are transcribed from specific genes. After being synthesized the miRNAs bind specific proteins from the Argonaute family, which have the capacity to degrade mRNA transcripts complementary to the miRNA sequence. Since each miRNA sequence can be complementary to several mRNA transcripts, several types of mRNAs can be degraded in this manner (5). Recent discoveries show that this form of gene silencing mediated by miRNA can establish a repressed chromatin state through back-signaling to DNA. Such chromatin state represents a more stable gene silencing mechanisms since it can be propagated through several cell divisions (22). Additionally, small nuclear RNAs can also operate through transcriptional gene silencing (TGS, see Figure 1.2) pathways. It has been proposed that in order to defend the organism from harmful DNA elements introduced by viruses, small RNAs can interact with effectors of histone lysine methylation pathway to silence the extraneous DNA (22).

The third type of epigenetic modification, the DNA methylation, has been widely studied. In humans methylation occurs primarily in the 5' carbon of cytosines followed by guanines (CpGs) and because of the large effect of that methyl group in gene expression, methylated cytosines have been referred by some as the 5th nucleotide (25). In fact, DNA methylation plays important roles in several genetic processes such as

repression of transposons and repetitive elements, genomic imprinting and X-chromosome inactivation, where one of the X chromosomes is silenced in every somatic cell of female mammals (21, 26).
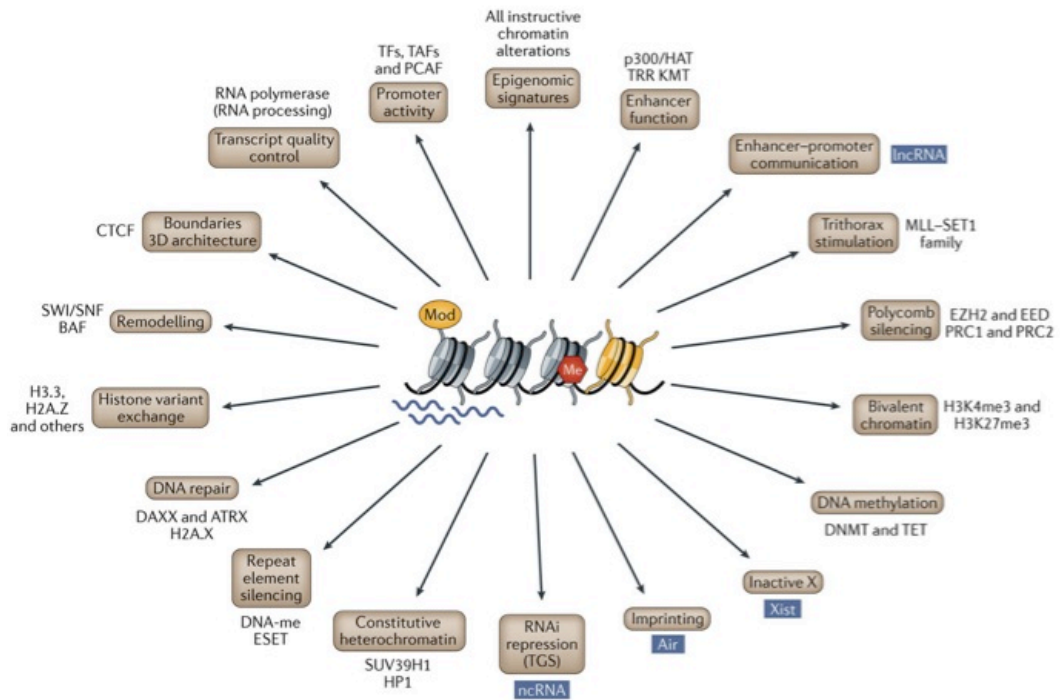
Figure 1.2 – Schematic representation showing all the genomic and epigenomic factors and mechanisms involved in gene regulation. The center of the figure represents 4 nucleosomes and the main mechanisms for epigenetic regulation such as DNA methylation (red circle with Me), histone modifications (Mod), histone variants (yellow nucleosome) and non-coding RNA (wavy blue lines). The brown boxes display genomic functions that affect epigenome modifications and help in its regulation. The main factors needed for genomic functions are shown near the boxes. Adapted from (22)

The close physical relationship between DNA and histones in the nucleosomes precludes possible correlations between the different states of chromatin and DNA methylation. Ongoing research aims to further draw a mechanism of how such interactions affect gene expression. In 2010 direct evidence was found that connect the methylation status of CpGs to recruitment of proteins able to mediate modifications in the

histone tails. One example is the specific recognition and binding of protein CXXC-type zinc finger protein 1 to unmethylated CpGs which promotes the recruitment of enzymes called histone lysine methyltransferases (KMTs). Methylation of histone tails on their lysine residue by KMTs seems therefore to be influenced by unmethylated cytosines on associated DNA regions (27).

G. DNA Methylation

The genetic diversity observed in humans is not only a result of DNA polymorphisms in the DNA sequence that accumulate over generations, but also due to a more transient modification – DNA methylation – that is altered during the life of one individual and accounts for phenotypic changes (24). In the 1980s the link between DNA methylation and gene expression was established. After blocking DNA methylation in vitro using the drug 5-azacytidine, researchers were able to establish the link between irregular DNA methylation and the onset of cancer. The connection between anomalous DNA methylation and such a devastating disease such as cancer greatly increased research interests towards studying DNA methylation and the enzymes responsible for maintaining a stable methylome (22).

Typically, the cytosines that can undergo methylation are followed by a guanine and therefore they are commonly referred as CpGs. The 'p' between the C and G highlights the phosphate link between the two nucleotides on the same strand of DNA as opposed to the complementary cytosine-guanine hydrogen bond in the Watson-Crick base-pairing. Cytosines in other dinucleotides such as CpA, CpT or CpC can also undergo methylation, however those forms are less common.

The DNA methyltransferases (DNMTs) promote the catalysis of cytosine methylation (28). Methylation in somatic cells of mammals is most predominantly seen in CpGs as a result of an increased affinity (40-500 fold) of DNMTs to that dinucleotide, when compared to the dinucleotides CpA, CpT or CpC (29). Non-CpG methylation is therefore a rare event in somatic cells. A different scenario is present in pluripotent cells such as stem cells where non-CpG methylation accounts for about 25% of all methylation seen in those cells. Interestingly, the levels of non-CpG methylation for *in vitro*-induced pluripotent cells are similar to those of embryonic stem cells, indicating that non-CpG methylation levels are re-established when cells are in their undifferentiated state (28, 29). The exact molecular mechanisms that allow the non-CpG methylation to occur only in pluripotent cells can be of great interest to understand how DNA methylation is involved in establishing cell type identity and to help future determination of genomic regions that are involved in cell differentiation.

The successful establishment of an embryo depends on how gene expression and silencing is modulated between the two progenitor's DNA. All stages from oocyte development, maturation and after fertilization are marked with several epigenetic modifications. Whole genome demethylation occurs in two specific events in the embryo development – during germ cell migration and immediately after fertilization. Demethylation of DNA in these stages is crucial to develop the new embryo's epigenome by first erasing the progenitor's epigenome and thus establishing state of pluripotency necessary for cell development (30). After embryonic implantation, a *de novo* methylation is generally observed in the whole genome occurs and a new methylome is established (31). In contrast, *de novo* methylation is not so prominent in specific tissues

during differentiation, establishing a link between low level of new methylation in the genome and cell differentiation. At the tissue level, *de novo* methylation is therefore rare in normal development and cell differentiation but is widely seen in the development of cell lines *in vitro* and in cancer (32). Research using pluripotent stem cells (ESC) have observed that unlike somatic cells, ESC have the ability to *de novo* methylate exogenous DNA introduced in the cells. At this stage of organism development, unique motifs in the DNA sequence determine which regions are methylated. It has been observed that at the time of implantation, some CpGs such as those of housekeeping genes, are protected from de novo methylation. Most of those unmethylated CpGs are located near places where the RNA polymerase binds (transcription start sites) marked with H3K4me3, suggesting a role for RNA polymerase to bind the DNA and protect it from de novo methylation (31).

Regarding the distribution of DNA methylation in adult human genomes (somatic cells), it has been established that there is an equal distribution on gene bodies, transposons and intergenic DNA (28). From all the CpGs in the human genome, those that are unmethylated regardless of gene expression constitute only 1-2% of all CpGs and they are located mostly in CpG islands (CpGi). CpGi are areas with higher-than-normal density of CpGs and they are mostly located near or at gene promoters (33). The physical proximity between CpGi and promoter regions agrees perfectly with the notion that changes in DNA methylation can lead to changes in gene expression (28, 34). Moreover, it has been proposed that when CpGi regions become methylated during development, the gene associated is silenced permanently (32). Furthermore, Quante and Bird (2016) (35) recently suggested a connection between DNA sequence and DNA methylation.

Such connection was established due to the observation that sequence-specific proteins such as those with CXXC domains, because they recognize unmethylated CpGs, they are preferentially recruited to CpGi, which in turn preserves their unmethylated status. The revolutionary view that sequence and epigenome are intertwined has been supported by some studies, but has not been established yet (35).

Apart from CpGi, other genome regions have been shown to have specific methylation patterns. Such regions are linked to tissue type and are therefore named tissue-specific DNA methylation regions (tDMRs). Initially it was thought that tDMRs would be more prevalent near gene promoters, in or around CpGi and thus research was focused on such loci. However, in 2012 Fernandez and colleagues (36) looked at the methylation status of 1505 CpGs for 434 DNA samples donated by healthy individuals and determined that transcription start sites, and not promoter regions of genes are the DNA regions more predominantly linked to tDMRs. The discovery that transcription start sites coincide with tDMRs was further corroborated by an array study that interrogated more than 450,000 individual CpGs reported by Slieker *et al.* (2013). In this study, Slieker and colleagues confirmed that tDMRs are commonly associated with CpG-poor regions, not CpGi and are intimately linked to alternative transcription sites (37).

Even though the specific mechanisms by which the distribution of DNA methylation is established in the embryo and maintained throughout the organism's life are not known, most of the enzymatic pathways that catalyze methylation have been studied and are related to different family members of the DNMTs. The *de novo* methylation established in the embryo after fertilization occurs due to the action of DNMT3A and DNMT3B which catalyze the addition of a methyl group to cytosines. The

23

methyl group is donated by s-adenosyl-L-methionine which specifically functions as a

methyl donor (38). In more detail, DNMT3A seems to be highly expressed in germ cells

whereas expression of DNMT3B peaks in early development following fertilization (39).

Both DNMT3A and 3B contain a c-terminal catalytic domain and an N-terminal

regulatory domain. Next to the regulatory domain a conserved motif Pro-Trp-Trp-Pro

(PWWP) has been described, which appears to provide a link between the DNMTs to

histone tails that show particular types of modifications. For example, a point mutation

on the PWWP motif reduces the capacity of DNMT3B to recognize and bind the

H3K36m3- modified nucleosome, leading to a loss of methylation which in turn leads to

the pathological phenotype called Immunodefficiency Centromeric instability and Facial

anomalies syndrome (40). A third enzyme – DNMT3L – co-localizes with both

DNMT3A and 3B and works as an enhancer of the activity of DNMT3A and 3B. DNA

methyltransferase 3L does not possess a PWWP motif, and even though it has a catalytic

domain, it is unable to catalyze the methylation reaction on its own. Instead, two

monomers of DNMT3L and two monomers of DNMT3A form a tetramer complex that

catalyzes the methylation reaction. Because of the size of the catalytic tetramer,

periodicity of *de novo* methylation occurs on every 8-10 base pairs (41). In summary, the

factors determining which CpGs will be *de novo* methylated leading to a functional

methylome are the spacing necessary for the tetramer DNMT3L:3A to bind and the

histone modifications ability to bind the PWWP domains.

After the establishment of the methylome, maintaining the same patterns of DNA

methylation in healthy cells and tissues relies on the genome's ability to methylate

specific CpGs after each cell division. Before cell division and after the DNA sequence is

duplicated, it becomes necessary to copy the methylation pattern to the newly formed strand. The enzyme responsible for the addition of methyl groups on specific cytosines of the newly formed strand is called DNMT1. The specific mechanism depends on a 30-40 fold preference of DNMT1 to hemimethylated DNA. Hemimethylated DNA is DNA where only one strand (the template strand) has a methyl group on the cytosine, whereas the complementary strand (the newly synthesized strand) has an unmethylated cytosine. The catalytic domain in DNMT1 is longer than that in the DNMT3 family and it also contains a motif that binds to the replication machinery (specifically to PCNA). The proper functioning of DNMT1 depends on an accessory called UHRF1 (ubiquitin-like, containing PHD and RING finger domains 1). In more detail, the SRA domain within the UHRF1 flips the DNA strand open and exposes the hemimethylated site, which in turn allows the unmethylated cytosine to fit into the catalytic pocket of UHRF1 and therefore be methylated (41). Somatic cells are rich in DNMT1 and poor in DNMT3A and 3B which further establishes the importance of such methyltransferase in maintaining the methylome in post-embryonic development. However, the division of roles in DNMTs is not clear-cut. Some results have shown that DNMT3A and 3B can aid the maintenance of methylation patterns in somatic cells (39). The experimental design showed that when embryonic stem cells that lack DNMT3A and 3B expression are subject to successive *in vitro* multiplication, a decrease in DNA methylation is observed with cell division. The embryonic stem cells contain healthy levels of DNMT1, which should be sufficient to maintain the DNA methylation pattern even if no DNMT3A or 3B are present to perform *de novo* methylation. After these observations, it has been suggested that the presence of DNMT3A and 3B is essential to aid DNMT1 in the maintenance of DNA methylation

patterns (39). The interplay amongst DNMTs and other factors creates a pathway which define and maintain stable DNA methylation patterns in somatic cells.

In mammalian development, the male genome is actively demethylated within hours of implantation, whereas the female genome is only passively demethylated with each successive cell division (30, 31). Demethylation is therefore another important biochemical process necessary to maintain proper levels of methylation and to allow successful development of embryos. Initially it was thought that demethylation occurred uniquely in a passive manner when successive cell divisions occurred without being accompanied by an adequate increase in DNMT activity. However, such passive loss of methyl groups could not account for demethylation of the entire genome post-fertilization, especially since at that stage of embryonic development there is no cell division. For that reason, another mechanism was proposed and it was called active demethylation (42). Nowadays, both mechanisms are generally accepted and research is focused on determining at which stages of cell division or development one is prevalent over the other.

Passive demethylation occurs when the DNMT1 fails to recognize, bind or process a hemimethylated site and thus the newly formed strand is kept unmethylated. When the next cell division occurs and if not corrected, the copy of that strand will perpetuate the unmethylated state in that particular CpG. One example of how DNMT1 may not be able to bind DNA is if the binding site for DNMT1 on the DNA strand is occupied by a transcription factor. Unmethylation of DNA in this fashion may be a controlled event in the cell or tissue as a way to respond to specific gene expression due to the encounter with a new environment (28).

Active demethylation on the other hand depends on enzymatic reactions catalyzed

by TETs (ten-eleven translocation). The TET family of enzymes is composed of three

members thought to be involved in embryonic development, meiosis, stem-cell

reprograming, maintenance of imprinting as well as maintaining the correct DNA

methylation patterns which leads to proper gene expression (28, 43). Such enzymes

successfully oxidize 5-methylcytosines (5-mC) to 5-hydroxymethylcitosine (5-hmC), 5-

formylcytosine (5-fC) and 5-carboxylcytosine (5-caC). Experimental data has not

established yet whether these oxidized forms of 5-mC are only intermediates on the

demethylation process or if they have an active role on the genome. However, if indeed

they are mere intermediates in an enzymatic reaction, then it would be extremely difficult

to find stable forms of such oxidized methylcytosines *in vitro*. However, there are

sufficient numbers of those methylcytosines to permit their quantification. Genome-wide

studies show a lower abundance of these oxidized derivatives (less than 30%) when

compared to 5-mC (above 80%), but they are nonetheless a quantifiable presence in the

human genome (25). The fact that stable oxidized forms of methylcytosine are easily

found and quantifiable on the human genome may indicate a biological role yet to be

discovered (42). Another indication that 5-hmC may indeed have some undiscovered role

in the genome comes from the fact that the 5-hmC is not the TET enzymes' preferred

substrate. Moreover, complete demethylation via this pathway depends on Base Excision

Repair (BER) enzymes that can further treat the oxidized cytosine as a DNA lesion and

repair it to then become an unmethylated cytosine (Figure 1.3). Also the enzyme

thymine-DNA glycosylase which is an important part of BER-dependent demethylation,

is much more efficient at recognizing 5-fC and 5-caC when compared to 5-hmC. The

combination of both factors - the low level of oxidation of 5-hmC to 5-fC and 5-caC by

TET and the low excision and repair activity of 5-hmC by the BER glycosylase - can lead

to an increase in the presence of 5-hmC in the genome. At this point the role of 5-hmC

remains elusive but its presence in both dividing and non-dividing cells suggests a

potential role in gene regulation (42).
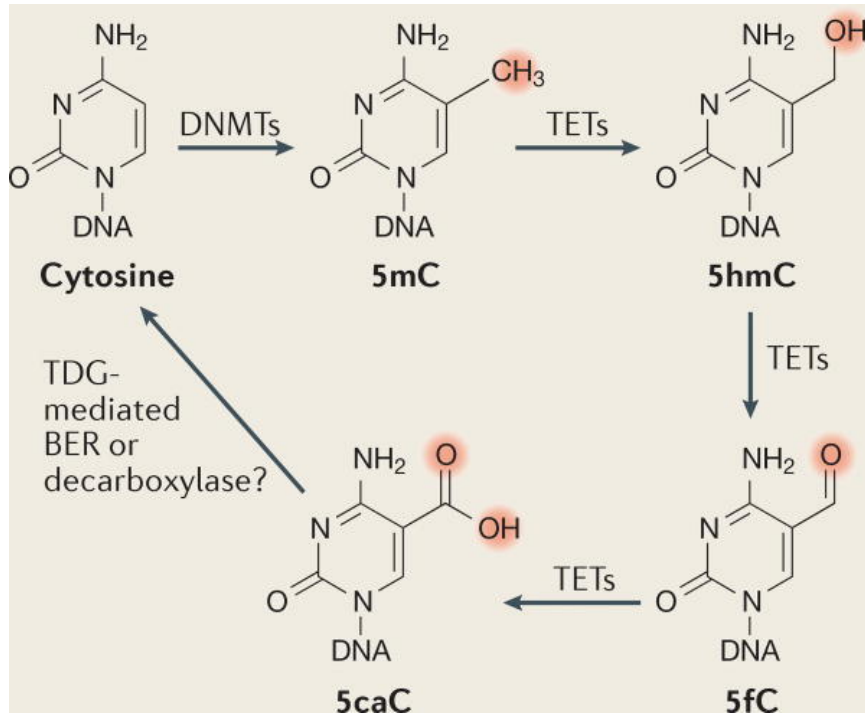
Figure 1.3 – Schematic representation of the enzymes and the chemical modifications they catalyze. Cytosine is modified to 5-methylcytosine by DNMTs. The methylated cytosine can undergo modification by one enzyme of the TET group which converts it to 5-hydroxymethylcytosine, which can be further oxidized to 5-formylcytosine (5-fC) and 5-carboxylcytosine (5-caC). The conversion back to unmethylated cytosine can be catalyzed by thymine-DNA glycosylase (TDG) – mediated base-excision repair (BER) or by decarboxylases. Adapted from (25).

The initial studies that linked DNA methylation with gene expression observed

that tDMRs are unmethylated on their specific tissues and highly methylated in other

tissues leading to gene silencing. Comparisons between methylation in DNA and mRNA

levels also showed that the expressed genes are unmethylated. One of the first mechanisms proposed for gene silencing as a result of DNA methylation included the inability of transcription factors to bind methylated DNA, thus preventing the transcription of such gene (31). Alternatively, other proteins called methyl-CpG binding proteins specifically bind methylated CpGs and are mediators in gene silencing (32). Even though such mechanism may be present for some genes and may constitute the main mechanism for transient gene expression, most transcription factors do not have CpGs in their binding sites which indicates the existence of other mechanisms.

Research on gene expression has demonstrated that DNA methylation may affect chromatin packaging. With most of the methylated CpGs located in the core center of the nucleosome and not on the DNA in between nucleosomes, the chromatin adopts a condensed state. Moreover, the recruitment of proteins capable of modeling chromatin structure may depend on the presence of methyl groups in the DNA. Combined, these mechanisms explain better the observed phenomena of some genes being transiently silenced and others keeping their repressed state through the entire lifespan of the organism. Certain parts of the genome can undergo changes in methylation status after the embryonic stage of implantation. One of these changes involves the loss of the pluripotent state of embryonic cells as they get differentiated in specific tissues. Another major event is the inactivation of one of the X chromosomes in females. In both cases modifications of the chromatin structure as well as changes in gene expression and replication timing occur prior to de novo methylation of CpGs. The observation that DNA methylation occurs after chromatin structure modification at certain genome

locations suggests that DNA methylation locks the gene repression phenotype and is not responsible for causing it (31).

During the lifetime of an organism, changes in DNA methylation occur resulting from exposure to the environment and as a direct consequence of molecular ageing. The epigenome of an organism is a dynamic system that allows the modulation of the molecular environment due to factors present in the external environment. Several bodies of research have focused on the impact of changes in the environment and its consequences on DNA methylation or chromatin remodeling (44-46). Environment constitutes everything surrounding an organism, including diet, smoking habits, physical exercise, pollution, contact with known carcinogenics, etc. Some of those factors are known to promote general health and increased life-span in humans (type of diet and physical exercise), however it becomes difficult to pinpoint the exact molecular mechanisms that affect an individual's health. For example, the regulation of folate in diet can alter the levels of s-adenosylmethionine, which is the main methyl donor for the DNMTs, thus changing the levels of DNA methylation. Also, the hypomethylation of genes related to inflammatory processes is delayed when moderate physical exercise is adopted in the lifestyle of an individual (47). Additionally, ageing is a natural phenomenon for all living organisms, and several studies have looked at DNA methylation as a prediction of biological age with varying degrees of success (48-52). However, biological age itself is dependent upon habits of each individual that range from diet to physical exercise. At the molecular level, it is generally accepted that any change in DNA methylation is transient until it gets locked-in by chromatin modifications that can be transmitted to the next generation. Since in the gamete formation there are

waves of demethylation and *de novo* methylation, most of such modifications are erased from the genome (30).

Interestingly, recent studies looked at the epigenetic modifications of the offspring of women that were pregnant during the Dutch Hunger Winter and a link was found between changes in insulin resistant cells and epigenetic modifications. Even though studies such as these are limited due to availability of cohorts, research in mouse models has demonstrated a link between changes in DNA methylation and behavior which can be perpetuated to later generations. For example, pregnant rats in the presence of predators, display high levels of stress and show a decrease in grooming and nurturing activities towards their pups. At the molecular level, it has been demonstrated that such behavior causes an increase in DNA methylation at the glucocorticoid receptor gene which is directed by a low level of serotonin in the pup's hippocampus. The levels of glucocorticoid receptors are maintained throughout the lifespan of the pups which will result in the same lack of nurturing behaviors towards their offspring (53). In humans the complexity of behaviors increases greatly since they depend not only on nurture but also on acquired information during adult life, which may further influence the levels of epigenetic modifications. The increased complexity observed in humans can create difficulties in linking molecular phenomena with phenotypic or behavior.

With respect to the exposure to pollutants and other detrimental environmental factors, changes in the epigenome will become stable with the persistence of the environmental insult. As seen in the embryonic studies, some genome regions are greatly protected from epigenetic changes thanks to the presence of condensed chromatin states and the binding of proteins to the DNA. For that reason, tDMRs are expected to maintain

a stable level of DNA methylation throughout the lifespan of an healthy individual (44, 45, 53).The cause-effect relationship of changes in global methylation as a result of the exposure of environmental pollutants or a specific diet are difficult to prove. The difficulty in proving such correlation is mostly because these changes are often subtle and difficult to be directly observed by the appearance of an external phenotype in humans (32).

As a consequence of the link between DNA methylation and disease, most research is focused on the understanding of how changes in DNA methylation impact disease. As mentioned before, one such diseases is cancer, which is also linked to age. In fact, the mechanisms of age have been the topic of many research papers and it is generally accepted that aging is characterized by an increase of damage to an organism, whether because of a decrease in adaptability or as a result of accumulation of 'mistakes' in the genome. The accumulation of mutations can result from the presence of reactive oxygen species or because of alteration of DNA repair mechanisns, such as is the case of Werner or Cockayne syndromes where patients appear older than they truly are. Also, the telomere shrinkage due to a decrease in the action of telomerase has been pointed to as an additional cause of aging in humans. Even in the case of monozygotic twins, which share the same DNA sequence, the process of aging is different, further emphasizing the role of environmental factors in the process (47, 50, 54). Interestingly, aging seems to lead to a global hypomethylation and localized hypermethylation in the human genome. The regions hypomethylated are mostly those present within transposable repetitive elements, such as Alu and LINE, and CpG-rich promoters of imprinted genes. The hypermethylation seems to occur mostly in CpGi in promoter regions and in genes

related to cell apoptosis and senescence. Additionally, the link between aging and heterochromatin alterations has been known for over 20 years. However, the exact mechanisms are still unknown. As more studies unveil the links between DNA methylation, histone modification and heterochromatin stability, the picture will become clearer on which molecular pathways contribute to aging. Nonetheless, an increasing number of studies link altered levels of DNA methylation to age in humans, even allowing the prediction of biological age (48, 55-57). However, in such studies the accurate prediction of age is dependent on an initial identification of tissue of origin of DNA, which indicates that DNA methylation differences between tissues are higher than DNA methylation differences between the same tissue collected from individuals of different age (47, 58).

CHAPTER II - FORENSIC DNA ANALYSIS

A. History and Advancements in Technology

In 1985 a publication in the journal Nature denoted that the analysis of several tandem repeats on the DNA sequence of humans "…are a powerful method for paternity and maternity testing, can be used in forensic applications and might also be useful in detecting inbreeding…". In the same article, Dr. Alec Jeffreys called these sequences "DNA fingerprints" (59). Following this discovery, 'DNA fingerprinting' was first used in a murder case the U.K. by Sir Alec Jeffreys to help determine who committed the rape and murder of two young girls. DNA was able to connect both crimes showing that the body fluids found at the crime scene were from the same perpetrator. Dr. Jeffreys was the pioneer of 'DNA fingerprinting' and its use spread to Europe and all over the world (60). The DNA fragments, called fingerprints, are also known as RFLP (restriction fragment length polymorphisms) and were obtained by incubation of DNA with restriction enzymes that specifically cut the DNA in the same consensus sequences throughout the DNA strand of interest. In the USA the enzyme most commonly used was H*ae*III which recognizes and cuts the sequence 'GGCC' (61). The RFLP method as used by Dr. Jeffreys first extracts the DNA by lysing the cells and disrupting membranes and proteins attached to the DNA. Next, the restriction enzymes are incubated with the extracted DNA and the fragments they produce are separated by size through the use of an agarose gel. In order to determine which fragments were present, the gel product was transferred to a nylon membrane, the double stranded was then denatured by a strong alkaline solution and the bands hybridized with DNA probes. The specific base pairing between the bands

in the nylon membrane and the nucleotide sequence in the radiolabelled probes allowed

the detection of specific RFLP fragments. The membrane was placed next to X-ray film

and the location of the RFLP bands were photographed (62). In the original method, the

probes hybridized to several DNA bands creating a unique pattern that could be

compared with a control sample. However, since forensic samples often are composed of

mixtures, the bands would be the result of DNA from different contributors, which could

be difficult to analyze. The technology then evolved to use single-locus probes that

hybridize only to a single locus and give one band for homozygous or two bands for

heterozygous individuals (60). To analyze the presence of another RFLP, the process was

repeated by first removing the first single-locus probe by a alkaline solution washing, and

then hybridizing the membrane with new single-locus probe and a photograph taken. The

analysis of multiple RFLPs could take several weeks or even months to be completed.

Besides the low throughput, this method was also limited to samples with high yields of

DNA (>50 ng) and good quality and high molecular weight (>12 kb) DNA (63).

Kary Mullis had already invented the method referred to as polymerase chain

reaction (PCR) in 1986 (64). However it was not immediately used in forensics. PCR has

a great advantage since it mimics *in vitro* what occurs with DNA replication *in vivo*.

Briefly, the extracted DNA is incubated with a DNA polymerase, primers that flank the

region of interest to be copied, deoxyribonucleotides (dNTPs) and ions such as

magnesium which are necessary for enzymatic activity. The PCR product consists of

multiple copies of the region of interest, which is delimited by the primer sequences.

In 1990, some laboratories were using dot blot tests which work by hibridizing the

denatured PCR product with specific probes captured in a nylon membrane (65).

Colorimetric detection allowed the appearance of blue dots for positive hybridization. Even though there were issues with low power of discrimination of the method and the possibility that the PCR products would re-anneal prior to hybridization with the probes, the use of the method allowed the introduction of PCR-based technology in the court system.

Once the sequence of the regions used in RFLP were further explored, several attempts tried to perform PCR to amplify each of the RFLPs. In 1994, the D1S80 kit allowed the amplification by PCR of the locus D1S80 which spanned 400-800 bp, depending on the number of repeats present in the individual DNA sample (66). The PCR products were analyzed by a vertical polyacrylamide gel followed by silver stain detection. However, the relatively large length of fragments, made it still difficult to type samples with degraded DNA. The main limitations of the methods at that point were either the need to use long sequences of DNA making it problematic to analyze degraded samples and presenting a low power of discrimination. Most importantly, none of the methods studied were easily automatable or allowed a high throughput analysis.

Even though short-tandem repeats (STRs) were known since the late 1980's, in 1995 the United Kingdom Forensic Science service launched the use of 6 STRs for suspect match in forensic analysis (60). STRs are regions in the genome that show smaller repeat units (2 to 7 bp), therefore called 'microsatellites' when compared to the 9-100 bp found in the minisatellites typed with RFLP. In 1997, the FBI laboratory selected the 13 USA core STRs as the necessary loci to be typed for DNA forensic analysis and stored in a national database. Even though the small size of STRs compared to RFLP allowed the use of degraded DNA, the analysis step still relied on silver staining of

polyacrylamide gels thus not allowing high throughput. Another issue was that the same

DNA sample would have to be divided into 13 separate PCR reactions for each one of the

13 loci and their analysis was done separately on in groups of two or three. To overcome

such difficulty, a multiplex PCR reaction was developed allowing the amplification of

several STRs. Presently the multiplex amplification and analysis of 15 STRs allow a

power of discrimination of 1 in a trillion or greater (Table 1.1(67)).

Table 1.1 Summary of characteristics comparing RFLP and methods based on PCR, such as STR typing. Adapted from (67)

Characteristic	RFLP Methods	PCR Methods
Time required to obtain results	6-8 weeks with radioactive probes; ~ 1 week with chemiluminiscent probes	1-2 days
Amount of DNA needed	50-500 ng	0.1-1 ng
Condition of DNA needed	High-molecular-weight, intact DNA	May be highly degraded
Capable of handling sample mixtures	Yes (single-locus probes)	Yes
Allele identification	Binning required since a distribution of sizes are observed	Discrete alleles obtained
Form used in analysis	DNA must be double stranded for restriction enzymes to work	DNA can be either single stranded or double stranded
Power of discrimination	~ 1 in 1 billion with 6 loci	~ 1 in 1 billion with 8 to 13 loci (requires more loci)
Automatable and capable of high-volume sample processing	No	Yes

B. Methods used in Forensic Laboratories for DNA analysis

The forensic analyst performs analysis of available evidence left behind at crime scenes, in order to establish what cannot be directly observed. The analysis of such evidence may allow detectives to establish a timeline for the series of events that lead to the crime. Biological fluids, fingerprints, documents, electronic disks, letters, magazines, lost phone calls or text messages, are all examples of evidence that can be collected during an investigation. Regardless of their origin, all evidence must be handled according to established guidelines. Rules including chain of custody assure that the origin of such evidence is not questioned in court and also that its integrity has been preserved as much as possible. Similarly, all the methodology for handling, storing and analyzing evidence has been established in the form of validated protocols that forensic laboratories strictly follow.

B.1. Sample Collection

Several types of evidence can be collected as samples. For example, for DNA analysis even as little as 0.1 ng of DNA can in certain circumstances, provide important information (Table 1.1) for investigations. Because of the high sensitivity achieved nowadays, even touching surfaces can leave behind enough DNA to form a full or partial profile of an individual. Examples of samples that provide enough DNA to construct a profile are urine in snow, tears on a tissue and even fragments of contact lenses found in a vacuum cleaner bag (68).
The increase in sensitivity of methods for DNA analysis means that there is a higher probability that instruments can detect the contamination of evidence as a result of

improper collection and handling. To prevent such contamination, DNA evidence must be collected and preserved properly from the point of origin. The first responders to a crime scene may be law enforcement personnel; therefore training is important so that they follow strict rules while at the scene. People collecting the evidence should wear gloves and utilize equipment and tools that prevent the contamination of evidence with the collector's DNA. All evidence must be packaged in sterile bags, boxes or other appropriate materials. DNA evidence must be stored in paper bags since the moisture that gets trapped inside plastic bags can cause degradation of samples, if stored for long periods of time. All samples collected must be sealed inside the package used to transport them, dated and signed to maintain chain of custody (69).

A different type of sample is collected for DNA analysis - reference samples. DNA reference samples are collected from suspects, potential innocent donors to serve as a comparison to the DNA found at a crime scene. Reference samples can then be collected from convicted felons or arrestees to be kept in a national database or can also be collected from family members/ housemates of the victim and law enforcement personnel to exclude any potential contribution of their DNA to the samples collected at the crime scene (elimination standards). Typically the reference samples are collected by a specialized investigator via blood drawn or a buccal swab which consists of a sterile cotton swab rubbed against the inside of one's cheek. Since they are collected directly from the source, and not from different physical materials found at crime scenes, the DNA content is higher and the likelihood of degradation is lower since their integrity is protected while they are transported to a laboratory (70).

For all samples collected, it is important that they are fully dried and maintained in a refrigerated environment, diminishing the action of bacterial species, hydrolysis by humidity and the action of digesting enzymes such as DNases. When arriving at the laboratory, samples can be kept refrigerated, frozen or at room temperature until they get processed and analyzed (68).

B.2. DNA Extraction

Body fluids are the most common source of DNA from crime scenes. Body fluid samples arriving at a forensic DNA laboratory may have been deposited in different physical matrices. For example, a bloodstain collected from a crime scene may have been deposited in bed sheets or other type of fabric, it may have been left behind in a paper towel or it may be in a floor tile. Regardless of what material the biological sample has been deposited on, there are fundamental steps that are required for DNA extraction. As mentioned before, genomic DNA is inside the nucleus of the cell in a compacted form associated with proteins such as histones. The first step of any DNA extraction procedure is to disrupt the cell membrane so that the cell contents are free in solution. The lysis solution contains a buffer that maintains a proper pH for the proteinase K activity, which breaks down the proteins associated with DNA. Commonly, lysis buffers also contain detergents such as sodium dodecyl sulfate to aid the breakage of the lipidic cell membrane. The result of this step is a solution containing a mixture of all cell components such as proteins, DNA and cytoplasm and their constituents (71). The second step consists on the DNA purification or extraction, which renders the isolation of DNA from the proteins, lipids, ions and organelles in the lysis mixture.

40

Finally, the DNA is purified, consisting of a liquid form that can stored at -20 ºC or -80

ºC depending on the expected length of storage. The frozen extracted DNA can last for

years and because it is kept in small tubes (of about 1.5 mL of capacity or lower), the

storage space necessary for each sample is small compared to the space necessary to store

the initial evidence (72).

The purification step can be performed by several different methods. The methods

presented here are the ones that were used for the project described in this dissertation.

One of the methods used was the Phenol-Chloroform extraction. The method is

also called organic extraction because it uses the properties of organic-aqueous mixture to

separate DNA from other cell components. The phenol-chloroform extraction is one of

the most popular DNA extraction methods and it has been used in forensic sciences since

the beginning of DNA analysis (73). A phenol-chloroform-isoamyl alcohol mixture

(25:24:1 v/v) is added to the aqueous medium containing the lysis product and following

centrifugation, the DNA moves to the aqueous phase, whereas other cell components

move to the organic phase. The aqueous phase is then removed from the top of the tube

and purified with ethanol precipitation or filters. The use of hazardous reagents and the

fact that PCR-inhibiting phenol residues can potentially remain on the extracted DNA

constitute two important limitations of this method (74, 75).

The other method used is called purification by affinity or solid-phase extraction.

The novelty of these methods when compared to the phenol-chloroform extraction is the

fact that the separation of DNA from the cell debris is performed through the use of

affinity substrates in solid form that specifically bind to the DNA. The binding does not

depend on solubility of DNA in the aqueous phase, but instead it depends on the negative

charge of DNA when bound to a silica-based column. The binding of DNA to the silica

column depends on several conditions. First the presence of chaotropic salts such as

guianidinium chloride at pH 6 in the lysis buffer causes protein denaturation and a

disruption of stabilizing forces on the DNA molecule, such as hydrogen bonds and van

der waals forces (76). Second, the pH of 7.5 added to the lysis solution, increases the

adsorption between the silica membrane and the DNA (72). The increase of entropy

observed by the removal of solvated water on the DNA molecule and also the increase of

hydrophobic forces between DNA and silica, allow the specific binding of DNA, whereas

the other debris in solution are eluted through the membrane (77, 78). After several

washing steps, an alkaline buffer is added to the column and the DNA is eluted. Certain

commercial kits contain silica coated magnetic beads instead of a fixed column which

allows the automation of the extraction procedure increasing sample throughput.

B.3 DNA Quantitation

Determining the amount of DNA that is present in the extracted solution is

important to obtain optimized results for analysis. If too much DNA is added to the PCR

reaction, the reaction may not be successful or peak overloads will occur when the

samples are analyzed by capillary electrophoresis. Amounts between 0.5 to 2 ng of

human DNA added to the PCR reaction usually lead to good results in electrophoresis if

the DNA is not badly degraded (79).

There are several things to take into consideration when selecting a quantification

method. First, samples collected in crime scenes may contain a mixture of DNA from

several species, for example, the presence of DNA from pets or bacteria deposited in the

surface prior to the crime. Also, DNA from certain body fluids like saliva and vaginal fluid, contain bacterial and yeast DNA as a consequence of the commensal microbiota population that inhabit the mouth and female reproductive tract. The DNA from blood collected as a reference sample, on the other hand, should not contain non-human DNA. Since forensic laboratories work with both crime scene samples and reference samples, it is simpler to use a single quantitation method that is human specific. For that purpose, several commercial kits have been developed and are currently used in forensic laboratories in order to quantify human DNA (80-82).

Another consideration to take is the quality of DNA from crime scenes. If the DNA was exposed to the elements for too long it may be fragmented. A quantification method will perform better if it can provide the analyst with an idea of DNA quality and not just quantity (83). Because of the specific limitations of forensic sciences and the methods used, quantification kits have been optimized in order to standardize the methodology between laboratories. The optimization of such kits together with the amplification kits for PCR allows a forensic laboratory to validate protocols for processing DNA in order to obtain a strong enough signal on the capillary electrophoresis (79).

Typically the quantification kits in forensic laboratories work with a real-time PCR instrument where the samples are ran as unknowns (unknown concentration) and standard solutions with known concentration of human DNA are ran as standards to calibrate the reaction (84). The number of PCR cycles after which the fluorescence increases exponentially is called cycle threshold (C_T). Standards with higher concentration of DNA will have a lower C_T since fewer cycles are necessary to observe

43

fluorescence from the PCR products. A graph that plots the C_T versus the log of the

concentration of the standards is determined and by linear regression, the concentration

of the samples is established (Figure 2.1, (79)).
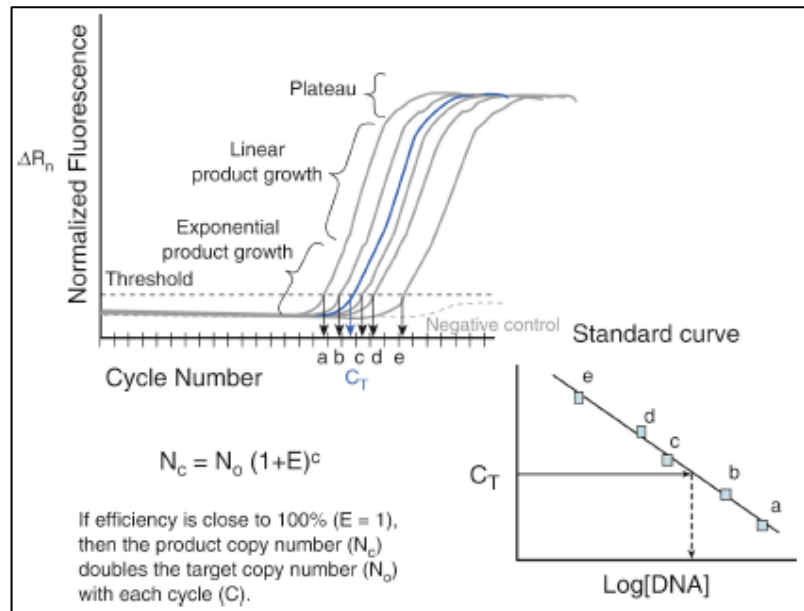
Figure 2.1 – Schematic representation of the calculations performed for DNA quantification using methods based on real-time PCR. The increase of fluorescence as a result of increasing copies of DNA are displayed in a graph of normalized fluorescence versus cycle number (CT). On a second graph the CT is plotted against the Log of the concentration of DNA for the standard solutions (with known DNA concentration). The concentration of samples can be extrapolated from the linear regression line and the efficiency is also calculated. Adapted from (79).

Using real-time PCR for quantification relies on the assumption that the reaction

conditions are similar for the standard solutions of DNA and for the DNA extracted from

the crime scene samples, however, that may not always be the case. For example, if the

primers anneal to a region of DNA to produce a lengthy PCR product (>300 bp),

degraded DNA may not have fragments of that size, which would cause the amount of

DNA in the unknown samples to be underestimated due to degradation. Most kits

overcome this limitation by including primers that produce amplicons of different sizes,

thus providing the analyst with information about DNA quality (80, 85).

Another difference between samples and standards is that samples may include a mixture

of DNA from several individuals, for example in sexual assault cases a mixture of male

and female DNA. To overcome some of these issues, commercial kits were developed in

order to quantify autosomal and Y-chromosome DNA. For example, Quantifiler Trio (85)

(Thermofisher Scientific, USA) uses 3 sets of primers; one that amplify longer DNA

sequences (LA, Large Autosomal, 214 bp); another set of primers that produce smaller

amplicons (SA, Small Autosomal, 80 bp) and a third set of primers that amplifies 75 bp

region on the Y-chromosome. All three primers amplify several regions on the autosomal

or Y-chromosomes so that higher detection is achieved for samples with small amount of

DNA. In order to detect the amount and type of PCR product (or amplicon) formed

during the PCR, fluorescent probes are added to the PCR mixture together with the

primers. The probes are called TaqMan, which consist in small oligonucleotides that are

complementary to the target region located between where each forward and reverse

primer anneal (Figure 2.2). The probes contain a fluorescent dye on the 5'-end and a

quencher on the 3'-end. When intact, the quencher is close enough to avoid any

fluorescence to be emitted by the probe. With the progression of the PCR reaction, the

polymerase which has a 5'-3' strand displacement ability, is able to perform similarly to

what is observed *in vivo* with Okazaki fragments, destroying the probe with the results

that fluorescence is no longer quenched. The increase in the number of amplicons results

in an increase of fluorescence detected by the real-time instrument.
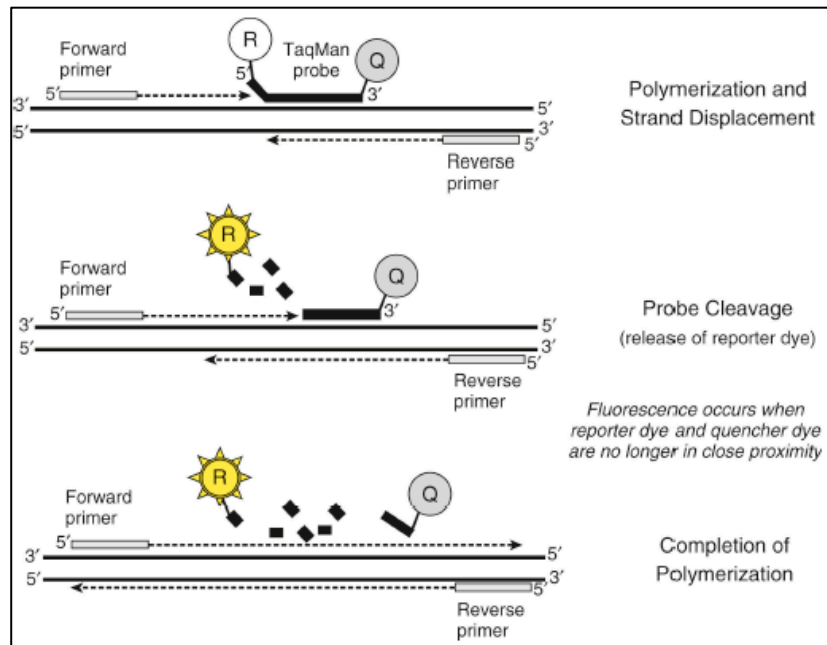
Figure 2.2 – Schematic representation of TaqMan probes and their mode of function. Adapted from (79).

Another commercial kit used in forensic laboratories is the Plexor HY (Promega, USA)

that contains one set of primers for several copies of a 99 bp autosomal region on

chromosome 17 and another set of primers for multiple copies of a 133 bp on the Y-

chromosome (82). The fluorescence detection chemistry is different from the TaqMan

since there is no use of a probe to produce and quench the fluorescence (Figure 2.3).

Instead, the primers are designed to have a modified nucleotide (iso-dC) at the 5'-end that

is labeled with a fluorophore. As the amplification progresses, the polymerase will add an

iso-dG which is the only nucleotide capable of making hydrogen bonds with the iso-dC

on the template strand.
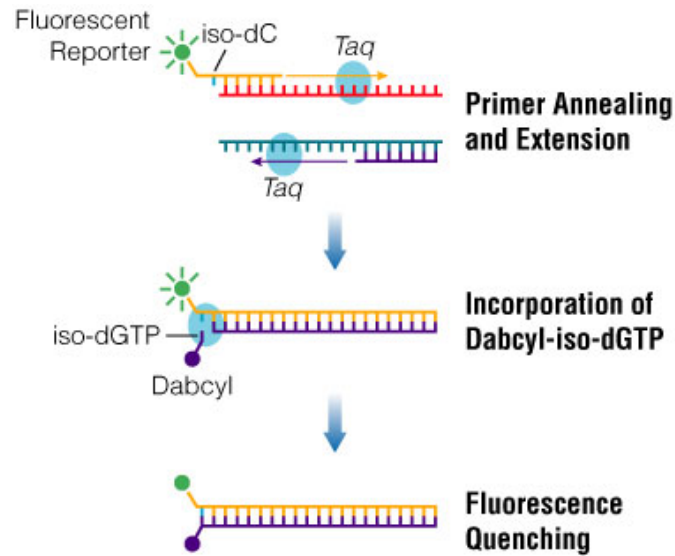
Figure 2.3 – Schematic representation of how Plexor chemistry fluorescently labels each amplicon produced from primers tagged with the same fluorophore. Adapted from (82)

The iso-dG contains a quencher dye thus a decrease in fluorescence is observed with the increase of number of amplicons in the PCR reaction. The unique chemistry of Plexor also allows the analysis of melt curves which are able to determine if there is an inhibition of the PCR reaction (86) or the formation of unspecific products, which will have a different melt curve compared to the target amplicon.

A third factor that can be different between DNA extracted from crime scene samples and standard DNA is the presence of inhibitors that are co-extracted with the DNA. While it is more common to find co-extracted inhibitors when the phenol-chloroform is the method of extraction (phenol itself is a PCR inhibitor), silica-based procedures can result in inhibition for certain type of inhibitors especially if present in

high quantities (87). To tackle this issue, an internal control is added to the PCR mixture together with the samples. If an inhibitor is present, the amplification of the internal control is also impaired and the analyst is able to determine the best option of PCR additives before performing the amplification that will be used for STR analysis (88).

Even though the use of commercial kits is highly advantageous in forensic laboratories, in the scope of this project, using commercial kits for quantification of DNA would be prohibitively expensive. For that reason, two main protocols were used to quantify DNA. Both protocols use a standard curve that matches fluorescence units to concentration. The Alu Quantification protocol uses amplification by a polymerase in the presence of an intercalating dye such as SybrGreen (89), whereas the PicoGreen method uses an intercalating dye to quantify the presence of DNA without any amplification performed(90, 91). The AluQuant method is a primate-specific method since the primers used only anneal to regions in the primate DNA. The primers selected amplify numerous Alu repeats that are widely present in the human genome. The fluorescence increases as the amplification progresses due to the capacity of the intercalating dye SybrGreen to fluoresce intensely when intercalated in the double stranded DNA. The background signal of SybrGreen fluorescing because of thr intercalation in the template double-stranded DNA is normalized by the instrument that uses a low gain of signal settings based on the fluorescence read at the beginning of the run. As the amplification continues there is an increase in DNA concentration in the reactions because of the accumulation of amplicons. The standards can then be plotted similarly to what is observed in Figure 2.1 and the concentration of the DNA in the samples is determined (92). The AluQuant method provides reliable quantification and its use is adequate to the research done in this

project because all the samples used were collected from volunteers and not from crime scenes. Considering the origin of samples, we do not expect DNA degradation or contamination by inhibitors present in the initial sample. Regarding potential inhibitors such as phenol carried over from extraction, the analysis of the melt curve as well as the steepness of the curve allows DNA analysts to determine if significant inhibition is present (93).

The PicoGreen method measures fluorescence of all DNA present in the samples, regardless of its origin. Because it does not rely on primers or amplification, the fluorescence observed is merely the result of how much fluorescence is emitted by picogreen (an intercalating dye), which is proportional to the amount of genomic DNA present in the sample (90, 91). For samples such as blood collected in sterile conditions, this method represents a fast, reliable and cheap option to quantify genomic DNA. Similarly to other quantification methods, the fluorescence measured in the samples with unknown concentration of DNA is compared to the fluorescence of a standard curve and the concentration is calculated by extrapolation.

B.4 Polymerase Chain Reaction (PCR)

Polymerase Chain Reaction somewhat mimics what happens *in vivo* during DNA replication. A designated set of primers added in solution establishes the DNA sequence to copy. The other components necessary for PCR are the extracted DNA template, the DNA polymerase, dNTPs, magnesium and a buffer capable of keeping a stable pH for a wide range of temperatures. Some commercial kits used in forensic laboratories also include additives that minimize the effects of inhibition present in the samples.

One of the first thermo-resistant DNA polymerase used in PCR was extracted from the bacterial species *Thermus aquaticus* (Taq) which inhabits in hot springs. The Taq resistance to high temperatures is important because in PCR, the DNA is firstly melted with an increase in temperature which would degrade most DNA polymerases (94). Nowadays the polymerases used for PCR are still called Taq polymerases, even though most of them are genetically modified enzymes descendent from Taq. Another important characteristic of Taq polymerases used in forensic sciences is the fact that they require an activation step, usually performed at the beginning of the PCR reaction. The need for an activation step prevents unspecific amplification by the polymerase at room temperature. Enzymes that require the initial activation step are called Hot Start polymerases (95).

Despite of the differences in polymerases developed due to biotechnological advances, all polymerases need magnesium to properly amplify DNA, therefore an optimal concentration of magnesium is necessary in PCR (96). The primers commonly used are small oligonucleotide sequences (18-32 bp) that are complementary to the flanking regions of the DNA sequence to be copied. For each sequence to copy two primers are necessary – a forward and a reverse primer – both flanking the beginning and end of the region to be amplified. Typically the primers should have a GC (guanine-cytosine) content around 50% and a melting temperature of around 60 ºC (95). The melting temperature (T_M) of a DNA template or of a primer:template region is the temperature necessary to provide to the system in order to cause the separation of half of the primer from the template DNA, respectively. The melting temperature depends on the GC content and the length of the primer. Primers offer the 3'-OH necessary for polymerase recognition of a substrate and to perform the extension of the strand copying

50

the template by addition of dNTPs to the nascent strand. For STR analysis, the primers anneal either to microsatellite regions of the genomic DNA and some to specific regions of the Y or X chromosome.

The PCR reaction is carried on through several steps, namely denaturation, annealing and extension. Each step has a designed temperature and duration, which are precisely achieved by instruments called thermocyclers. The design of thermocyclers allows each PCR tube to be surrounded by metal that quickly transitions its temperature at a rate of several degrees Celsius per second. The tubes are composed of thin plastic that allows an improved conductance of hear to the entire volume of the PCR solution. During denaturation, the solution is heated to 95 °C or above to ensure that the double stranded DNA is melted to single stranded DNA. This melting of DNA mimics *in vitro* the unwinding of DNA by the replication machinery that occurs *in vivo*. The next step, called annealing, happens at an optimal temperature determined by the primer sequence and length. Typically, the annealing temperature can be set on the PCR instrument as 5 °C below the T_M of the primers. At this temperature the primers anneal to the template DNA thus creating the template:primer substrate for DNA polymerase. The instrument will then raise the temperature to 72 °C which is the optimal temperature for Taq polymerase to extend the primers and form a new copy of the template. The cycle of temperatures is repeated several times until several copies of the same sequence have been made (95). Each double stranded PCR product will denature in the next cycle and serve as a template for the polymerase. For that reason, in a reaction that is occurring in optimal conditions, the number of PCR products increases exponentially with the number of cycles. For the first cycles, the DNA serving as template is the genomic DNA added to

51

the mixture. With the increase in PCR products, the template DNA soon becomes the PCR products themselves. The fact that PCR products from the initial cycles become a template is an important issue in the use of PCR in forensic laboratories since any amplification errors made in the first few cycles are perpetuated until the end of the PCR (97).

In forensic laboratories several DNA regions are amplified in one reaction, which is called multiplex PCR. The simultaneous amplification of multiple DNA regions is possible through the use of primers which have a similar T_M and are designed to not anneal to each other. Since the primers are labeled with fluorophores, the PCR products corresponding to each set of primers can be distinguished from each other after PCR and during analysis. Besides the use of multiplex PCR, another feature of PCR in forensic laboratories is that it uses low amounts of DNA. When a low amount of template is added to the PCR, it is possible that not all primers will be able to anneal to it with the same efficiency. Therefore some products may be more predominant at the end of the reaction, causing an artifact known as allele dropout. If the template added to the PCR is degraded, meaning the genomic DNA is fragmented due to the action of extreme heat or other environmental factors, the polymerase may not have an intact template to copy from. The bigger the region to copy, the more likely amplification can be interrupted because of fragmentation at that locus. The potential lack of amplification of larger amplicons as a result of DNA fragmentation can cause an imbalance of STRs in the final PCR product if the small STRs are properly amplified (95). The presence of inhibitors that are co-extracted with DNA, if in high enough concentration, may decrease the PCR efficiency. The only way to detect inhibition is during quantitation with the use of an internal

control. Some forensic commercial kits contain additives such as bovine serum albumin (BSA) that helps to diminish the effects of inhibition (95).

The commercial kits developed for STR amplification were designed to take these limitations into consideration and therefore they require a certain amount of DNA to be used as quantified by specified methods. Before being commercialized most of those kits were tested by multiple laboratories to better predict the type of products they will produce, thus diminishing the risk of unspecific or unexpected results. However, since each DNA sample is unique because of its unique origin of crime scene, commercial kits continue to be developed in research every day.

B.5 Capillary Electrophoresis

The analysis of PCR products in forensic sciences depends on its separation by size. Electrophoresis is the method of choice since it allows the migration of PCR products based on size. The DNA is negatively charged due to its phosphate backbone and thus when an electric field is applied, it moves towards the positively charged pole. In gel electrophoresis, smaller DNA products will travel faster due to lower interaction with the gel, whereas bigger fragments travel slower and appear as bands closer to where the sample was loaded (98).

Both agarose and polyacrylamide gels can separate DNA by electrophoresis with different resolutions. Agarose gels cannot distinguish between PCR products of very similar sizes, whereas polyacrylamide allows separation of products with 1 bp difference. The resolution can be improved if the DNA is in its single strand form. Denaturing DNA to its single strand form can be achieved by addition of formamide or by loading it in a

denaturing gel. Denaturing gels have saturating concentrations of urea. Both urea and formamide form hydrogen bonds with DNA, preventing the complementary bases to anneal to each other (99).

The use of electrophoresis gels require time since each gel needs to be assembled, manually loaded with the samples and can only be used once. The low throughput of electrophoresis gels greatly impairs the usability of DNA in forensic sciences since a single crime scene may have hundreds of DNA samples retrieved as evidence. Additionally, gel electrophoresis may require a higher quantity of PCR product since the bands are visualized by naked eye and the loading of samples is limited to how little an analyst can pipette (99).

The use of capillary electrophoresis tackles the low throughput limitation since the PCR products are ran through a capillary that can be reused many times and the instrument can load each PCR product automatically, which allows for less sample usage. The capillary contains a polymer that acts similarly to an electrophoresis gel, separating the PCR products by size. The two ends of the capillary are immersed in buffers which have opposite electrical charges thus creating a current within the tube. Because the capillary is very thin, the heat caused by the increase of voltage is easily dissipated (100, 101). The PCR products pass through a detection window in the capillary which is placed near a laser. The passage of the fluorophores from the PCR products (from the labeled primers) through the window of the capillary, the laser excites them and a detector registers the fluorescence and time of passage. The detector is connected to a computer that has installed software capable of analyzing the data. A size standard ran at the same time as the PCR products allows the software to match time of passage through the detection

54

window with a size in base pairs. The software will then display each PCR product detected to a peak in a electropherogram. An electropherogram is a graph that displays peak height and width, which is proportional to the amount of PCR product, *versus* amplicon size. The units of an electropherogram are RFU (relative fluorescence units) *vs* size in base pairs (101).

Each PCR product is composed of multiple fluorescent products that originated in each of the primer pairs extended during the multiplex PCR. The capillary electrophoresis instrument can read several fluorescent products at the same time. Peaks with different fluorescence but same size will be displayed on the electropherogram as overlapping peaks. The software analysis includes the separation of peaks by fluorescent colors and the calling of peaks depending on sizes. Presently forensic laboratories use capillary electrophoresis instruments capable of independently running and reading several capillaries at the same time, greatly increasing the throughput of analysis (102).

Several commercial kits for multiplex PCR have been validated to use in forensic sciences (103). To allow analysis of PCR products, those kits were designed to have the STRs being amplified in a range of sizes per fluorophore color. Figure 2.4 displays all the possible repeat numbers for alleles of each STR. For example, the blue panel corresponds to a fluorophore that emits in the blue wavelength and because of the differences in each PCR product size, 4 STRs can be detected without overlapping peaks. There are approximately 12 possible alleles for D8S1179 according to population studies done in humans (104). Any of those 12 possible alleles are smaller in length (base pairs) than any of the 24 possible alleles of the D21S11, thus they appear first on the electropherogram. Since both D8S1179 and D3S1358 overlap in size, they were designed to amplify with a

55

different fluorophore colors so the software was able to separate them. In summary, the

analysis software indicates the number of repeats for each STR by analyzing the base pair

sizes for each PCR product and each color.
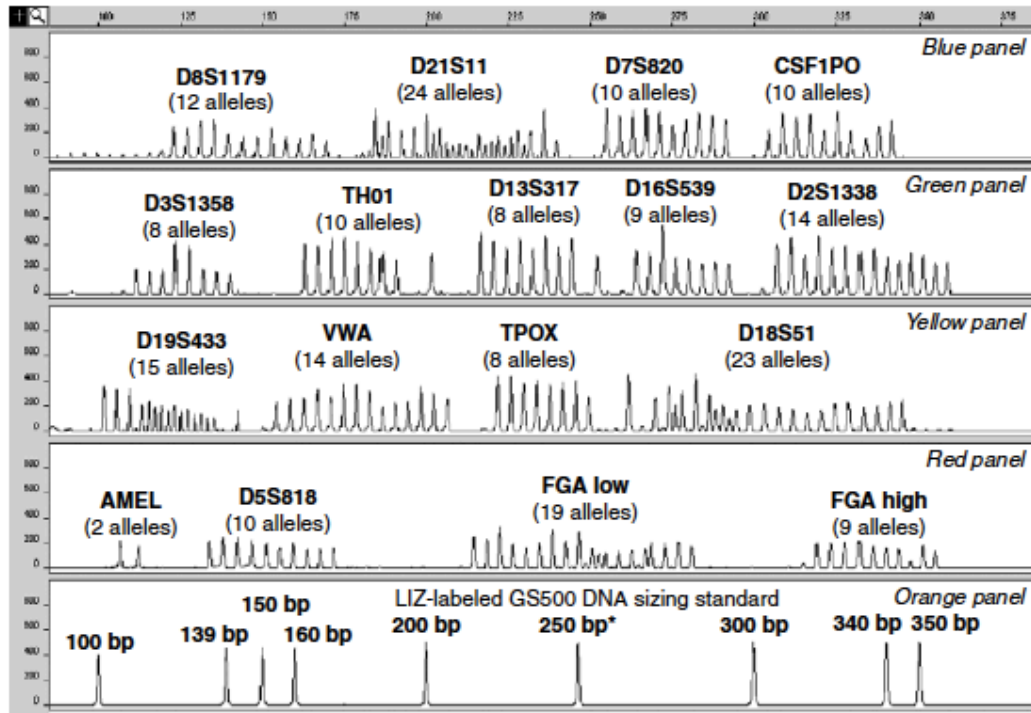
Figure 2.4 – Schematic representation of all possible alleles for 16 STRs plus the sex marker amelogenin. The STRs are displayed in electropherograms separated by fluorophore colors and sizes in base pairs. Adapted from (103)

An allele corresponds to the same region of DNA located in homologous chromosomes,

one donated from the mother and one from the father. Therefore a DNA sample that

originates in a single individual will have 1 or 2 alleles per STR locus, depending if the

person is heterozygous or homozygous for that particular locus. A person is said to be

heterozygous for an allele if the chromosome donated from the mother has different

number of repeats than the chromosome donated from the father in that specific STR.

For heterozygous individuals, the software will display 2 peaks with 2 different sizes

within the region of that specific STR. After the analysis of each STR, the software

produces a list of all number of repeats for all STRs, which constitutes the DNA profile

of that person (105). If all the steps in the STR analysis, from multiplex PCR

amplification to analysis with capillary electrophoresis are calibrated properly, two

heterozygous peaks in an electropherogram must have similar peak heights. The presence

of a third peak or unbalanced heterozygous peaks on a known sample indicates issues

with amplification, analysis or with the sample itself (106). If the sample is a known

control DNA that was commercially bought, its molecular weight and quality should be

high so that no fragmentation is expected and the issue is most likely in the multiplex

PCR or electrophoresis analysis. If the DNA is from a crime scene, the third peak or two

imbalanced peaks may mean that a second contributor is present in the sample. For that

reason, control DNA samples are run at the same time as the evidence samples and the

commercial products available for amplification and capillary electrophoresis are

carefully tested and pass rigorous quality control tests (107, 108). When the analyst trusts

the commercial products they are using, the appearance of artifacts give them the

confidence that what is seen in the electropherogram is most likely due to the quality of

DNA or the presence of a second donor, which allows the analyst to employ standard

operation protocols to re-analyze the sample according to the perceived issue observed in

the results.

Since the entire process from sample injection in the capillary to data collection by the

software has been automated, the analyst is left with more time to analyze at the

electropherograms and interpret the results to tackle potential issues with the samples.

There have been many advances in the technology for DNA analysis in the forensic sciences since its first use in 1985 by Sir Alec Jeffreys. Most importantly the quantity of DNA necessary to produce a profile has been greatly reduced. In addition, automation has greatly increased the number of profiles that can be obtained in a short period of time. It is quite remarkable how much improvement has occured within the life span of one generation. One of the most positive characteristics of DNA analysis for forensic purposes is the constant push of limits in the current technologies.

B.6 Methods for Body Fluid Identification

Samples collected from crime scenes can be of various origins. It is important to determine the exact location of a body fluid in a piece of evidence. Identifying the presence of a body fluid aids the analyst to determine where to extract sample for DNA analysis and determining the type of body fluid can help investigators determine the relevance of such evidence. A number of serological tests have been developed to determine the presence and origin of biological fluids from crime scenes samples (108). The Scientific Working Group on DNA Analysis Methods (SWGDAM) classifies the tests based on their ability to determine the presence or identify the type of body fluid. The SWGDAM has a Quality Assurance Committee that drafts revisions to the original FBI's Quality Assurance Standards that each forensic laboratory must meet to be part of the national DNA index system. The committee meets twice a year and provides recommendations to the FBI to update the quality standards as necessary. Furthermore, the committee publishes online the new recommendations thus their website (109) is the most up to date resource to determine what new standards will most likely be adopted in

forensic laboratories in a near future. According to SWGDAM, serological tests can be visual which allow the analyst to determine possible areas of where the biological fluids have been deposited, including the use of alternate light sources. Chemical tests use properties of the body fluid to change the color, fluorescence, solubility or other properties of a chemical used to react with the fluid. Histological tests are those that use a microscope to identify specific elements on the body fluid and immunological tests are the ones that make use of an antigen-antibody reaction to determine the presence of specific proteins in the fluid (107).

Alternative light source tests use light in different wavelengths such as the ultraviolet or infrared ranges which cause certain body fluids to glow or darken (for example, blood). Alternative light sources are mainly used to detect the presence of body fluid stains in large pieces of evidence and is not able to identify the body fluid present (110).

Two types of tests exist to determine which body fluid is present in a piece of evidence - presumptive and confirmatory tests. Presumptive tests can simply be used to tentatively determine a specific body fluid. They can be applied in large areas but they have limited specificity, meaning, other substances may react with the same test. Presumptive tests to determine the presence of blood generally rely on the detection of hemoglobin which is present in the red blood cells which are the most abundant cell types in blood. The Kastle Meyer test relies on the oxidation of phenolphthalein in the presence of hydrogen peroxide catalyzed by the hemoglobin present in blood, which causes the solution to turn pink (110). Another presumptive test is the use of Luminol which fluoresces in the presence of blood (68). Luminol consists in a mixture of 3-amino-

phthalhydrazide, sodium carbonate and sodium perborate, added immediately before the solution is used (111). The reaction occurs due to the presence of Iron on the hemoglobin and the solution can be sprayed in large areas. The UV light will then make the fluorescence visible and able to be photographed in a dark room. Luminol does not cause any inhibition of PCR therefore samples in contact with it can still be used for DNA typing.

Presumptive tests to determine the presence of saliva can be quite useful since unlike blood, saliva has no color and its presence may not be as easily detected. The presumptive test for saliva is called Phadebas and it relies in the detection of a salivary enzyme called alpha-amylase. Since this enzyme is also found in fungi, bacteria and saliva from other animals, this test is not specific to human saliva. Moreover, close variants of the saliva alpha amylase are also present in the pancreas and thus it can be secreted in other body fluids like urine (110, 112).

The presence of semen and seminal secretions is also difficult to determine without the help of presumptive tests. The tests used can detect proteins that are specific to semen such as acid phosphatase which is an enzyme 400 times more abundant in seminal fluid than in other body fluids (68). A dark purple color is observed when acid phosphatase is exposed to sodium alpha naphtylphosphate and Fast Blue B. The prostate specific antigen (PSA) test detects p30 which is a protein present mostly in semen but has also been observed in lower concentrations in other body fluids namely breast milk. Semen from azoospermic males are positive for p30 which makes this test useful to determine the presence of semen even if no sperm cells are present (113).

Confirmatory tests on the other hand, are designed to conclusively identify a body fluid and to not cross-react with other body fluids. They can be more expensive and their use be limited to smaller areas. Typically the portion of evidence used for confirmatory tests cannot be used for DNA analysis, which further limits the applicability of these tests especially when limited amounts of evidence are available.

Confirmatory tests to determine the presence of blood include the Takayama test where the formation of crystals occur when a solution of dextrose, sodium hydroxide and pyridine reacts with blood. The crystals are visible under a microscope and they exhibit a salmon-pink color (111). Despite of its broad use, this test is not human-specific therefore it has limited use in aiding the forensic analyst to determine whether to consider the stain a viable evidence to collect DNA from (114). The ABA card HemaTrace strip kits have an antibody against human hemoglobin which is mobile upon agglutination with hemoglobin in blood. The result of the test is visible as bands on a test window. Cases of false positives have been reported when samples of saliva where identified as blood using this test (115), which makes its reliability limited. The RSID™-Blood test is a commercially available test that detects human GlycophorinA but not hemoglobin. Glycophorin A is a protein present in the cell membrane of erythrocytes (116). The capacity of RSID™-Blood to detect other protein different than hemoglobin makes it not susceptible to the high dose hook effect present in some tests that detect hemoglobin. The high dose hook effect occurs when a test shows negative results because too much hemoglobin is present because of too much sample being loaded into the test (117). RSID™-Blood is commercialized as being blood-specific and not showing cross-reactivity with other body fluids. Recently RSID™-Blood was reported to display some

limitations such as not being able to detect blood after it has been cleaned with detergents that possess active oxygen agents. The test also showed false positive results with saliva samples collected from 15 individuals (115) and also presents low sensitivity when compared to presumptive tests such as luminol and others (118). In summary, confirmatory tests for blood available for forensic laboratories either render the sample useless for STR analysis and show high dose hook effect, or have lower sensitivity like the RSID™ and provide false negative in the presence of blood that has been cleaned or in contact with active oxygen detergents.

In 2009 Virkler & Lednev (119) reported the absence of confirmatory tests to determine the presence of saliva but mentioned RSID-Saliva as an emerging method. Instead of detecting alpha-amylase based on its activity, the RSID™-Saliva uses two monoclonal antibodies that specifically detect the presence of the enzyme. As a result, the test is more sensitive and does not rely on enzyme activity, which can be impaired by pH, temperature and other environmental factors. The test design consists of a lateral flow strip test that contains the two monoclonal antibodies which migrate in the presence of alpha-amylase. Validation experiments determined that the limit of detection for the test is 0.05 μL of saliva. The same report published by the company that commercializes RSID™ established that the test shows negative results when tested with a mix of urine, blood or semen at different ratios but it shows faint positive results when tested with breast milk. The test showed stability since it was able to detect saliva after being exposed to 37 °C for 11 minutes and 56 °C for 30 minutes, it also showed primate-specificity since it only reacted positive to gorilla saliva and also proved able to detect saliva in several forensically relevant items such as aluminium cans, plastic cups, etc

Another commercial test that detects the enzyme alpha- amylase is the SALIgAE®

which produces results in the form of a colorimetric solution. According to the

information available on the tests' website, the SALIgAE® requires 10 µL of saliva and

storage conditions should be around 2-8 ºC. No scientific literature has been cited by the

company indicating which validation tests were performed by them (120). However,

Meyers *et al.* (2008) determined that the SALIgAE® shows a lower sensitivity to saliva

since it is only able to detect it in dilutions of 1:10, whereas the Phadebas can detect

dilutions of 1:200 (121). Moreover, in the same study, the SALIgAE® showed positive

results for saliva swabs of guinea pig and weak positive results for saliva stains in

aluminum cans and other mock crime scene evidence. Another study aimed to compare

the RSID™-Saliva and SALIgAE® and it determined much lower sensitivities for both

(122). The RSID™-Saliva showed a sensitivity of 0.01 µL of saliva (122) in the report by

Pang and colleagues (2008), which is 5 times lower than the reported on the validation

studies (123). Overall, Pang *et al.* (2008) reported that the RSID™-Saliva shows superior

capacity to determine the presence of saliva whereas the SALIgAE® did not perform so

well. On the same report, both tests show false positives when tested with rat saliva,

human pancreatic alpha-amylase, stains of breast milk, liquid samples of urine (but not

stains of dried urine) and fecal swabs. It is important to mention that oftentimes

validation reports show different results for the same type of samples which may be due

to different preparation of samples. For example, on the report by Meyers *et al.* (2008),

the dilutions of saliva used to perform the sensitivity tests were dried in cotton fabric

prior to testing. It is not clear whether the samples were extracted from the cotton

material which would indicate further dilution. The limitations of both methods must be

taken into consideration when considering their classification as confirmatory tests and when reporting results used in the court of law.

The most widely test for the confirmation of semen is the microscopic observation of sperm cells after Christmas tree staining. The staining makes the head of the sperm cells red and the tails green allowing an easier observation (119). The two main issues that arise with this method is that the semen of azoospermic males do not contain sperm cells and the fact that microscopic analysis is time consuming. In order to detect the presence of semen even from azoospermic males and to do so in less time, the company Independent Forensics developed the RSID™-Semen strip test. Instead of identifying the protein p30 or sperm cells, the test has two monoclonal antibodies specific for semenogelin, which is a protein produced by the seminal vesicles. The validation report published by the company determines the strip test to be sensitive to 2.5 nL of semen but also gave false positives due to high dose hook effect when 3 µL of semen are added to the sample chamber. The RSID™-Semen showed specificity for human semenogelin and in the same report the test was able to detect semen in several mock crime scene samples where stains of semen were deposited in different materials and fabrics. Besides the high dose hook effect, mixtures of semen and some body fluids give false negative results and also the RSID™-Semen seems to show poor results in post-coital samples and diluted samples (124). Since it is impossible to know if evidence collected from a crime scene consists of a mixture of body fluids or how much semen is present, the RSID™-Semen test presents severe limitations for use in forensic evidence.

Even though vaginal secretions, urine and sweat constitute important evidence and are commonly found at crime scenes, there are no established presumptive or

confirmatory tests for the detection of vaginal secretion, urine and sweat. Some serology tests are available but are not widely used in forensic laboratories.

The acid-Schiff reagent selectively stains glycogenated epithelial cells present in vaginal secretions. However, the level of glycogenation depends on the menstrual cycle and age of the donor. False positives from samples collected from the mouth or urethra of males have been well documented for this test (119). In addition, the test is time- and sample-consuming; therefore its use is very limited. One other test such as the detection of isoenzymes 4 and 5 of lactate dehydrogenase which is merely presumptive since it is not specific for vaginal secretions. The limitations of these tests leave this forensically-important body fluid unable to be identified (125).

To presumptively determine the presence of urine, some colorimetric tests have been used. The detection of urea, creatinine, uric acid or urease can be of value to determine the presence of urine, however all the different tests used present limitations and false positives with other body fluids. RSID™-Urine was developed to detect Tamm-Horsfall glycoprotein, however the presence of vaginal fluid or blood can impair the applicability of the test (119, 125).

Sweat existing in clothing and other items recovered from crime scenes can contain enough DNA to provide a full profile. Regarding the serological identification of sweat, the use of ELISA to specifically identify the protein G-81 has been reported with successful results, however its use has not been widely implemented (119, 125). Another method consists in the use of electron microscopy in combination with energy dispersing X-ray analyzer that allows determining the chemical composition of very small stains.

However the use of this technique is restricted to compare the evidence stain with another stain of known origin, therefore its use for body fluid determination is very limited (119).

B.7 Description of the problem

The current confirmatory serology tests used in forensic sciences are used to detect proteins and present several limitations. Proteins are the effector cells in the tissues, and therefore in theory it is expected that their presence and relative abundance can establish which body fluid is present. Proteins can be carriers or catalysts of biochemical reactions and a test can use the proteins' substrate as an indicator of an active enzyme. However, for such detection to be possible, the protein must maintain activity otherwise it will consist of a disorganized active center that does not recognize its substrate. Extreme temperatures, changes in pH and moisture can easily denature a protein, rendering its activity non-existent and therefore its detection impossible for body fluid identification. Moreover, some enzymes lose its catalytic capacity due to the presence of inhibitors present at the crime scene and collected with the body fluid stain. These aspects are extremely important since the samples that arrive at a forensic laboratory have not been in environments with controlled temperature, pH or clean from other contaminants. Most of the confirmatory methods mentioned above have been validated in mock crime scene samples, most consisting of samples placed in different materials. Some have been exposed to long term storage or high temperatures but such studies are limited.

Forensic serologists have studied the proteins that are specific to a tissue and developed presumptive methods to identify tissues based on the presence or quantity of a

single protein (110). According to the central dogma, proteins are translated from mRNA transcripts. Therefore, measuring specific mRNA transcripts instead of proteins may provide a method with higher sensitivity since unlike proteins, levels of mRNA can be directly assessed without the need for catalytic activity. The transcriptome of a tissue is defined as the pattern of gene expression for the cells that compose such tissue. Forensic research has aimed to determine the transcriptome of body fluids commonly found at crime scenes and identify which mRNA transcripts are specific for which tissues (126). What is attractive about measuring levels of mRNA compared to proteins is the possibility to use automated methods similar to those used for DNA.

To measure the levels of one specific mRNA transcript, first the total RNA of a tissue must be extracted and quantified. The extracted RNA is then incubated with DNase I, which will degrade any DNA co-extracted with RNA so it would not interfere with the reverse transcriptase reaction. Using random primers that anneal in the poly-A tail of mRNA or in other locations, a reverse transcriptase reaction is performed. The reverse transcriptase reaction is similar to PCR but uses enzymes called reverse transcriptase, which are capable of reading RNA, and synthesizing DNA (called cDNA), which is complementary to the RNA. Following this reaction, a quantitative PCR or end-point PCR will amplify the selected regions that correspond to the mRNA transcripts of interest for the target tissue (127). Additionally, a housekeeping gene is also amplified to allow the normalization of mRNA transcripts quantified. The amount of mRNA produced in a cell or tissue may be momentarily increased as a result of the cell cycle or other factors inherent to the status of the tissue (10). For that reason, the concentration of the specific mRNA transcript is always normalized with the concentration of a housekeeping gene

(128). The transcripts of a housekeeping gene are considered stable because of a constant

need of such genes by the cells. Lack of normalization between the mRNA transcript of

interest and the transcripts of housekeeping genes would lead to several errors while

assessing the concentration or presence of a specific mRNA transcript since some are

only expressed at specific points of the cell cycle, and when analyzing a tissue, several

cells in different points of cell cycle are present.

For a quantitative PCR approach, the quantity of each transcript after normalization with

the housekeeping gene is determined through comparison of fluorescence of unknown

samples to those of standard solutions that have known concentrations (128). If end-point

PCR is used, the amplicons may be analyzed by capillary electrophoresis (CE) similarly

to what is done for STRs. Additionally, the use of high resolution melt analysis has also

been tested for analysis of mRNA transcripts in forensic samples (129). Multiplexing

several primers is also a possibility for this approach as long as they can be distinguished

by the real-time instrument or by the CE application (130, 131).

One of the first reports to quantify mRNA for body fluid characterization in

forensics was published in 1999 by Bauer *et al.* (132). In that report, the authors aimed to

determine the presence of epithelial cells in menstrual blood in stains stored for 6 months.

The positive results achieved initiated a body of research in stability and usefulness of

mRNA transcripts for the identification of body fluids. Juusola and Ballantyne (126, 130,

133) were very active in the identification and in performing multiplex reactions of

several mRNA transcripts specific for different tissues. Initially the analysis of transcripts

was merely to determine its presence in only the target body fluid and it was achieved by

gel electrophoresis (126), later with capillary electrophoresis (130) and in the last report

by the two authors the use of a real-time instrument with TaqMan® technology (Figure

2.2). In real-time PCR, the determination of C_T values for all transcripts in all tissues of

forensic interest allows the establishment of a reference table that would permit forensic

analysts to identify the tissue of interest. The number of fluorophores available and the

possibility that a mixture of body fluids and not a single source sample is present may

limit the analysis of mRNA transcripts by this method. Additionally, individual variation

in the mRNA expression of certain genes may also lead to false negatives (133).

Other reports (134) have used a multiplex approach similar to that of Juusola and

Ballantyne (2007) using different mRNA transcripts with varying specificities for the

target body fluid. With the aim of identifying the transcriptome of forensically-relevant

tissues, in 2013, Park and colleagues (135) performed a whole-genome microarray for

blood, saliva, semen and vaginal secretion in order to detect which mRNA transcripts

were present in higher quantities in which body fluids. In total, they were able to select

137 candidate genes from which a second selection of 41 was made. Some of the body

fluid-specific mRNA transcripts already reported by other groups also showed as

candidate genes for those body fluids in the microarray, further confirming the previous

work by others.

The main concerns with the use of RNA is the low stability of the molecule over

time and when exposed to the elements, variation in its abundance on the tissue of origin

and the need to use extra portions of evidence to extract RNA for body fluid

identification (136). Most of the forensic research using mRNA transcripts have done

some level of stability studies and developed mock crime scene samples to determine the

ability to detect mRNA in challenging samples. Despite the overall results being positive,

environments with high humidity can cause issues in the detection of mRNA and more studies are needed until this approach is robust enough for body fluid identification. Regarding the use of extra evidence to perform RNA extraction, even though the simultaneous extraction of DNA and RNA from the same piece of evidence is possible (137, 138), handling RNA in a laboratory must be accompanied with extra care because of the ubiquitous presence of RNases (including in human skin). For example, all water used, including that used to dilute buffers and other solutions must be treated with diethyl pyrocarbonate (DEPC), which includes long incubations so that the DEPC may combine with all RNases and a final autoclave step to remove the DEPC from the water(128). The limitation in the use of mRNA can also result from where the sample itself was collected, since the presence of RNases is ubiquitous in the human body (139).

Other issues related to the use of mRNA for body fluid identification include the use of DNase I to remove any DNA contaminating the extracted RNA (133, 140). The accidental release of this enzyme in a forensic laboratory could signify the impairment of DNA typing methods for several samples, even for samples processed several weeks after the release happens. To tackle some of the stability issues inherent to mRNA transcripts, some research has suggested the use of microRNA (112, 141). Because they are smaller, microRNA molecules present higher stability when compared to mRNA (142, 143), however the issues with DNase I and the ubiquitous presence of RNases are still a concern even for these smaller RNA molecules.

Using the central dogma as a reference, it makes sense to look at DNA instead of RNA or proteins to the identification of body fluids. DNA presents higher stability, it is the source of the code for all genes and for differential gene expression in tissues.

Moreover, the extraction, use and storage of DNA in forensic laboratories has been already validated for several years (68, 72).

The reliability of a confirmatory test for body fluid identification for forensic purposes depends on its capacity to accurately determine which body fluid is present regardless of its concentration, its purity and whether its mixed or not with other body fluids. The accurate determination of which body fluid is in the origin of a DNA samples can be extremely important to solve some crimes, namely those that rely on determination of the suspect's involvement in the crime. If a serology test shows false positives or false negatives in one of the validation studies performed, it creates reasonable doubt for any piece of evidence analyzed with such test. Adding to this, the amount of evidence recovered from the crime scene may already be very limited, therefore any portion of it being used for a test known to show false results may be considered a waste of sample. For this reason there is the need to establish a body fluid determination test that can achieve similar degree of certainty as those methods currently observed in DNA typing technologies. Due to the fact that all forensic laboratories are well equipped to collect, extract, store and analyze DNA, it seems logical to choose DNA as the biomolecule for body fluid discrimination in forensics.

CHAPTER III – FORENSIC EPIGENETICS

A. DNA Methylation Analysis for Forensic Applications

Since the mid-1990's that DNA has been used to match a suspect to a crime scene (60). The optimization of methods and techniques from collection to analysis hasn't stopped since then. Most forensic laboratories are equipped with and have a set of validated standard operation procedures to collect and store DNA. On the biochemical research side, more information is known about DNA methylation presently than it was 10 years ago and several commercial kits have been launched to analyze DNA methylation of loci known to be implicated in pathologies. DNA samples are already commonly retrieved from crime scenes and stored after DNA typing. Oftentimes in the investigation, the need to determine body fluid type emerges after DNA typing. The use of DNA methylation methods in forensic laboratories would for example allow the forensic analyst to simply retrieve a stored DNA sample rather than going back to the evidence collected at the crime scene (shirt, sheet, etc) in order to analyze another biomolecule (RNA or protein), thus saving precious evidence and time.

B. Methods to Analyze DNA Methylation Patterns

To determine which CpGs of a specific DNA region are methylated or unmethylated several methods can be employed depending on several parameters such as total amount of DNA template available, initial quality of the DNA template available, etc (144). Most of the methods currently available for analysis of DNA methylation patterns have been developed for health sciences such as for the study of gene expression

in cancer patients. Researchers in health sciences can depend on high amounts of intact DNA since most samples are biopsies from patients. For that reason, some of those methods present limitations regarding their potential use in forensic sciences and will must be adapted or improved to achieve higher sensitivity. For example, since DNA typing methods can match samples with as little as 100 pg of DNA, a method for body fluid identification should aim to reach that sensitivity. Another example are the high throughput and robustness of the body fluid identification method should be equivalent to DNA typing methods so they can both be used with samples from crime scenes.

DNA typing methods have been validated according to the SWGDAM guidelines (109). Even though the guidelines were designed specifically for DNA typing and address some things specific for that analysis, any validation methods covered in the SWGDAM will apply to body fluid identification using DNA methylation. Different types of methods have been presented to study the DNA methylation pattern of body fluids. Such methods start by using either restriction enzymes that recognize and cleave the DNA in presence of methylated cytosines or methods that rely in the bisulfite conversion of DNA which distinguishes unmethylated cytosines by chemically converting them to uracil while leaving the methylated cytosines untouched. The analysis downstream depends on which method was used but should also consider factors such as throughput, automation, robustness and sensitivity.

B.1 Restriction Enzymes

Restriction enzymes are proteins capable of cutting DNA near or at specific recognition sequences. There are over 50 restriction enzymes that recognize DNA at a

CpG site. The cleavage of DNA will be dependent upon the recognition of unmethylated

cytosines in the dinucleotide CpG (145). Initially the loci of interest were studied and a

determination to which restriction enzyme would cut the DNA near the CpG of interest

was assessed. In the case of body fluid identification, the CpG of interest may be one that

is methylated in one body fluid (the target body fluid) and unmethyated in all others.

Alternatively, the target body fluid may have an unmethylated CpG whereas all other

body fluids present methylation at that specific dinucleotide. After the appropriate

restriction enzyme is selected, primers are designed to flank the region where the enzyme

cuts the DNA. The restriction enzyme *HhaI* has been commonly used in forensic studies

for the use of DNA methylation analysis in body fluid identification (146, 147). *HhaI*

recognizes and cleaves within the sequence GCG↓C, where the arrow pointing down

represents where the sequence is cleaved if the cytosines are unmethylated (147, 148).

For a body fluid with methylated cytosines in the recognition sequence, there will be no

cut in the DNA and therefore both forward and reverse primers will be able to anneal and

amplify the region through PCR. If the cytosines are unmethylated the DNA is cleaved

and two fragments are formed preventing the amplification. The analysis downstream can

be done by capillary electrophoresis where a peak corresponding to the size of amplicon

is visible if the CpGs are methylated. Similarly to STRs, a multiplex PCR reaction can be

designed if all the loci contain the recognition sequence and by labeling each primer with

different fluorophores. However, the absence of a peak may indicate unmethylated CpGs

or incomplete cleavage by the restriction enzymes (146, 147). The fact that incomplete

digestion can lead to false results severely limits this methodology. Incomplete digestion

can occur when the temperature or time of incubation is not ideal; or due to an

inssuficient concentration of enzyme relative to the quantity of DNA present. One possible method to control for temperature and time of incubation would be to run known controls at the same time as the forensic samples. However, if the incomplete digestion occurred as a result of the inaccurate quantification of DNA or because of the presence of other contaminants co-extracted with DNA, the use of a control DNA would not necessarily indicate incomplete digestion. Another major drawback for the use of restriction enzymes in forensic identification of body fluids is the necessity to identify restriction enzymes able to recognize sequences at or near the loci of interest (149).

B.2 Bisulfite Modification

In forensic sciences, a more robust method is to use bisulfite modification to maintain the patterns of DNA methylation (149). The deamination of cytosines to uracil in the presence of sodium bisulfite was first described in 1970. The chemical reaction consists on a deamination promoted by sodium bisulfite, in a slightly acidic (pH 5-6) media heated to 66 ºC. The sulfite group is then removed by a reaction of desulphonation at pH 10 resulting in uracil. No other nucleotides besides unmethylated cytosines are affected by the sodium bisulfite. The presence of the methyl group on the 5'- carbon of the methylated cytosines inhibits the capacity of a nucleophilic attack at the double bond shared between carbons 5 and 6 in the cytosine ring (150, 151). DNA polymerase recognizes the uracils resulting from this reaction by adding an adenine in the newly forming strand. Bisulfite conversion can cause fragmentation of DNA which may be of concern for forensic sciences. For that reason, it is advisable to amplify relatively short loci as larger fragments of DNA may cease to be intact following conversion.

After bisulfite conversion, the DNA is amplified by PCR and can be analyzed by a veriety of methods capable of determining the status of individual CpGs or the relative abundance of methylated versus unmethylated CpGs. The methods utilized in this dissertation include pyrosequencing, high-resolution melt (HRM) analysis and SNaPshot.

B.3 Pyrosequencing

The quantification of methylation for each individual CpG in a targeted DNA sequence can be performed using pyrosequencing, which was firstly described as an alternative to Sanger sequencing. In 1986 Pål Nyrén developed a way to follow in real time the release of the pyrophosphate by DNA polymerase during the incorporation of nucleotides (152). Briefly, pyrosequencing uses a cascade of enzymatic reactions following addition of a base by DNA polymerase to produce light emission that is detected by a camera and displayed in a graph called a pyrogram. Prior to pyrosequencing, a PCR is setup with primers that amplify a larger DNA region than the one to be sequenced. The region to be sequenced is called target region and is located within the amplicon of the PCR reaction. One of the PCR primers contains a biotin molecule at their 5'-end. As amplification progresses the PCR product will consist of double stranded molecules with one biotin-labeled strand. The amplicons are mixed with streptavidin-coated beads and captured by vacuum on filter probes. The captured DNA is then washed in successive solutions, including a denaturing one that separates the double strand DNA, releasing the non-labeled strand and retaining the biotin-labeled one. The beads with the single stranded DNA are then released in a solution containing a third primer, called sequencing primer. After the sequencing primer anneals to the single-

stranded DNA, the mixture is added to the pyrosequencing instrument and sequencing begins (153). The instrument is composed of a cartridge that contains all 4 nucleotides, an enzyme mixture and substrate mixture in separate wells and a place for a sequencing plate. Due to increase in pressure, the contents of each well in the cartridge can be individually dispensed on the plate. At first the enzyme mixture and the substrate mixture are dispensed in all wells of the plate. The nucleotides are then added one at a time, according to a specified order, to each well. If the added nucleotide is complementary to the DNA strand being sequenced, the DNA polymerase will add the nucleotide and release pyrophosphate, which is similar to what happens on any reaction catalyzed by DNA polymerase. Next, adenosine 5'-phosphosulfate in the presence of the released pyrophosphate is converted to ATP by the enzyme sulfurylase. Luciferase can then catalyze the oxidation of luciferin to oxyluciferin due to the presence of ATP in solution, emitting light. The amount of light generated is proportional to the amount of ATP and it is detected by charge coupled device (CDD) and displayed as a peak on a pyrogram (153). Before the next nucleotide is added to the mixture, apyrase degrades any ATP and nucleotides that were not used. The peaks in the pyrogram have a height proportional to the number of each nucleotide added. For example, on the sequence 'GCAGGCCT', the pyrogram would be similar to what is shown on Figure 3.1.
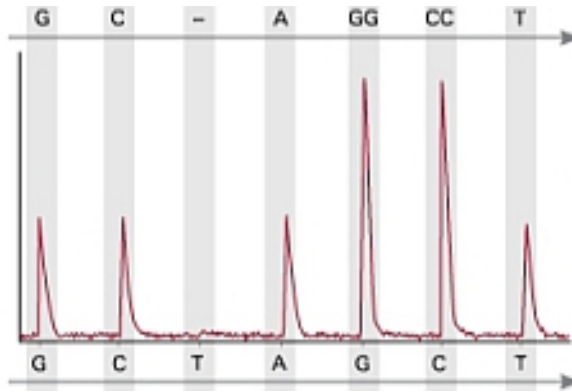
Figure 3.1 – Pyrogram depicting the relative peak height for all the nucleotides dispensed when sequencing the fragment 'GCAGGCCT'. The pyrogram only shows peaks when the dispensed nucleotide is in the sequence. The relative peak heights for a position with two repeated nucleotides ('GG' and 'CC') are double the height of peaks in positions that contain only single nucleotides. Adapted from (154).

When performing analysis for DNA methylation, for each CpG the instrument will dispense a T and a C and measure the light emitted after each dispensation. Since the light is proportional to the amount of each nucleotide, quantification is possible. On the Qiagen Pyrosequencing® method multiple assays with different PCR products and sequencing primers can be ran in the same plate. The PyroMark® software used to design each run allows the customization of each individual plate by a drag-and-drop method of assays in the plate layout (155). The user needs to input the sequence to analyze and the software will create the dispensation order. If the user also inputs the 'sequence to analyze prior to bisulfite conversion', which corresponds to the genomic DNA sequence, the software highlights possible locations for bisulfite controls in red.

The peak height will provide information about the relative amount of cytosines and thymines for each CpG. Since bisulfite reaction does not modify methylated cytosines, a high peak when a cytosine is dispensed corresponds to a high methylation percentage. On the example shown on Figure 3.2, the first CpG shows 2% of methylation because the peak height for 'C' is 2% of the total peak height for a single nucleotide. The PyroMark® software determines the peak height for all individual base pairs, such as the first 'G' shown in the pyrogram (Figure 3.2, position 4). On the variable positions (CpGs) the sum of peaks of 'C' and 'T' should be equivalent to the height determined by the remaining sequence.
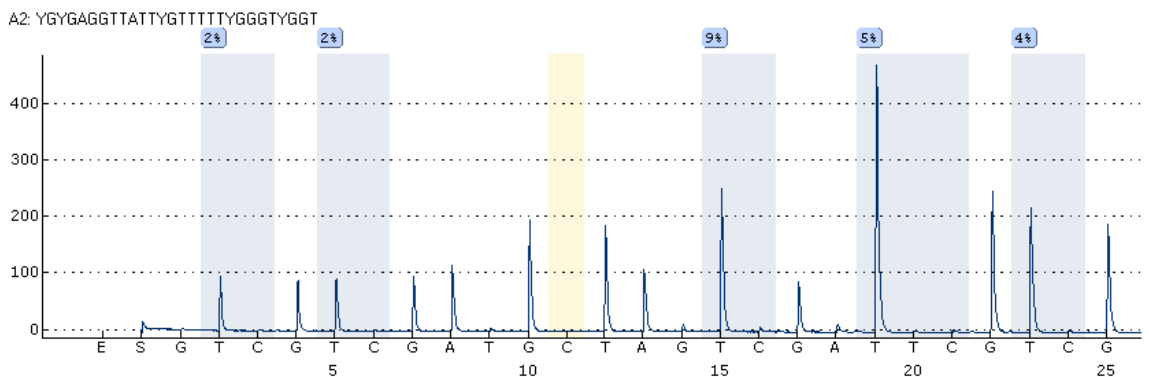
Figure 3.2 –Pyrogram showing the percent methylation in 5 CpGs (highlighted in blue). The peaks corresponding to 'C' are very low when compared to 'T', which makes the methylation level to be below 10% for all peaks. The pyrogram shows relative light units on the y-axis and position on the target sequence on the x-axis. The 'E' and 'S' on the beginning of the x-axis correspond to the dispensation of enzyme and substrate mixtures, respectively. The 'C' highlighted in red shows the bisulfite control. On top of the pyrogram is shown the sequence to analyze, where 'Y' corresponds to a CpG

The percent methylation is then calculated from the relative peak height of 'C' versus the expected peak height for a single nucleotide. To confirm that the correct PCR product is being sequenced instead of a random sequence that annealed to the sequencing primer,

the instrument dispenses negative controls, that is, nucleotides that are not complementary to the strand being sequenced and these should not display any peak when dispensed. One example of a negative control is the first 'G' dispensed after the substrate (S) on Figure 3.2. The sequence to analyze (Figure 3.2, on top labeled as 'A2') starts with a CpG (labeled as 'Y' on the sequence to analyze) and thus either a 'C' or a 'T' should be dispensed, however the instrument dispensed a 'G' to confirm that the sequencing primer annealed where expected. Comparing the sequence to analyze and the nucleotides dispensed on the x-axis will provide information on which ones are negative controls.

The bisulfite control is shown in the image as a dispensed 'C' on position 11. The PyroMark® software assumes that all 'Cs' not followed by 'Gs' on the genomic DNA sequence are unmethylated and thus will be converted to 'Ts'. As we can see on the sequence to analyze on Figure 3.2, after position 11 there is a string of 2 'Ts', which result from bisulfite modification of unmethylated 'Cs'. If bisulfite modification had not been completely successful, at least one of those 'Ts' would still be a 'C' and thus a peak would show on the pyrogram. Dispensing a 'C' before or after any 'T' that is the result of bisulfite conversion will therefore serve as a bisulfite control for pyrosequencing. When the 'sequence to analyze prior to bisulfite conversion' is included in the assay design of the Pyromark® software, it highlights in red the potential positions on the sequence where bisulfite controls can be added by the user. Otherwise, the nucleotide dispensation can be modified to include as many bisulfite controls or negative controls as desired or depending on the time available to sequence each well, since more nucleotides dispensed results in longer dispensations and more time needed for each well (155).

The first time that pyrosequencing was used for body fluid identification for forensics application was in 2012 and the results are displayed in the report of Madi *et al.* (156). In their experimental setup, DNA extracted from several samples of body fluids and methylation was quantified for 4 separate genome locations. The authors selected which CpGs in each location show methylation differences between body fluids. When multiple samples from different donors of each body fluid were ran with each primer pair, the percent of methylation for each CpG of the 4 separate genome locations was recorded. Thus each body fluid has an expected mean percent of methylation for each genome location and a corresponding standard deviation.

The limitations of this method being implemented in forensic laboratories is the fact that it requires new equipment and training of personnel, which may be a deterrent for some laboratories. However the great advantage of pyrosequencing is its capacity for high throughput and its ability to be automated (153). The forensic analyst can be confident in which CpGs are methylated since the pyrogram reveals all issues associated with the sequencing process. Furthermore, even though the method has not yet been designed for multiplexing PCR, it may be possible to have multiple primers amplified in a single PCR reaction and then sequenced separately depending on which sequencing primer is added for each well.

B.4 High-Resolution Melt (HRM) Analysis

High-resolution melt (HRM) for DNA methylation analysis was first described in 1997 by Wojdacz and colleagues (157). After bisulfite conversion and PCR, an unmethylated amplicon will be mainly composed of double hydrogen bonds between

81

adenine and thymine thus having a lower melting temperature when compared to an amplicon derived from a methylated template (158) (Figure 3.3). The melt analysis is performed for the PCR products in a real-time instrument with melt analysis capacity, in a closed tube, immediately after the end of PCR. Like many other real-time amplifications, the mix contains an intercalating dye such as EvaGreen that emits fluorescence when in the presence of double stranded DNA. With the completion of the PCR reaction, the melting protocol begins by heating the PCR products in increments of 0.1-0.3 °C and reading the decrease in fluorescence. The gradual heating results in a melt of the double stranded DNA, which causes a release of the intercalating dye and a decrease in fluorescence. As the temperature increases, the fluorescence decreases at a steady pace, followed by a sudden decrease around the melting temperature (T_M) of the PCR product (Figure 3.4). The T_M is defined as the temperature at which half of the double stranded DNA has been melted, which roughly corresponds to half of the fluorescence in a graph like the one on Figure 3.4. To facilitate the visualization of T_M, another graph depicting the first derivative in fluorescence per temperature, that is, the observed change of slope (defined as '-dF/dT') of Figure 3.4 is then plotted *vs* temperature (T) and the peak observed corresponds to the T_M of the amplicon (159, 160).
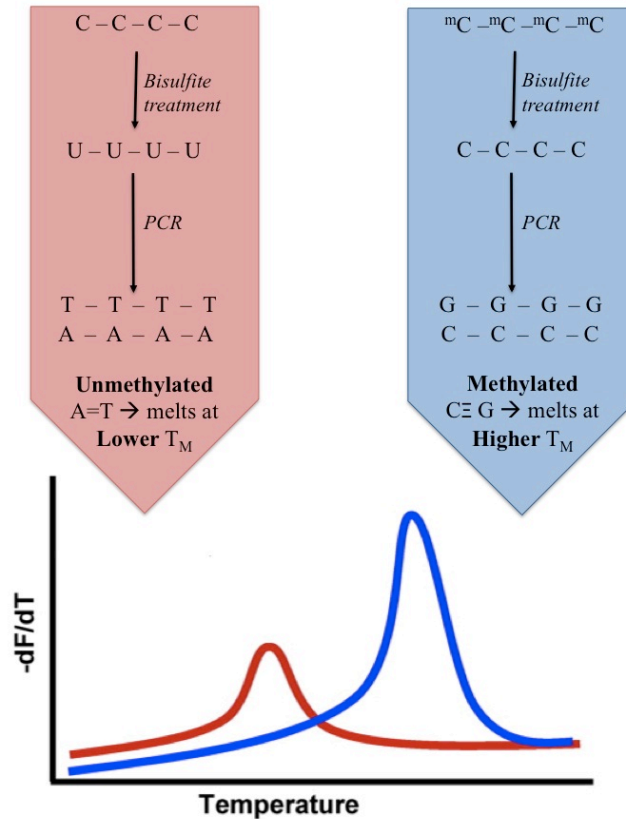
Figure 3.3 – Schematic representation of differences in melt curve analysis for one DNA target with different methylation status. The melt peaks will locate at different temperatures depending on methylation. Adapted from (161)

The method is quick and nondestructive, allows the characterization of PCR products (162) derived from bisulfite-modified DNA templates with different levels of methylation. It uses an instrument available in most forensic laboratories, which facilitates the adoption of this methodology by smaller forensic laboratories. HRM analysis has some disadvantages when compared to pyrosequencing such as the fact that it does not allow individual CpG quantification and that some amplicons may become

difficult to analyze using melting properties because of potential secondary structures formed after bisulfite modification (163).
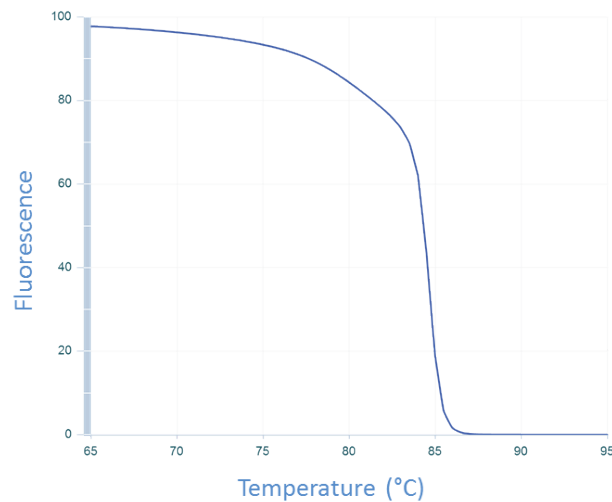
Figure 3.4 – Schematic representation of the decrease in fluorescence resulting from intercalating dye release during the separation of the double strands of a PCR product. The slope resulting from the sudden drop in fluorescence is plotted versus temperature to better represent the T_M of the PCR product. Adapted from (160).

However, HRM analysis can also present several advantages when compared to pyrosequencing. One advantage is the overall lower cost that results from a faster turnaround because of the melt analysis after amplification being automatic. Moreover, less expensive reagents and lower investment in a new instrument is necessary for those cases where the laboratory needs to acquire a real-time instrument with melt analysis. Another advantage of HRM is that the assay only needs a forward and reverse primers and not a third sequencing primer. As a result of bisulfite modification, some templates

become very rich in 'A' and 'T' and also become very similar amongst themselves which may impair primer design. HRM analysis can be multiplexed if the amplicons display different melting temperatures, which correspond to different peaks, or if the melt curves can be distinguished by a fluorescent method .

Overall as a result of the quick turnaround, HRM analysis is a method that can easily be optimized and is familiar to most forensic DNA analysts. For those reasons there are many advantages in investing in HRM research and analysis for body fluid identification.

B.5 Methylation-Sensitive Single Nucleotide Primer Extension

The analysis of methylation after bisulfite conversion can also be achieved using capillary electrophoresis. The methylation-sensitive single nucleotide primer extension (ms-SNuPE) method is similarly to SNaPshot, which has been widely used for SNP analysis (164). Following bisulfite modification, the DNA is amplified with primers specific for bisulfite-modified DNA and flanking the region of interest. The PCR product is then purified to remove extra primers and dNTPs that were not used in the amplification. The ms-SNuPE primers are designed to anneal to the bisulfite modified DNA immediately upstream of the CpG of interest (164). During the single base extension reaction, which is prepared similarly to a PCR, a dideoxy-modified nucleotide is added to the nascent strand of DNA (165). Since the modified nucleotide does not allow further amplification and because each dideoxy-nucleotide is labeled with a specific fluorescence color, the fragments can be separated in a capillary electrophoresis instrument. The capillary electrophoresis instrument determines the presence of a 'T' or a

'C' on the template strand. The peak height provides the relative amount of each

nucleotide, which are labeled with different fluorophores. Even though this method

cannot provide information regarding the whole sequence, it does have single CpG

resolution and allows quantification (164). The instrumentation necessary for ms-SNuPE

is familiar to the forensic analyst which represents an advantage over pyrosquencing.

However, the procedure includes steps that are time consuming and not automatable,

such as enzyme incubation. Moreover, each sample requires two reactions to be set - the

initial amplification and the single-nucleotide extension reaction - which may limit

throughput. Ms-SNuPE has been explored in body fluid identification for forensic

sciences purposes with mixed results. For example, in the report by An and colleagues

(146) a set of 3 semen-specific markers (USP49, DACT1 and PRMT2) were analyzed by

ms-SNuPE (which is referred as SNaPshot by the authors). A fourth marker PFN3 which

has vaginal-specific methylation and was also included in the analysis. The authors were

able to achieve high sensitivity by obtaining results from samples with 500 pg of DNA.

However when semen samples of vasectomized and non-vasectomized men are analyzed,

the peaks are different amongst themselves at the DACT1 marker but more importantly

the peaks for all markers are different between the vasectomized males (n=5) and non-

vasectomized males (n=35) when compared to the 29 semen samples (not labeled as

vasectomized or not). Even though the authors did not address the contradicting results,

the fact that the semen samples were pooled together may explain the disparity of results

and also the high sensitivity of 500 pg for this protocol. SNaPshot was used in two other

reports (166, 167) by the same team of researchers to validate results obtained from a

bead array study designed to identify new genome locations able to distinguish body

fluids commonly found in forensic sciences. In both reports, the expected DNA methylation values obtained by the bead array can differ when a PCR-based method, such as SNaPshot, is used to analyze samples. However, on the second report (167) the markers tested by SNaPshot showed higher specificity for each body fluid, probably due to an increased number of samples analyzed *in silico*. Regarding the SNaPshot method, only the 2015 report performed sensitivity studies. The results vary depending on how much DNA was bisulfite modified even if the same amount was added to the PCR reaction. Additionally, it is not clear how many samples were used for sensitivity studies and even the authors agree that more work needs to be done (166).

One of the issues that can arise with SNaPshot results is the different sensitivity of markers may mean that some will present high peaks whereas others may present peaks so low in the electropherogram that they are not detected. If one PCR product amplifies more efficiently than others in a multiplex PCR reaction, that peak may be so high that smaller peaks will be undetected due to issues with dynamic range (99).

C. State of the Art

Even though the analysis of DNA methylation patterns in other fields begun several years ago, their applications in forensic sciences is still at its infancy. The interest in the use of DNA methylation for forensic science applications started only in 2010 with a report from Frumkin and colleagues (168), which aimed to show that artificial DNA amplified by PCR and left deliberately at a crime scene, can be distinguished from authentic DNA from the victims and perpetrators based on methylation. Since amplified DNA is all unmethylated, studying DNA regions naturally methylated would allow the

87

distinction between the two types of DNA. During their work, Frumkin and colleagues utilized methylation-specific restriction enzymes to observe that some of the loci used presented different methylation levels for different body fluids examined. The difference in methylation observed led to a second report specifically directed towards body fluid identification using DNA methylation patterns (147). In the 7 years that have passed since that first report, the number of loci described for body fluid identification analysis has been limited. Research efforts have been made to identify more genome locations capable of distinguishing body fluids for forensic purposes, particularlly using array-based methods. However, each array can run a limited number of samples and is still prohibitively expensive for most forensic research laboratories. Moreover, such arrays cover mostly CpGs that provid specific methylation site for certain types of cancer or pathologies and the CpGs are mostly located in CpGi, which are not necessarily co-localized with regions of the genome more relevant for body fluid identification. Nonetheless in the past 7 years a balance between the validation of existing loci for body fluid identification and the search for new loci has taken place. In the future, as this technology becomes more reliable, it is expected that funding agencies will recognize the value to invest in whole-genome studies aimed specifically for forensic samples which have their own composition regarding the type of cells and the amount of DNA present.

CHAPTER IV – FORENSIC DISCRIMINATION OF VAGINAL EPITHELIA BY

DNA METHYLATION ANALYSIS THROUGH PYROSEQUENCING

A. Abstract

The accurate identification of body fluids from crime scenes can aid in the discrimination between criminal and innocent intent. This research aims to determine if the levels of DNA methylation in the locus PFN3A can be used to discriminate vaginal epithelia from other body fluids. In this work we bisulfite-modified and amplified DNA samples from blood, saliva, semen and vaginal epithelia using primers for PFN3A. Through pyrosequencing we were able to show that vaginal epithelia presents distinct methylation levels when compared to other body fluids. We also show that the sensitivity of the method was determined to be 5 ng, that mixtures of different body fluids have methylation values that correlate with single-body fluid samples, and that the primers for PFN3A are specific for primates. We successfully demonstrated that analysis of methylation in the PFN3A locus can be used for vaginal epithelia discrimination in forensic samples.

B. Introduction

In forensic cases, determining the presence of vaginal secretions can aid in the distinction between innocent and sexual contact. The current method used in forensic laboratories to presumptively identify vaginal fluids depends on successful histological staining of the glycogen-rich vaginal cells by the Schiff reagent. However, the method's sensitivity depends on the reproductive age and menstrual cycle of the female and it easily gives false positives by unspecific staining of male buccal and urogenital cells (169). More sensitive and reliable methods are those that rely on the quantification of certain RNAs as a product of cell-specific gene expression. Gene expression can occur at

90

the transcription level which is quantifiable by the levels of messenger RNA (mRNA)

present in a specific type of cells (133); or at the post-transcription level that occurs

through micro RNA (miRNA) presence in specific tissues (170). The most successful

reported methods that are able to quantify mRNA or miRNA are capillary electrophoresis

(171) and high-resolution melt analysis (172). Both methods show promising results

since they are automatable, easy to implement in a forensic laboratory and sensitive.

However, the use of RNA requires normalization of transcript levels with those of a

housekeeping gene in the same sample. Often that normalization has to be performed by

quantifying RNA levels prior to cDNA synthesis. Quantification methods available do

not discriminate between human and bacterial RNA, which is commonly present in most

vaginal samples (169). Since most transcripts are present at a basal level in body fluids,

the quantification of RNA for body fluid identification relies on levels of specific

transcripts being higher than a certain threshold in specific cells. Lack of an accurate

quantification of RNA in the sample can lead to false results since it influences the

threshold for the target transcript. Because of the aforementioned difficulties, we propose

the analysis of different DNA methylation levels as a way to discriminate body fluids.

DNA methylation is one of the epigenetic mechanisms for gene regulation in the human

body. Different levels of DNA methylation in certain loci are thought to control gene

expression by silencing or activating specific genes. The presence of a methyl group on

the 5' carbon of a cytosine belonging to the dinucleotide CG (CpG) is believed to prevent

the binding of the transcription machinery in the promoter of genes (47). Some loci on

the genome called tissue-specific differentially methylated regions (tDMRs) can therefore

be used for body fluid discrimination (136, 147, 173) since they present different levels

of DNA methylation depending on the cell studied (156, 174). To determine the pattern

of DNA methylation in a locus, the most commonly used methods include the bisulfite

modification of genomic DNA that chemically converts the unmethylated cytosines to

uracils but does not react with methylated cytosines. During the polymerase chain

reaction (PCR) uracils become copied as thymines and the resultant amplicons can then

be sequenced to determine the presence of a cytosine or a thymine at each specific CpG.

One of the most commonly used methods for DNA methylation analysis is

pyrosequencing because it allows individual quantification of CpG's and a percent

methylation value is displayed on a pyrogram created by the instrument software (154).

In order to successfully use pyrosequencing in forensic laboratories, more loci capable of

discriminating several body fluids, including vaginal epithelia need to be identified. One

study that aimed to uncover more loci was the work of Lee *et al.* (2010) where samples of

blood, saliva, semen, menstrual blood and vaginal epithelia were collected from

volunteers. After extraction, the DNA from each body fluid was pooled so that the DNA

yield was sufficient for sequencing post bisulfite conversion. Sanger sequencing was

performed and the results compared to *in silico* converted sequence in order to obtain a

comparison of methylation levels. The methylation map revealed that the tDMR named

PFN3 presented an overall different DNA methylation level for vaginal fluid when

compared to the other body fluids (175). To determine if this tDMR can be used to

discriminate vaginal epithelia from crime scenes, we performed pyrosequencing on a sub-

region of PFN3 - PFN3A. Contrary to the method used by Lee *et al.* (2010),

pyrosequencing does not require DNA samples to be pooled. The region selected for this

study – PFN3A - includes only the CpGs that in the work of Lee *et al.* (2010) showed the

highest methylation difference when compared to the other body fluids. The aim of this work was to explore the potential of PFN3A to become a biomarker for vaginal fluid discrimination in forensic laboratories.

C. Materials and Methods

Blood, buccal, vaginal swabs and semen samples were collected from volunteers according to the approved IRB-13-0555 from Florida International University and Pontifícia Universidade Católica do Rio Grande do Sul (CONEP #723.619/ CEP #845.747). Swabs were air-dried and DNA extraction was performed using the EZ1® DNA Investigator kit (Qiagen, CA) and the BioRobot® EZ1 automated purification workstation (Qiagen, CA) according to the manufacturer's specifications. Quantification was performed using the PicoGreen® method (Life Technologies, CA). Fifty nanograms of DNA were bisulfite modified using the EpiTect® Fast DNA Bisulfite Kit (Qiagen, CA) according to manufacturer's instructions. For the sensitivity studies, six vaginal epithelia samples were serial diluted to obtain 10, 5 and 1 ng of input DNA to bisulfite conversion. For the mixtures studies, DNA samples of blood, vaginal epithelia and semen were quantified and mixed in different ratios of each two body fluids in order to have 100 ng total of DNA input to bisulfite modification. For the species-sensitivity studies, DNA from cat, dog, chicken, cow, mice, bacterial pool (*Escherichia coli, Staphylococcus aureus, Enterococcus faecali* and *Pseudomonas aeruginosa*), horse, gorilla, orangutan and chimpanzee was collected and processed like described above. DNA from one sample of human vaginal epithelia was used as positive control.

DNA amplification reactions were performed using the PyroMark® PCR kit (Qiagen, CA) by adding 2 μL of bisulfite-modified DNA to each reaction and following manufacturer's instructions. Primers for PFN3A were designed using the PyroMark® Assay Design software version 2.0 (Qiagen, CA) and are displayed in Table 4.1. The PCR products and a low molecular weight ladder (New England Biolabs, MA) were run in 2% agarose gels, stained with ethidium bromide to confirm the presence and size of amplicons.

Table 4.1 – Sequence of PFN3A primers used in this study.

Primers	Sequence
Forward	5'- GTG TAT AGT TTT GTT GAG GAT GTT TT - 3'
Reverse[a]	5' - ACA AAC ACA CCT TCC TAC AA – 3'
Sequencing	5' - GTT TTG TTG AGG ATG TTT TT – 3'

(a) shows the biotinylated primer

Pyrosequencing reactions were performed using a PyroMark® Q24 pyrosequencer (Qiagen, CA) using 10 μL of PCR product and the protocol established by the manufacturer. The percent methylation for each individual CpG was automatically calculated by the PyroMark® Q24 software version 2.0.6 (Qiagen, CA).

Statistical analysis was performed using SPSS version 22. Methylation percent was compared between body fluids for each CpG performing a one-way ANOVA followed by Tukey-HSD post-hoc test. For the sensitivity studies, a dependent t-test was performed to compare the percent methylation of each CpG from each dilution to the 50 ng group. A p-value <0.05 represents that the difference observed is significant for all statistical tests performed.

D. Results and Discussion

We first aimed to determine if the DNA methylation levels in the PFN3A locus are different for vaginal epithelia when compared to blood, saliva and semen. To address this goal, we performed pyrosequencing of 8-12 DNA samples from each body fluid. The results from each DNA sample from each body fluid are displayed by the Qiagen software as a pyrogram that calculates percent of methylation for 10 individual CpGs within the PFN3A region (data not shown). The percent methylation for each CpG from each body fluid was recorded and the arithmetic average and standard deviations calculated (Figure 4.1). Statistical analysis was performed and only the CpGs that had a significant difference ($p<0.05$) between vaginal epithelia and all other body fluids were reported. Our analysis showed that 9 out of 10 CpGs have an intermediate level of mean percent methylation for vaginal epithelia that is significantly different when compared to the other body fluids.

The only CpG (CpG9) that presented a p-value larger than 0.05 demonstrates that methylation levels between blood and vaginal epithelia are very similar, therefore differentiation between the two is not possible using this CpG. The similar methylation percent on CpG9 demonstrate that for the locus PFN3A this particular CpG will not be used for discrimination of vaginal epithelia, therefore it's percent methylation value will not be displayed in the following results.

Forensic evidence often contains low amounts of DNA that can severely hinder the ability to obtain reliable values for percent of methylation. In order to determine the sensitivity of our method, six DNA samples of vaginal epithelia were quantified and

95

diluted to obtain inputs of 10, 5 and 1 ng to bisulfite conversion, followed by PCR

amplification and pyrosequencing.
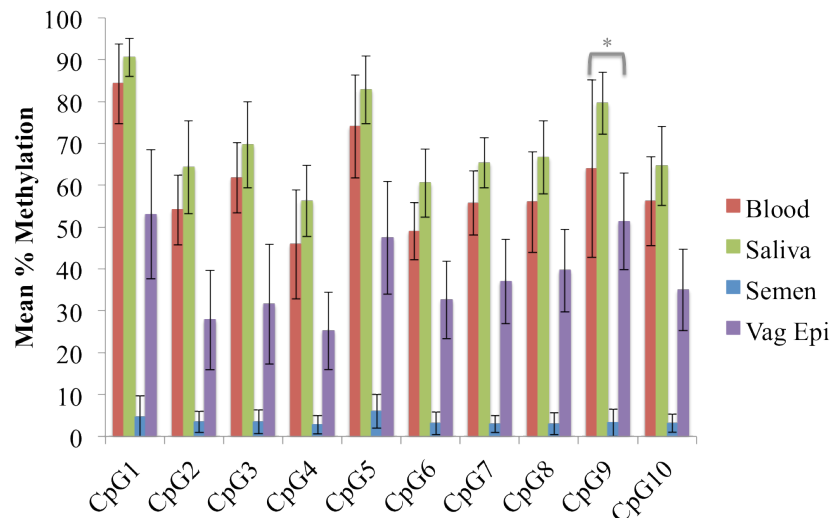
Figure 4.1 - Graph showing mean percent of methylation for samples of blood (n=8), saliva (n=11), semen (n=12) and vaginal epithelia (n=10) with 50 ng input to bisulfite. * CpG where the difference in methylation levels is not statistically significant (p<0.05) between vaginal epithelia and blood.

The mean percent of methylation of the diluted samples were then compared with those

of samples with 50 ng input to bisulfite conversion. The results obtained (Figure 4.2)

show that for samples with 5 ng only one CpG (CpG6) showed levels of methylation

different when compared to the 50 ng samples. When samples of 1 ng of input DNA to

bisulfite were pyrosequenced, less reliable results were obtained. Both CpGs 2 and 3

show differences in methylation levels when compared to 50 ng input samples and all the

pyrograms generated for the 1 ng samples present warnings due to low peak height (as

example, Figure 4.3).
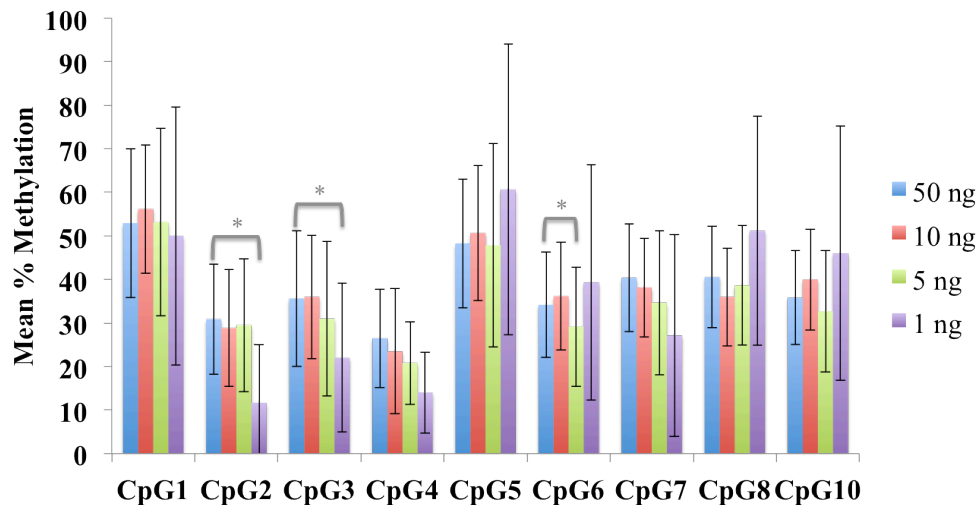
Figure 4.2 - Graph showing mean percent of methylation for vaginal epithelia samples (n=6) with different input of DNA. *CpGs and nanograms of DNA that present methylation levels with statistically significant differences (p<0.05) between groups.

Sensitivity studies show that some data can be obtained with as little as 1 ng of DNA input to bisulfite since the methylation levels are the same as the ones presented by the 50 ng samples for 7 out of 9 CpGs. However the data showed that as the input of DNA decreases so does the peak height in the pyrogram, which leads to warnings recorded by the software. Such warnings are the result of a minimal peak height threshold set by the manufacturer (20 relative light units). For that reason we suggest the use of at least 5 ng of DNA input to bisulfite with this method.

Another common issue in forensic samples is that mixtures of body fluids are commonly present in crime scenes samples. This work aimed to identify vaginal epithelia in samples collected from specimens that are part of cases of suspected sexual assault. For that reason we decided to ascertain how the methylation levels of the PFN3A locus change when different ratios of blood, vaginal epithelia and semen are present in one sample.
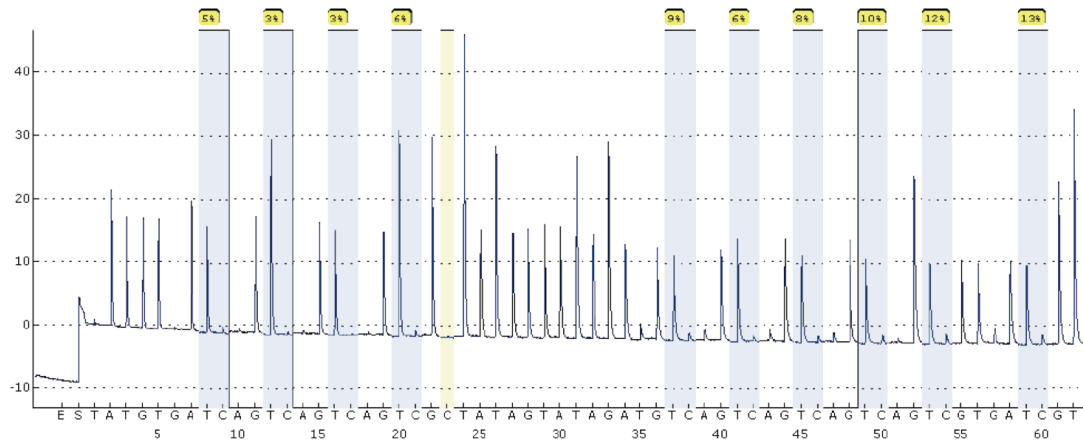
Figure 4.3 - Pyrogram showing a sample with 1 ng input to bisulfite. Peak height for the majority of peaks corresponding to single nucleotide dispensations are lower than 20 light units (y-axis) which creates a warning recorded by software.

Three randomly selected DNA samples, one from each body fluid, were quantified for

DNA methylation levels at the PFN3A locus and compared to the average percent

methylation for each body fluid observed in Figure 4.1. The same randomly selected

DNA samples were mixed with each other in ratios of 75:25, 50:50 and 25:75 and their

methylation levels assessed. Figure 4.4A shows that samples containing mixtures of

blood and vaginal epithelia have methylation values intermediate to those of samples

containing only blood or vaginal epithelia. Moreover there is a decrease in methylation

percent as the ratio of blood to vaginal epithelia decreases.

A similar correlation is also observed for the samples containing blood and semen

(Figure 4.4B) as well as samples where vaginal epithelia and semen are in the mixture

(Figure 4.4C).

The results displayed in Figures 4.4 are promising in the use of PFN3A as one of the

vaginal markers for forensic discrimination of body fluids. More genome locations

capable of discriminating vaginal epithelia based on DNA methylation have been

reported (176). This research provides the first piece of data for methylation values using

PFN3A in mixtures of blood, semen and vaginal epithelia measured by pyrosequencing.
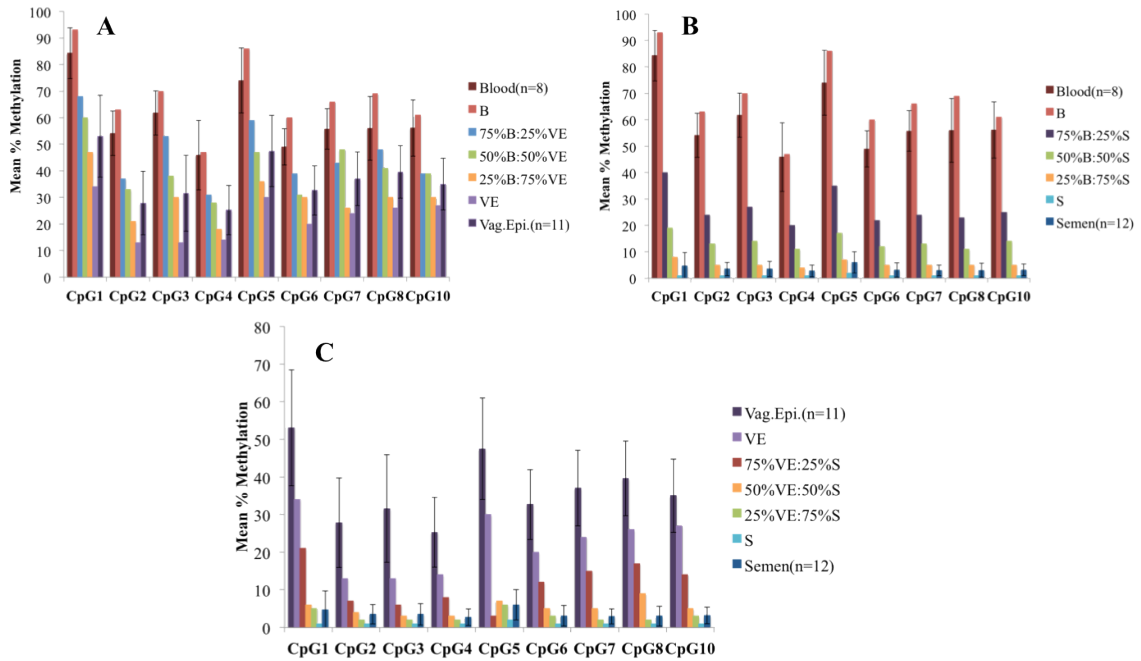
Figure 4.4 - Mean percent of methylation for samples containing different ratios of DNA that was extracted from different body fluids. (**A**) shows the mixture of blood and vaginal epithelia blood. As the DNA content of blood decreases compared to vaginal epithelia, the percent of methylation also decreases. The average percent methylation for the blood (n=8) and vaginal epithelia (n=11) samples are shown as control. The samples labeled 'B' and 'VE' are the DNA from blood and vaginal epithelia, respectively, used to make the mixture. (**B**) shows the mixture of blood and semen. With the decrease in the content of DNA from blood compared to semen, a decrease in the percent of methylation is also observed. The average percent methylation for the blood (n=8) and semen (n=12) samples are shown as control. The samples labeled 'B' and 'S' are the DNA from blood and semen, respectively, used to make the mixture. (**C**) shows the mixture of vaginal epithelia and semen. As the DNA content from vaginal epithelia decreases, so does the methylation percent of the mixture.. The average percent of methylation for the vaginal epithelia (n=11) and semen (n=12) samples are shown as control. The samples labeled 'VE' and 'S' are the DNA from vaginal epithelia and semen, respectively, used to make the mixture.

To determine whether this method could also be hindered by the presence of non-human

DNA, we tested DNA from cat, dog, chicken, cow, mice, bacterial pool (*Escherichia coli,*

Staphylococcus aureus, Enterococcus faecali and *Pseudomonas aeruginosa*), horse, gorilla, orangutan and chimpanzee. One human DNA sample was used as a positive control. Even though in the agarose gel only the samples from cat, cow, orangutan and gorilla showed faint bands and only the human and chimpanzee showed bright bands (data not shown), all PCR products were pyrosequenced to confirm the absence of peaks on the pyrograms. We believe that the presence of bands in the agarose gel results from non-specific amplification since 45 PCR cycles are necessary to amplify bisulfite-modified DNA. If that is the case, when pyrosequencing is performed, the sequence should not correspond to the expected sequence and the software should display red warnings.

The results showed that only the samples from human and chimpanzee present a pyrogram without any warnings recorded by the software (Figure 4.5A and B). For the samples of cat, cow, gorilla and orangutan the pyrogram showed peaks that do not correspond to the expected DNA sequence (as example Figure 4.6A), which means that the DNA sequenced was not the specific for the PFN3A locus. Pyrosequencing with respect to DNA methylation analysis requires the user to input the target sequence to analyze so that the software can automatically compare the peaks on the pyrogram with the expected peaks from the sequence to analyze. Comparing Figures 4.5 and 4.6 one can see that the observed peak height for the tested sample (Figure 4.6A) does not correspond to the peak height of the expected sequence. This difference in peak height agrees with the notion that the bands observed in the agarose gel are due to unspecific amplification during PCR. All other samples provided negative results with no peaks present in the pyrograms. One of those samples is the DNA from bacterial pool (Figure 4.6B). The

absence of peaks is expected since no bands were observed in the agarose gel for these samples.
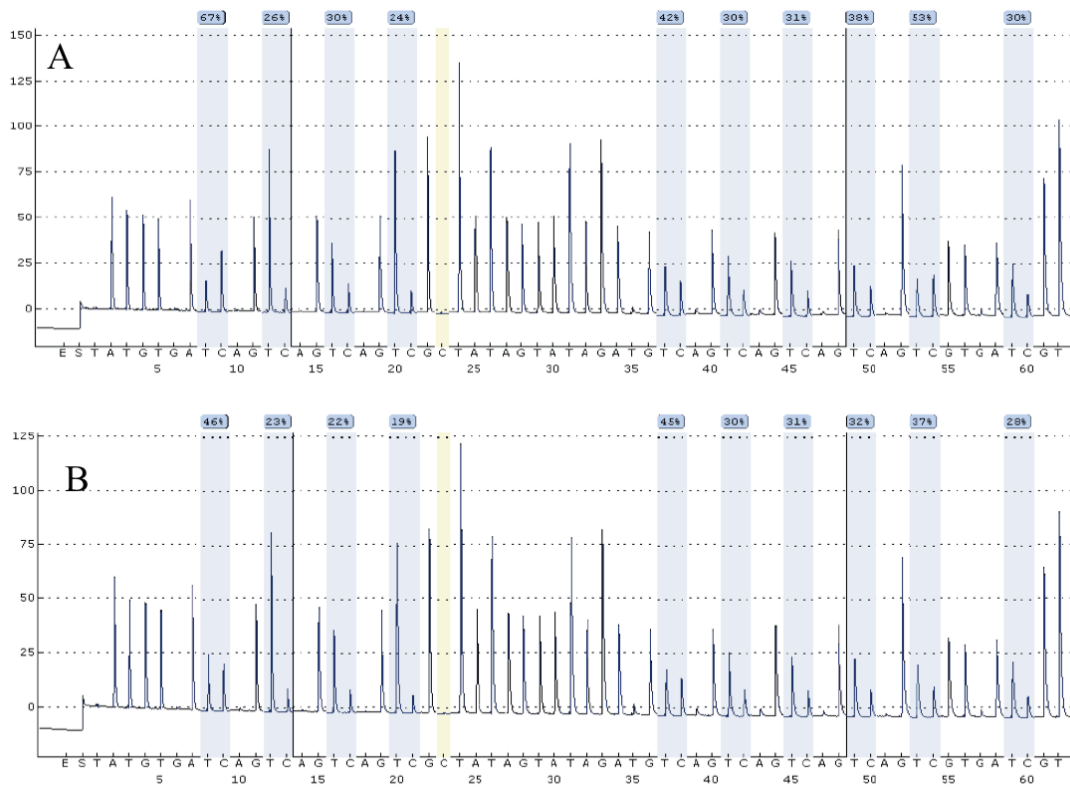
Figure 4.5 – Pyrograms for the PFN3A locus showing DNA samples from (**A**) chimpanzee and (**B**) human after pyrosequencing. Both samples show good peak height and expected sequence therefore all CpGs analyzed passed the quality test imposed by the software.

Contrary to methods that use RNA, body fluid discrimination using DNA methylation methods do not depend on accurate quantification of human DNA present in the sample. Through the use of specific primers for bisulfite-converted DNA, we can obtain amplicons that correspond to a genome region that are primate-specific. Forensic samples commonly contain DNA from other species, either due to exposure to the environment or as a result of the presence of several bacterial species that are part of the human microbiome. For that reason a method that uses nucleic acids for body fluid

discrimination must be specific enough to tolerate the presence of non-human DNA or

RNA.
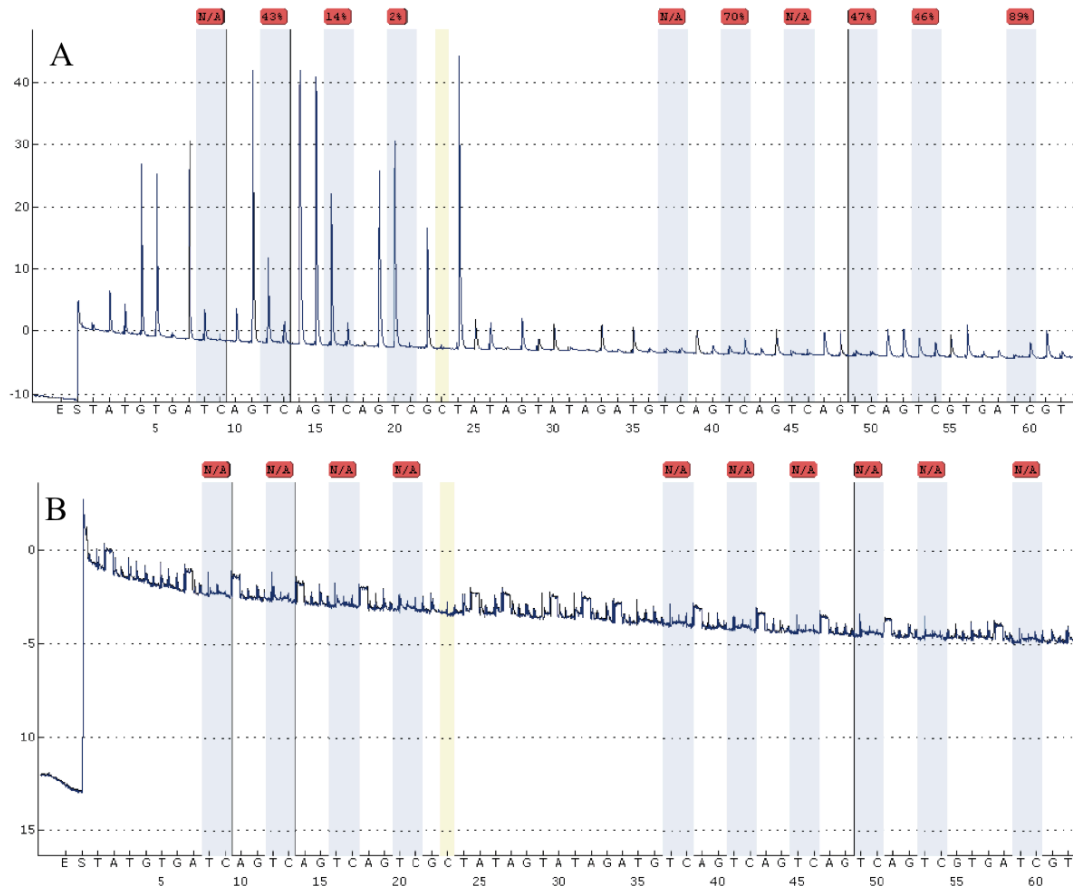
Figure 4.6 – (**A**) Pyrograms for the PFN3A locus showing a DNA sample from cattle as an example of how red warnings are displayed by the software when the sequence in the pyrogram does not correspond to the sequence expected. (**B**)Pyrogram for the PFN3A locus showing the absence of peaks from a DNA sample extracted from a bacterial pool containing the following species: *Escherichia coli, Staphylococcus aureus, Enterococcus faecali* and *Pseudomonas aeruginosa*. The DNA samples from dog, chicken, mouse and horse show similar pyrograms to (B)

Our data showed that the locus PFN3A is specific for primates and does not provide

pyrosequencing data for bacterial DNA. The fact that the chimpanzee DNA provided a

pyrogram comparable to the human DNA is expected because of the high level of genetic

102

similarity between the two species and does not constitute a major drawback in the application of this method.

E. Concluding Remarks

This report demonstrates how the quantification of DNA methylation levels measured by pyrosequencing on the PFN3A locus can successfully discriminate vaginal epithelia from samples with as little as 5 ng of input DNA. Moreover we were able to assess methylation levels when mixtures of body fluids were present and also performed species-specific studies which showed that DNA from bacteria are not amplified using the primers designed for PFN3A locus.

CHAPTER V – VALIDATION OF FORENSIC MARKERS BY PYROSEQUENCING AND SNAPSHOT

This chapter contains two validation studies performed with the purpose of determining the feasibility of two methods and several loci to be used in forensic sciences to determine body fluid of origin based on DNA methylation levels.

The first article focuses on the use of pyrosequencing to validate the loci ZC3H12D, BCAS4 and cg06379435 for samples that represent mock crime scene samples, according to the SWGDAM guidelines. I collaborated in this study by evaluating the stability of methylation in samples stored for several years at room temperature.

The second study uses SNaPshot methodology to determine methylation levels on 7 CpGs. The study was a collaboration exercise designed and led by Yonsei University College of Medicine in South Korea to evaluate whether the same set of results would be obtained when using SNaPshot in a set of samples prepared by the leading laboratory. I collaborated in this study by preparing and running the samples sent by the leading laboratory, followed by data analysis and reporting of results to be included in the manuscript.

A. Developmental Validation Studies of Epigenetic DNA Methylation markers for the detection of Blood, Semen and Saliva samples

A.1 Abstract

Determining the type and origin of body fluids in a forensic investigation can provide important assistance in reconstructing crime scenes. A set of epigenetic markers, ZC3H12D, BCAS4 and cg06379435, have been developed to produce unique and specific patterns of DNA methylation that can be used to identify semen, saliva, and blood, respectively. To ensure the efficacy of these markers, developmental validation studies were performed to determine the conditions and limitations of this new tool for forensic analysis. DNA was extracted from human samples and bisulfite modified using commercial bisulfite modification kits. Specific primers were used to amplify the region of interest and the methylation profile of the CpG sites were determined by pyrosequencing. The percent methylation values at each CpG site were determined in multiple samples and averaged for each tissue type. The versatility of these new markers is presented by showing the results of validation studies on sensitivity, human specificity, stability and mixture resolution. When testing the markers using different organisms, we did obtain positive results for certain non-human primate samples, however, all other tested species were negative. The lowest concentration consistently detected varied from

0.1 to 10ng, depending on the locus, indicating the importance of primer design and sequence in the assay. The method also proved to be effective when inhibitors were present in the samples or when samples were degraded by heat. Simulated case- samples were also tested. In the case of mixtures of different cell types, the overall methylation values varied in a consistent and predictable manner when multiple cell types were present in the same sample. Overall, the validation studies demonstrate the robustness and effectiveness of this new tool for body fluid identification.

A.2 Introduction

Body fluids recovered from crime scenes are considered among the most important types of evidence in forensic cases. They contain DNA evidence that may identify a suspect or victim and exonerate an innocent person. Moreover, determining the type and origin of a biological material can help reconstruct crime scenes (119, 175, 177). Identifying body fluids can be a critical factor in criminal investigations as the presence of skin cells may indicate innocent transfers of DNA while blood or other body fluids can indicate that a criminal act has taken place. However, many times it can be difficult to identify a forensic stain because of its similarities with other fluids or substances. Additionally, when trace levels of body fluids are present, visualization may be impossible and serological tests can fail to produce a result (119). Recent efforts in researching new assays to identify body fluids have focused on epigenetic DNA based markers. While RNA and protein markers may be used for body fluid identification, DNA methylation presents the ideal methodology since it provides quantitative results, is less prone to variations in expression, and has greater long term stability (140, 147). In

106

addition, as the extracted DNA target is already present in the laboratory, only minor

sample processing is needed. This is important, since in many cases the forensic sample

available for analysis is limited. DNA methylation is an epigenetic modification that is

involved in transcriptional regulation. A typical mammalian genome has approximately 3

billion base pairs, with a G/C content of about 40%. Methylated cytosines typically occur

as CpG dinucleotides and the presence of a methyl group at the C-5 position of the

molecule tends to reduce gene expression. The CpG dinucleotide distribution is uneven,

with several short DNA elements having a much higher density of CpG dinucleotides

than other regions of the genome, forming so-called CpG islands. Most of these islands

are located near transcription start sites. The methylated form, 5-methylcytosine (5-mC)

correlates with cytosine in the same way that thymine correlates with uracil, with no

effect on the base pairing (147, 178, 179). Although much is still to be understood about

the mechanisms by which methylation affects gene expression, it is known that

methylation is important in cell differentiation, and genomic loci are differentially

methylated between tissues. Because of this, differentiation of methylation patterns can

provide the basis of an assay for body fluid identification.

The most common body fluids found at crime scenes are blood, semen and saliva. A set

of epigenetic markers have been developed, ZC3H12D, BCAS4, and cg06379435, which

produce unique and specific patterns of DNA methylation that can be used to identify

semen, saliva and blood, respectively (156, 176). The process involves bisulfite

modification, followed by amplification by the polymerase chain reaction (PCR), and

pyrosequencing. The results permit quantitative determination of methylation patterns at

each locus for each body fluid type. The data demonstrate the applicability of epigenetic

markers as an assay for trace body fluid identification. However, to ensure the efficiency of these epigenetic markers, developmental validation studies need to be performed to determine the conditions and limitations of this new tool for forensic analysis.

A.3 Materials and Methods

Sample collection

Biological samples (semen, blood and saliva) were obtained from volunteer donors. Buccal cells were collected using cotton swabs. Blood was obtained by finger pricking and blood cells were collected using cotton swabs. Male volunteers were given a specimen jar and requested to donate a semen sample. All participants signed informed consent statements prior to sample collection. All samples were collected after obtaining appropriate Institutional Review Board approvals from Florida International University (IRB-13-0555), The University of Southern Mississippi (protocol #12010303) and Pontifícia Universidade Católica do Rio Grande do Sul (CONEP #723.619/CEP #845.747).

DNA extraction and modification

DNA extraction was performed using the BioRobot1 EZ1 automated purification workstation (Qiagen, CA) and the EZ1 DNA Investigator kit according to manufacturer's protocol. DNA was recovered in a final volume of 50 ml and quantified using PicoGreen1 fluorescence (Life Technologies, CA) (156). Because standard PCR amplification does not preserve DNA methylation sites, genomic DNA extracts were treated with sodium bisulfite in order to convert unmethylated cytosines to uracil using

the EpiTect® Fast DNA Bisulfite Kit (Qiagen Inc, CA). For all tests conducted in this research, except for the sensitivity test, the standard manufacturer's protocol for the conversion of around 400 ng DNA in a maximum volume of 40 ml was used. In this method the converted uracils are then replaced by thymine during the PCR amplification process. The bisulfite modified DNA was amplified by site-specific PCR primers designed to amplify the bisulfite modified target region.

Markers

As previously described, a set of epigenetic markers, ZC3H12D, BCAS4, and cg06379435 have been developed to produce unique and specific patterns of DNA methylation that can be used to identify semen, saliva, and blood, respectively. Markers ZC3H12D and BCAS4 were described by Madi *et al.* (156) and both targeted 5CpG sites in the sequence to be analyzed. Marker cg06379435 was described by Park *et al.* (176) and targeted only one CpG site. We have investigated this locus and added 4 additional CpG sites around the cg06379435 site. BCAS4 and cg06379435 PyroMark® CpG assays were custom designed in house using the Pyromark® assay design software (Qiagen). A predesigned ZC3H12D PyroMark® CpG assay was available from the Qiagen GeneGlobe Web Portal (Table 5.1). The reverse primers were biotin labeled allowing for the production of biotinylated PCR products necessary for the pyrosequencing reaction (156).

PCR and pyrosequencing reactions

Singleplex PCR reactions were performed using the PyroMark® PCR Kit (Qiagen) in a GeneAmp® PCR System 9700 (Applied Biosystems, Foster City, CA). For all samples, 1.2 ml of modified DNA template was added to the PCR mixture in a total reaction volume of 15 ml. The complete PCR protocol outlined by the manufacturer was followed (Qiagen). Pyrosequencing was performed following completion of the PCR reactions using a Pyromark® Q24 pyrosequencer (Qiagen) following the protocols recommended by the manufacturer. Once the pyrosequencing was complete, the percent methylation was calculated automatically by the Pyromark® Q24 software and was displayed as a pyrogram (156).

Validation studies

We tested 15 samples of each cell type (blood, saliva and semen) to be used in these studies. The pyrosequencing data generated for each marker was analyzed and compared to other samples in the validation studies performed to ensure they do not differ statistically.

Body fluid specificity

The specificity of the markers was previously tested against blood, semen and saliva (156, 176). In this study we tested the performance of the markers against 3 samples of each: menstrual blood vaginal epithelia and nasal epithelia/secretion.

Table 5.1. Panel of markers used in this study.

Marker	PCR and sequencing primers	CpG sites to be analyzed in the nucleotide dispensation order (underlined)
BCAS4	**Forward primer-** AGTGGGTGAGGTTGTGAAATGT **Reverse primer-** CCCATCCTACTAAAACATCTAATT **Sequencing primer-** AGTTTTTTGGTGAAGTTTAT	TAT**CG**TATCGAGAGAGATCGAGAT**CG**TATAT**C G**TATAGAT**CG**T**CG**AT
cg06379435[a]	**Forward primer-** AGTAGAGGTGGGGGTTAATAATT **Reverse primer-** ACCACACAACAAAACAACTATCTC **Sequencing primer-** GTTAGGAAAGAAAAATGTAATTTA	T**CG**GGATAAT**CG**GTGGAATTTTAGG**CG**TGGGA**CG**GTTGT**CG**GA
ZC3H12D	Proprietary[b]	T**CG**T**CG**AGTAT**CG**T**CG**T**CG**

[a] TargetID is a unique ID in the Illumina Human Methylation 450K bead array; 4th CpG site in the sequence provided is the target in Park et al (176).
[b] Primer sequences are proprietary (Qiagen Inc.) and not available.

Species specificity

Species specificity studies were performed with non-human DNA samples of dog, cat, mouse, chicken, bovine, equine, pig, mouse, chimpanzee, orangutan, gorilla and a microbial pool (*Escherichia coli*, *Staphylococcus aureus*, *Enterococcus faecali* and *Pseudomonas aeruginosa*). A human blood sample was used as positive control.

<u>Sensitivity studies</u>

According to the SWGDAM guidelines (108) it is important to evaluate the limits

of DNA quantities to be used in the reaction in order to obtain reliable results. Madi *et al.*

(156) processed a saliva sample five separate times and, for each replicate, the amount of

DNA subjected to bisulfite modification was varied (385, 100, 50, 10 and 5 ng). The

samples were amplified with the ZC3H12D PyroMark® CpG Assay. All replicates

showed no significant differences in the methylation levels observed at each CpG site.

Park *et al.* (176) tested marker cg06379435 and were able to obtain successful

pyrosequencing results when using 500–10 ng of premodified DNA. However, when

working with forensic samples, it is important to test the assay for its sensitivity limits

and to examine the lowest levels of DNA possible while maintaining reliability in the

results. For this validation study we therefore decided to test the following DNA

quantities: 10, 5 and 1 ng. Three distinct samples for each body fluid were processed for

each DNA amount tested.

<u>Stability studies and case-type samples</u>

Forensic samples are often exposed to environmental and chemical insults, and

may contain impurities, which can act as PCR inhibitors. In this study, we simulated

casework samples to evaluate the robustness and stability of the markers. To assess the

effect of the inhibitors on PCR, we tested 2 different solutions: hematin (100 mM in 0.1

N sodium hydroxide) and humic acid (1 mg/ml in water). From these stock solutions, we

prepared subsequent dilutions in water in a final concentration of 0.08 mM (hematin) and

0.24 mg/ml (humic acid) (180). Aliquots of 2 ml of the inhibitors were added before or

after bisulfite conversion. For this study, we tested three distinct samples for each body fluid. To test the stability of the markers under degraded conditions, samples with genomic DNA extracts were heated in deionized water at 95 ºC for 10, 15, 20, and 25 min to simulate natural DNA fragmentation (180). For this study, we tested three distinct samples for each body fluid. Also, DNA from the following simulated case-type samples was extracted and evaluated for all tested markers:

    i)   200 ml of blood in 100% cotton fabric

    ii)   200 ml of semen in 100% cotton fabric

    iii)   Saliva swab of lid of a coffee drink

Mixture studies

DNA mixtures were prepared in order to determine if there are differences in methylation values when multiple cell types were present in the same sample. DNA from blood, saliva and semen were tested in different ratios of mixtures:

    i)   Blood + saliva: ratios of 75% blood: 25% saliva—50% blood: 50% saliva—25% blood: 75% saliva

    ii)   Blood + semen: ratios of 75% blood: 25% semen—50% blood: 50% semen—25% blood: 75% semen

    iii)   Saliva + semen: ratios of 75% saliva: 25% semen—50% saliva: 50% semen—25% saliva: 75% semen

    iv)   Blood + saliva + semen: ratios of 75% blood: 12.5% saliva: 12.5% semen—50% blood: 25% saliva: 25% semen—75% saliva: 12.5% blood: 12.5% semen—50% saliva: 25% blood: 25% semen—75% semen: 12.5% blood:

12.5% saliva—50% semen: 25% blood: 25% saliva In addition to these

mixtures, each sample was analyzed individually as a positive control.

Reproducibility

This study was performed by testing 24 samples (saliva n = 8, blood n = 8, semen

n = 8). The pyrosequencing data generated by our laboratory (Florida International

University) and an independent laboratory (University of Southern Mississippi) was

analyzed and compared to assess the reliability and the production of concordant results

between different laboratories (181).

Data analysis

The CpG sites were examined and, for each marker, the methylation values were

averaged and listed for each cell type along with standard deviations. Using a one-way

ANOVA, the methylation values were compared across CpG sites of each marker

between controls and other samples to determine if they differed significantly. A t-test

was used to compare results obtained by the two different laboratories in the

reproducibility study. A p-value of <0.05 was assumed as significant for all tests

conducted.

A.4 Results and Discussion

Control samples

A set of 15 samples of each cell type (blood, saliva and semen) were analyzed and

methylation profiles of each locus were obtained to be used as controls. Table 5.2

presents the pyrosequencing data generated for each marker.

Table 5.2. Methylation profiles of the three loci investigated in this study.

Marker	Body Fluid	CpG (Mean % Methylation ± SD)				
		CpG 1	CpG 2	CpG 3	CpG 4	CpG 5
ZC3H12D	Semen	5.4±4.0	5.3±3.9	6.7±4.3	6.3±4.0	5.1±3.7
Specific for	*Threshold*	*14*	*13*	*15*	*14*	*13*
semen	Blood	94±1.6	94±2.7	100±0.0	97±7.9	86±3.5
	Saliva	81±4.1	78±4.5	99±2.1	79±4.9	82±3.4
		CpG 1	CpG 2	CpG 3	CpG 4	CpG 5
cg06379435	Blood	24±7.8	22±6.7	33±7.4	30±8.2	49 ±12
Specific for	*Threshold*	*8.2*	*8.4*	*18*	*14*	*25*
blood	Semen	3.4±1.7	2.4±1.8	2.7±1.1	1.8±1.3	3.13±1.2
	Saliva	8.7±7.0	2.6±1.4	6.0±3.9	3.5±2.6	7.7±4.7
		CpG1	CpG4	CpG5	CpG6	CpG7
BCAS4	Saliva	64±7.1	27±5.6	16±4.5	45±6.7	11±2.7
Specific for	*Threshold*	*49*	*16*	*7.2*	*31*	*6.1*
saliva	Blood	6.1±1.4	3.2±2.8	2.7±1.3	6.4±1.9	2.4±1.6
	Semen	3.9±1.6	2.5±1.1	3.0±1.1	5.6±1.6	2.3±0.9

*Each CpG site from each marker is given a specific threshold value due to the fact that each CpG has its unique level of methylation. A threshold value is assigned as the average methylation level ± 2 S.D depending on whether hypo or hyper methylation is present (182).

Body fluid specificity

Besides previously testing the performance of the markers against blood, semen and

saliva, we also tested the markers against menstrual blood, vaginal epithelia and nasal

epithelia/secretion (Table 5.3). Most of the CpG sites presented significant differences in

the methylation levels when compared with the different types of fluids. For marker BCAS4, the methylation level of CpG5 did not significantly differ when saliva samples were compared to vaginal epithelia samples ($p > 0.05$). For marker cg06379435, the methylation level of CpG2 did not significantly differ when blood samples were compared to menstrual blood samples ($p > 0.05$). However, all CpGs within each marker must be evaluated to obtain an indication of the body fluid present. Our results show that the majority of CpGs within each marker presents different mean methylation levels between the target body fluid and all other body fluids tested (Table 5.3).

In this study, we calculated the threshold as ±2 SD from the average methylation percent for each CpG, according to what has been published in terms of DNA methylation and threshold calculation (182). However, in our study multiple CpG sites are available at each locus for each body fluid; therefore we suggest that the user can adjust the threshold levels to ±1 SD, making the test more discriminatory. At present, in order to have a conclusive result for the identification of a body fluid, all CpG sites of a marker should reach the threshold value; otherwise the test should be indicated as inconclusive.

This study represents the first validation for the use of DNA methylation patterns in specific genome locations for body fluid discrimination. In order to achieve a higher discriminatory power for a wider range of body fluids, additional genome locations need to be determined using array technology and NGS and validated using pyrosequencing. With the advancements in next-generation sequencing we estimate that the number of research studies addressing this issue for forensic purposes will increase.

Table 5.3. Methylation profiles of the three investigated loci when testing menstrual blood, vaginal epithelia and nasal epithelia/secretion samples. Comparison with methylation levels of the three main fluids investigated in this study.

Marker	Body Fluid	CpG (Mean % Methylation ± SD)				
		CpG 1	**CpG 2**	**CpG 3**	**CpG 4**	**CpG 5**
ZC3H12D *Specific for semen*	Semen	5.4±4.0	5.3±3.9	6.7±4.3	6.3±4.0	5.1±3.7
	Blood	94±1.6	94±2.7	100±0.0	97±7.9	86±3.5
	Saliva	81±4.1	78±4.5	99±2.1	79±4.9	82±3.4
	Menstrual Blood	80±4.7	79±6.2	98±2.9	77±12	74.3±3.5
	Nasal Epithelia	83±6.4	83±6.8	99±1.2	78±11	77.0±6.9
	Vaginal Epithelia	77±9.7	77±10.0	97±2.5	73±10	80.0±4.0
		CpG 1	**CpG 2**	**CpG 3**	**CpG 4**	**CpG 5**
cg06379435 *Specific for blood*	Semen	3.4±1.7	2.4±1.8	2.7±1.1	1.8±1.3	3.13±1.2
	Blood	24±7.8	22±6.7	33±7.4	30±8.2	49±12.2
	Saliva	8.7±7.0	2.6±1.4	6.0±3.9	3.5±2.6	7.7±4.7
	Menstrual Blood	10±12	12±19[a]	12±18	8.0±12	23±30
	Nasal Epithelia	5.7±5.5	5±6.9	9.7±11	7.7±11	18±22
	Vaginal Epithelia	2.3±0.6	1.3±0.6	3.0±1.0	2.0±1.0	8.3±7.8
		CpG1	**CpG4**	**CpG5**	**CpG6**	**CpG7**
BCAS4 *Specific for saliva*	Semen	3.9±1.6	2.5±1.1	3.0±1.1	5.6±1.6	2.3±0.9
	Blood	6.1±1.4	3.2±2.8	2.7±1.3	6.4±1.9	2.4±1.6
	Saliva	64±7.1	27±5.6	16±4.5	44±6.7	11±2.7
	Menstrual Blood	35±8.5	15±4.6	11±3.1	27±5.5	8.0±1.0
	Nasal Epithelia	39±20.5	12±8.0	9±3.6	33±20	6.7±3.5
	Vaginal Epithelia	36±16.7	13±7.5	11±4.9[b]	30±11	6.7±3.2

[a]The methylation level of CpG2 did not significantly differ when menstrual blood samples were compared to blood samples (p>0.05).
[b]The methylation level of CpG5 did not significantly differ when vaginal epithelia samples were compared to saliva samples (p>0.05).

Species specificity

Several non-human DNA samples were tested to assess the specificity of the

markers. Non-human primate samples amplified the target regions but not all showed

pyrosequencing results. For marker cg0679435, an error occurred as the surrounding

reference sequence was not recognized in all primate samples tested. Also, the orangutan

sample did not sequence for marker ZC3H12D. All other primate samples presented

pyrograms similar to those obtained when testing a human sample. These results are

expected since there is a close evolutionary relationship between these animals and

humans (181). In contrast, the other species used in this validation study, which are more

evolutionarily distant to humans, failed to show any results. The specificity results

obtained here are sufficient to enable the application of these markers in forensic

casework.

Sensitivity studies

The sensitivity was assessed to determine the minimum quantity of DNA with

which accurate and reliable results could be obtained. For markers BCAS4 and

cg06379435, we were able to obtain successful pyrosequencing results when using 10 ng

of pre-bisulfate modified DNA. When the samples were amplified for the ZC3H12D

marker, good results were obtained using 1 ng of pre-modified DNA. When the input pre-

modified DNA was lowered to 0.1 ng, good quality methylation profiles were obtained

(Table 5.4). All low template samples for the three markers were compared with the

control samples and no significant differences were found. For most of the samples

tested, the amplification with the markers BCAS4 and cg06379435 was a success, and we

were able to see bands with the correct size on a 2% agarose gel. However, after

pyrosequencing, the percent methylation of samples with DNA inputs lower than 10 ng

differed from the values presented by the control samples. When working with bisulfite-

treated DNA, this so-called PCR-bias is often observed in the amplification step. When

using low amounts of DNA, a more efficient amplification can occur with unmethylated

alleles when compared to those which are methylated.

Table 5.4. Methylation profiles of the three investigated loci when using low amounts of pre-modified DNA.

Marker	Samples	CpG (Mean % Methylation ± SD)				
ZC3H12D		**CpG 1**	**CpG 2**	**CpG 3**	**CpG 4**	**CpG 5**
marker for semen	Semen control samples[a]	5.4±4.0	5.3±3.9	6.7±4.3	6.3±4.0	5.1±3.7
	Semen samples (DNA input=0.1ng)	4.7±4.0	4.7±3.8	4.3±3.5	3.7±2.1	3.0±4.4
cg06379435		**CpG 1**	**CpG 2**	**CpG 3**	**CpG 4**	**CpG 5**
marker for blood	Blood control samples[a]	23.8±7.8	21.8±6.7	33.2±7.4	30.4±8.2	49 ±12
	Blood samples (DNA input=10ng)	23±8.7	22±10	26.3±8.3	23±9.5	36.3±7.5
BCAS4		**CpG1**	**CpG4**	**CpG5**	**CpG6**	**CpG7**
marker for saliva	Saliva control samples[a]	63.6±7.1	27.5±5.6	16.3±4.5	44.5±6.7	11.5±2.7
	Saliva samples (DNA input=10ng)	65.7±6.0	29.3±4.9	16.0±1.0	44.0±9.5	10.3±6.5

[a]DNA input of control samples ranged from 100 to 500 ng.

Alternatively, in some cases, an inverse bias can occur with a deviation toward the

methylated alleles (183). With regard to the samples amplified with the ZC3H12D

marker, it was possible to keep the consistency in the results even when using low DNA

templates. According to Moskalev *et al.* (183) the primer design of some markers may

lead to a more unbiased amplification of bisulfite-treated DNA. However we cannot

prove this assumption for the ZC3H12D PyroMark® CpG assay, since its sequence is

proprietary and not provided by the manufacturer of the assay (Qiagen, CA).

Stability study and case-type samples
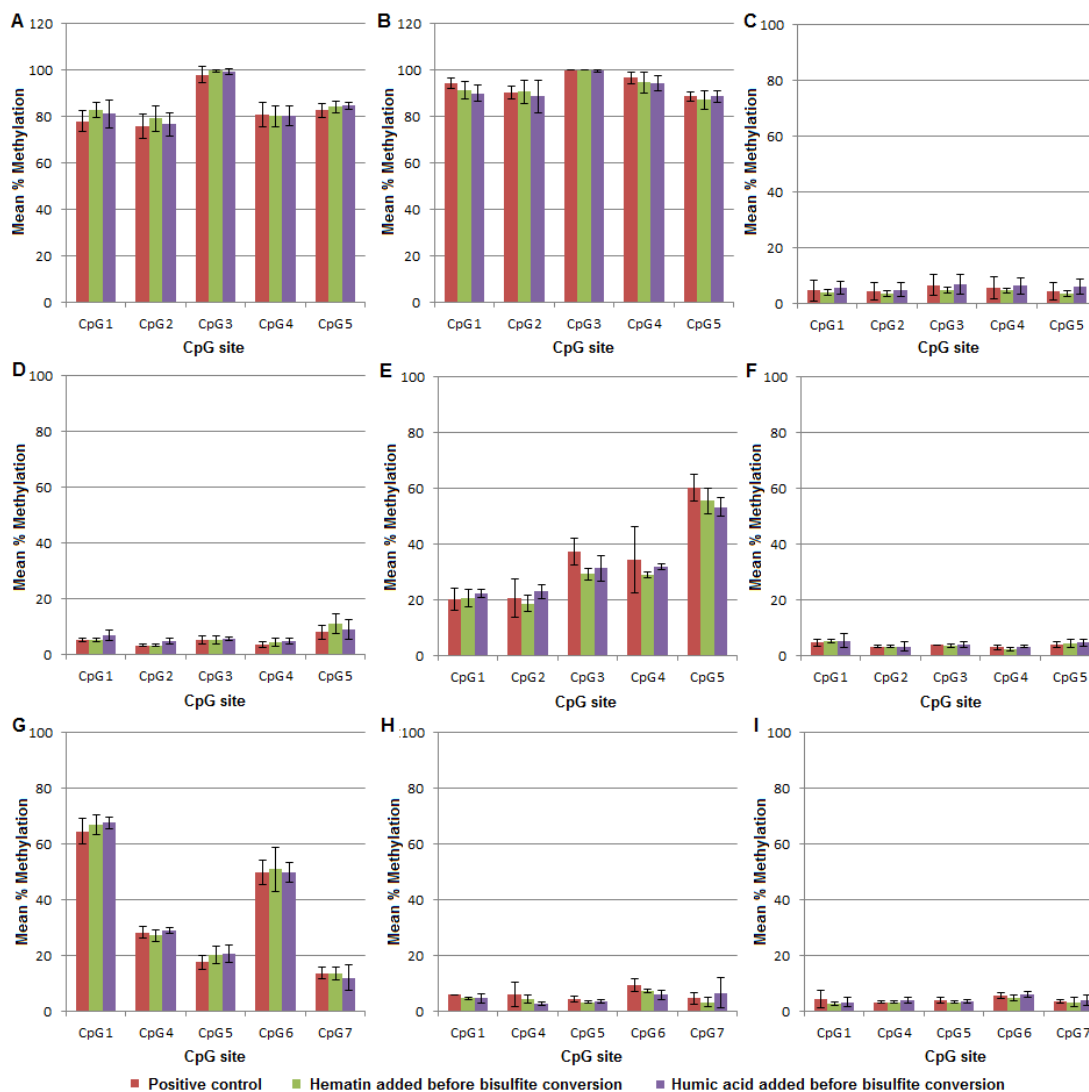
Figure 5.1. Mean methylation levels of samples in which PCR inhibitors were added before bisulfite modification. There were no significant differences between the control and the tested samples. **A)** Marker ZC3H12D: saliva samples. **B)** Marker ZC3H12D: blood samples. **C)** Marker ZC3H12D: semen samples. **D)** Marker cg06379435: saliva samples. **E)** Marker cg06379435: blood samples. **F)** Marker cg06379435: semen samples. **G)** Marker BCAS4: saliva samples. **H)** Marker BCAS4: blood samples. **I)** Marker BCAS4: semen samples.

When dealing with forensic casework samples, it is common to encounter

degraded and contaminated samples. Therefore, we have investigated the effects of PCR

inhibitors on the methylation levels. The final inhibitor concentrations tested were: hematin (0.08 mM) and humic acid (0.24 mg/ml).

We could observe that when these inhibitors were added prior to bisulfite modification, all samples were amplified and there was no significant difference between the methylation values of the control and the other tested samples (Figure 5.1). However, when inhibitors were added after bisulfite modification, the amplification process failed and the pyrosequencing results were similar to those obtained when testing a no template control (water).

According to the data provided by the manufacturer (Qiagen, CA), the bisulfite modification kit provides bisulfite conversion as well as cleanup of the modified DNA for subsequent methylation analysis. The cleanup step is likely the reason that PCR inhibitors had no discernable effect on the PCR amplification that followed bisulfite modification. When testing the stability of the markers under degraded conditions, we noticed that all primers were still able to amplify the target sequence. There were no significant differences in the methylation values for the degraded samples versus intact DNA (positive control) (Figure 5.2). Also, all three simulated case-type samples were successfully modified, amplified and pyrosequenced. Table 5.5 presents the pyrosequencing data obtained for each tested sample.
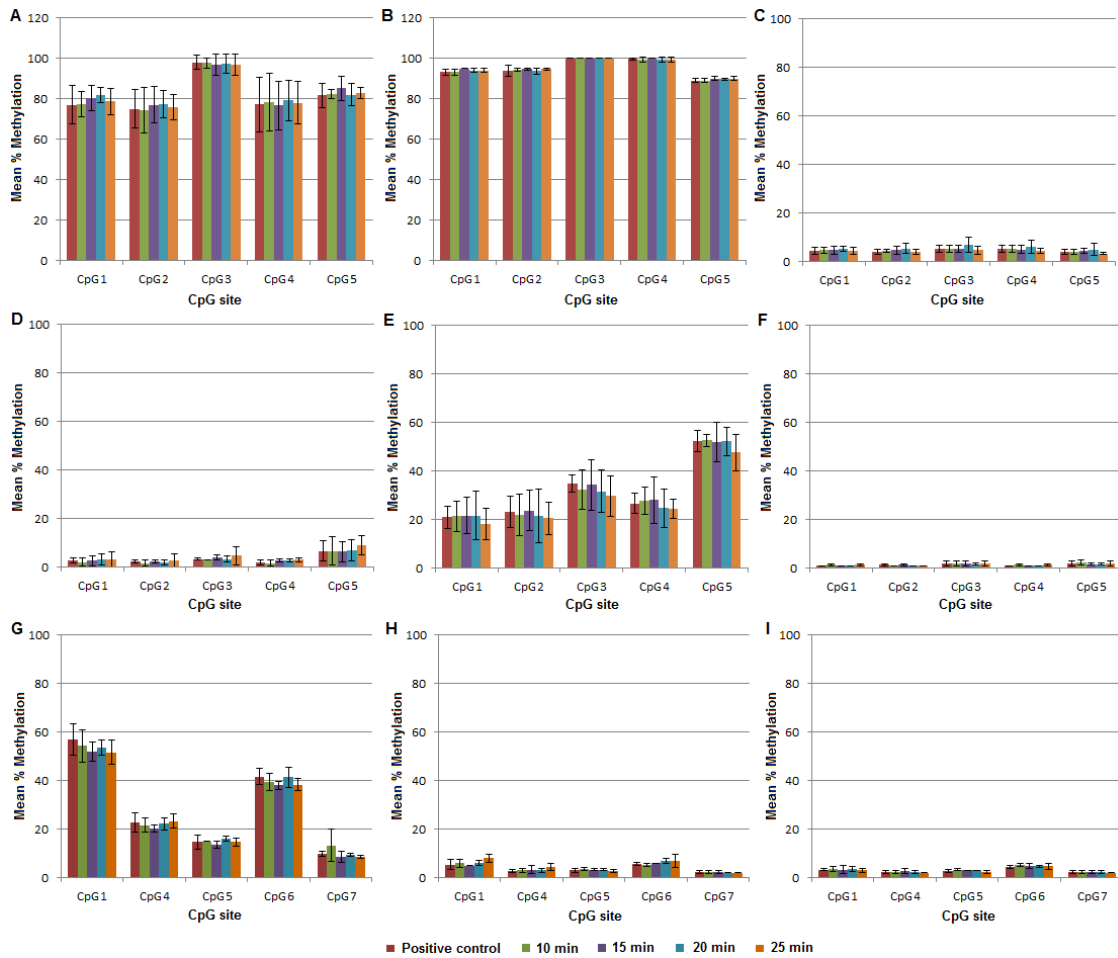
Figure 5.2. Mean methylation levels of samples degraded by heat with different time points. There were no significant differences between the control and the degraded samples. **A)** Marker ZC3H12D: saliva samples. **B)** Marker ZC3H12D: blood samples. **C)** Marker ZC3H12D: semen samples. **D)** Marker cg06379435: saliva samples. **E)** Marker cg06379435: blood samples. **F)** Marker cg06379435: semen samples. **G)** Marker BCAS4: saliva samples. **H)** Marker BCAS4: blood samples. **I)** Marker BCAS4: semen samples.

Table 5.5. Methylation levels of simulated case-type samples.

Marker	Samples	CpG (% Methylation)				
		CpG1	**CpG2**	**CpG3**	**CpG4**	**CpG5**
ZC3H12D *marker for semen*	200 µl of semen in 100% cotton fabric	15	10	16	14	13
	200 ul of blood in 100% cotton fabric	93	96	100	99	85
	Saliva swab of lid of a coffee drink	72	73	100	79	85
		CpG1	**CpG2**	**CpG3**	**CpG4**	**CpG5**
cg06379435 *marker for blood*	200 µl of semen in 100% cotton fabric	10	7	9	5	15
	200 µl of blood in 100% cotton fabric	23	20	36	33	49
	Saliva swab of lid of a coffee drink	5	6	10	7	17
		CpG1	**CpG4**	**CpG5**	**CpG6**	**CpG7**
BCAS4 *marker for semen*	200 µl of semen in 100% cotton fabric	5	5	5	11	3
	200 µl of blood in 100% cotton fabric	6	7	6	7	6
	Saliva swab of lid of a coffee drink	50	19	17	70	24

Mixtures

Determining the presence of a mixture is considered a difficult task, even in a case of human identification using short tandem repeats. To determine which body fluids are present in a mixture containing multiple tissue sources is even more arduous. In this study, we prepared different ratios of DNA mixtures to analyze if there were differences in the methylation values when multiple cell types are present in the same sample. For all markers, regardless of whether a mix of two or three body fluids are present (semen, blood and saliva) in one sample, it is possible to recognize that a mixture is present. The

methylation levels for mixtures show intermediate percentages when com-pared to pure samples (Figure 5.3).

In this study, with the analysis of a single specific marker for a particular body fluid, we were not able to perform complete mixture deconvolution. Additional genome locations will need to be identified and analyzed in order to fully deconvolute this data, however our results demonstrate that mixtures can be detected for sample discrimination.
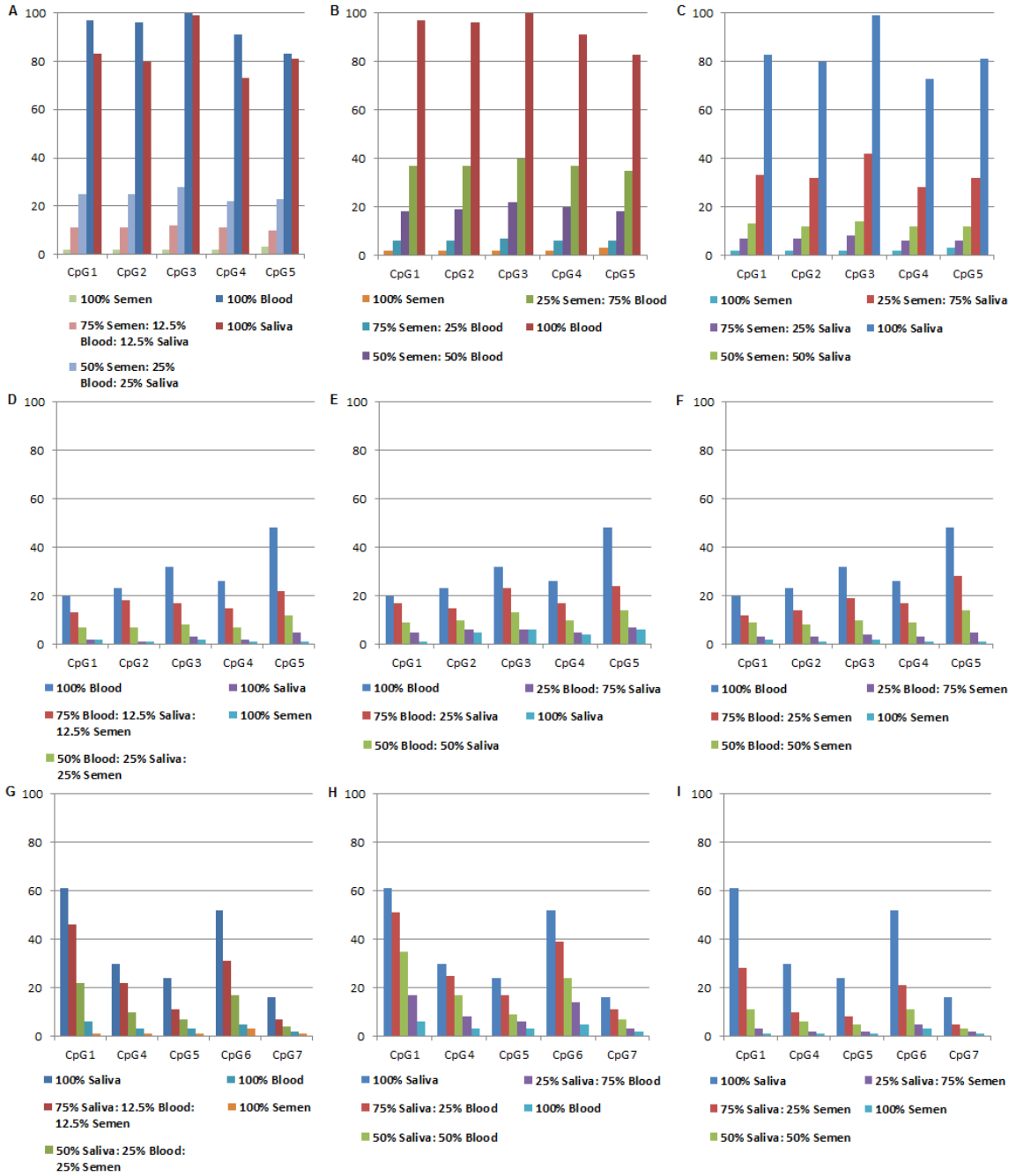
Figure 5.3. Mean methylation levels of different ratios of DNA mixtures. **A)** Marker ZC3H12D: semen + blood+ saliva. **B)** Marker ZC3H12D: semen + blood. **C)** Marker ZC3H12D: semen + saliva. **D)** Marker cg06379435: blood + saliva+ semen. **E)** Marker cg06379435: blood + saliva. **F)** Marker cg06379435: blood + semen. **G)** Marker BCAS4: saliva + blood+ semen. **H)** Marker BCAS4: saliva + blood. **I)** Marker BCAS4: saliva + semen.

A concordance study was performed in two independent laboratories by operators with different experience levels. No significant difference was found when we compared the methylation values for each CpG site obtained by both laboratories (t test, $p > 0.05$) (Figure 5.4).
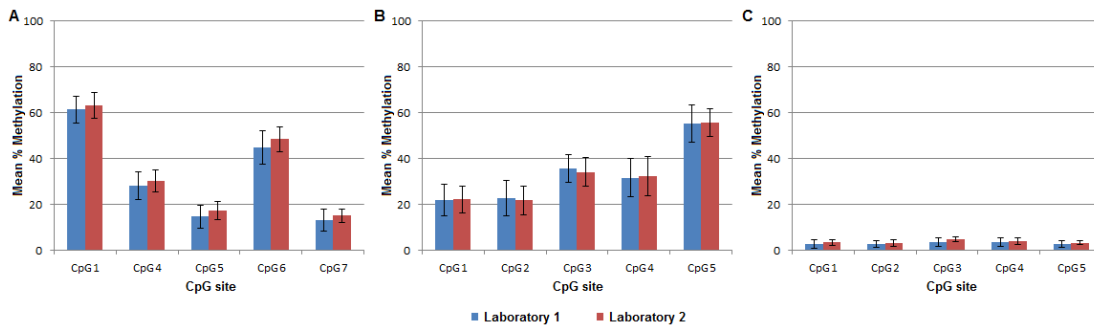
Figure 5.4. Comparison between pyrosequencing data generated by two independent laboratories: Laboratory 1 (Florida International University) and Laboratory 2 (University of Southern Mississippi). **A)** Methylation data of marker BCAS4 (saliva samples n=8). **B)** Methylation data of marker cg06379435 (blood samples n=8). **C)** Methylation data of marker ZC3H12D (semen samples n=8).

A.5 Conclusion

Body fluids recovered from crime scenes can provide crucial information for an investigation, particularly in situations where the presence of a suspects DNA may not in dispute, such as sexual assault and other types of violent crime. Epigenetic markers permit the identification of these biological fluids from trace levels of extracted DNA making them a powerful investigative tool. The validation studies presented here demonstrate robustness and reliability of the tested markers. We present results demonstrating, sensitivity, species specificity, resistance to degradation and the effect of mixtures. Our results demonstrate the capability of these new epigenetic markers in the

determination of trace levels of body fluids. They provide the community with specific and sensitive epigenetic methods that can add important information to the trier of fact in forensic analysis.

B. A Collaborative Exercise on DNA Methylation-based Body Fluid Typing

B.1 Abstract

A collaborative exercise on DNA methylation-based body fluid identification was conducted by 7 laboratories. For this project, a multiplex methylation SNaPshot reaction composed of 7 CpG markers was used for the identification of blood, saliva, semen, menstrual blood and vaginal fluid. A total of 30 specimens were prepared and distributed to participating laboratories after thorough testing. Required experiments included tasks divided into 4 parts of increasing complexity ranging from capillary electrophoresis of purified single-base extension reaction product, multiplex PCR of bisulfite-modified DNA, bisulfite conversion of genomic DNA and extraction of genomic DNA from body fluid samples. In steps dealing with bisulfite converted DNA, genomic DNA and body fluid samples, one or more mixtures were included, and specimens of indicated or unnamed body fluid sources were used. All of the laboratories, except one that failed to

set up appropriate condition for the capillary analysis of reference single-base extension products, generated consistent body fluid typing results for specimens of bisulfite converted DNA and genomic DNA. This result supports the successful incorporation of DNA methylation-based body fluid identification method to forensic laboratory workflow. Variation in the values obtained for DNA methylation between laboratories increases with the complexity of the required experiments, but all laboratories coincided in the interpretation of produced DNA methylation profiles. Although a test for overall sensitivity and the establishment of interpretational guidelines on DNA methylation-based body fluid identification remain to be performed, this study adds to the support for the possibility of successful application of DNA methylation profiling to forensic body fluid typing.

B.2 Introduction

Body fluid identification can be of significant importance in solving crimes by assisting the triers of fact in the determination of potential scenarios that may have led to the deposition of evidentiary material (119). Recently, molecular approaches that permit the detection of specific messenger RNA (mRNA) and micro RNA (miRNA) expression as well as differential DNA methylation patterns have been intensively investigated (184). Among these, the mRNA-based detection methods have been the most scrutinized, and performance in terms of reproducibility and sensitivity has been well evaluated through a series of large collaborative exercises (127, 184-189).

DNA methylation-based body fluid identification has many advantages when compared to mRNA-based methods, such as higher specificity, the ability to be

incorporated into current forensic DNA-based testing protocols, and the applicability to old cases where only DNA extracts are available (136). Many CpG markers have been identified which produce differential DNA methylation patterns when extracted DNA from body fluids are compared using genome-wide profiling and gene-specific analysis (147, 156, 166, 175, 176, 190). Three recent publications (166, 176, 190)are particularly notable for reporting a set of CpG markers that show a methylation signal only in the target body fluids, which can be beneficial to mixed sample analysis. In particular, two earlier studies (166, 176) reported two markers in common; cg17610929, suggested to distinguish semen and cg06379435, for blood. Moreover, further analysis of array data from the two studies revealed that they reported a third blood-specific marker in both studies, cg08792630. Because the identification and use of semen-specific CpG markers is not a problem because of the significant difference in DNA methylation between somatic cells and germ cells, further validation studies of CpGs specific to other body fluids, such as vaginal fluid, menstrual blood, saliva etc., would be needed for better application of DNA methylation analysis in forensic caseworks. To detect DNA methylation levels at multiple CpG sites simultaneously, methylation-sensitive restriction enzyme-PCR (MSRE-PCR) and the methylation-sensitive single nucleotide primer extension (SNuPE)-based approach, generally called methylation SNaPshot, have been utilized (166, 175, 190). In comparison to MSRE-PCR, methylation SNaPshot reactions provide better resolution, and have been adopted in more recent studies (146, 166, 190).

The present collaborative exercise was organized by Yonsei University College of Medicine in order to evaluate the robustness and reproducibility of DNA methylation profiling for body fluid identification in seven forensic laboratories using kits and

chemistries of their own choice and using their own instrumentation. The multiplex

methylation SNaPshot reaction used by the collaborative laboratories described in (166)

was modified to include seven CpG markers that show a methylation signal only in the

target body fluids. Since the participating laboratories had varying levels of experience,

especially with DNA methylation profiling using bisulfite conversion and with the

SNaPshot reaction, the goal was to implement the method and compare the

interpretational results of DNA methylation profiling from the various laboratories.

B.3 Materials and Methods

<u>Samples and materials provided</u>

The exercise was divided into four parts such that participants could easily check

the success of experiments performed after each step of the consecutive procedures

(Table 5.6). Parts 1 involved the capillary electrophoresis of purified single-base

extension (SBE) reaction products, part 2 involved multiplex PCR of bisulfite converted

DNA, part 3 involved the bisulfite conversion of genomic DNA and part 4 involved the

DNA extraction of body fluid samples.

As such, the complexity of the required experiments increased as the participants

proceeded through each part. Specimens, including blood, saliva, semen, vaginal fluid,

and menstrual blood, were collected from seven healthy volunteers using procedures

approved by the Institutional Review Board of Severance Hospital, Yonsei University in

Seoul, Korea.

Table 5.6 - Overview of the samples and experiments required in each part of collaborative exercise

Part	Samples	Required experiments
1	**Purified[a] SBE(single-base extension reaction) product[b]** 6 samples : for each cell type[c], single source samples, cell type indicated	Capillary electrophoresis
2	**Bisulfite converted DNA[d]** 5 samples : for each cell type[c], single source samples, cell type indicated 2 samples : single source samples, unspecified cell type 1 sample : mixture of two body fluids, unspecified cell type	Multiplex PCR, Multiplex SBE, Capillary electrophoresis
3	**Genomic DNA[e]** 5 samples : for each cell type[c], single source samples, cell type indicated 2 samples : single source samples, unspecified cell type 1 sample : mixture of two body fluids, unspecified cell type	Bisulfite conversion, Multiplex PCR, Multiplex SBE, Capillary electrophoresis
4	**Body fluid swabs** 4 samples : for each cell type[f], single source samples, cell type indicated 1 sample : mixture of two body fluids, cell type indicated 2 samples : single source samples, unspecified cell type 1 sample : mixture of body fluids, unspecified cell type	DNA extraction, DNA quantification, Bisulfite conversion, Multiplex PCR, Multiplex SBE, Capillary electrophoresis

[a] Purified with SAP-Recombinant enzyme
[b] Multiplex SBE reaction product
[c] Semen, blood, vaginal fluid, menstrual blood and saliva
[d] Bisulfite converted DNA using the Sigma's Imprint® DNA modification kit
[e] Genomic DNA extracted using the Qiagen's QIAamp DNA Mini kit and quantified using the ABI's Quantifiler® Duo DNA Quantification kit
[f] Semen, blood, vaginal fluid and saliva

Blood was collected by venipuncture with a syringe, and 200 µL aliquots from EDTA containing tubes were stored frozen at -20°C. Saliva samples were collected with the Oragene™ DNA self-collection kit (DNA Genotek Inc., Ottawa, Canada), and stored at

ambient temperature. Freshly ejaculated semen was collected in plastic cups, and 200 μL aliquots were frozen at -20˚C for storage. Vaginal fluid and menstrual blood were collected using sterile cotton swabs and allowed to dry at room temperature. The eight body fluid specimens used in part 4 included four single source samples and one mixture of indicated origin, and two single source samples and one mixture of unspecified origin. One hundred and twenty microliter of freshly ejaculated semen and blood were prepared separately in an eppendorf microfuge tube with indicated origin. A 200 μL of saliva sample was prepared in the Oragene™ DNA self-collection kit buffer with indicated origin. Two vaginal swabs and a menstrual blood swab were obtained from a volunteer. From the two vaginal swabs obtained from a volunteer, one was spiked with 70 μL of semen and both were provided to the collaborative laboratories with an indication of their origin. The menstrual blood swab was sent to the collaborative laboratories with unspecified origin. A semen swab was prepared by dropping 100 μL of semen on a sterile cotton swab, and was provided to the collaborative laboratories with unspecified origin. A body fluid mixture swab was prepared by adding 100 μL solution of blood, saliva and semen mixed in a ratio of 2:1:1.5 on a sterile cotton swab and was provided to the collaborative laboratories with unspecified origin. Treated swabs were allowed to dry at room temperature and stored at -80˚C until sent to participating laboratories.

For parts 2 and 3, 100 ng of bisulfite converted DNA and 200 ng of genomic DNA were prepared for five single source samples with an indicated source of origin, two single source samples of unspecified origin, and one mixture of unspecified origin. To prepare mixtures in parts 2 and 3, the same concentration of each bisulfite converted

DNA (from blood and saliva) or genomic DNA (from semen and vaginal fluid) were mixed in a ratio of 1:1.

For part 1, the final products from the methylation SNaPshot reaction obtained from one sample each of semen, blood, vaginal fluid and saliva, and two samples from menstrual bloods were prepared as reference materials.

All samples and PCR primer mixtures were thoroughly tested prior to shipment and sent to participating laboratories on dry ice taking from one to four days. When requested, additional reagents such as SNaPshot kit (Applied Biosystems, Foster City, CA, USA) or enzymes necessary for PCR product purification, SBE reaction and purification were sent together on dry ice, and the Imprint™ DNA Modification kit (Sigma-Aldrich, St. Louis, MO, USA) for bisulfite conversion was sent at room temperature.

DNA extraction, DNA quantification and bisulfite conversion

In the organizing laboratory, samples were subjected to DNA extraction, quantification and sodium bisulfite treatment. DNA was extracted from each aliquot of blood, saliva, and semen or from each swab of vaginal fluid and menstrual blood using a QIAamp® DNA Mini kit (Qiagen, Hilden, Germany) following the manufacturer's instructions. Extracted DNA was quantified using a Quantifiler® Duo DNA Quantification kit (Applied Biosystems), and 200 ng of genomic DNA was bisulfite converted using the Imprint™ DNA Modification kit and eluted with 20 μL of distilled water. For preparation of specimens to send to collaborative laboratories, an appropriate number of genomic DNA and bisulfite converted DNA batches were pooled and

133

redistributed into 10 µL aliquots, which contained 20 ng/µL of genomic DNA or approximately 10 ng/µL of bisulfite converted DNA. The specified genomic DNA input in bisulfite conversion of part 3 was 100 ng of genomic DNA, but in part 4, the input amount was not suggested. Example protocols were provided to participating laboratories as a reference, but the chemistries and instrumentation to be used were left for each laboratory to decide.

Multiplex methylation SNaPshot

The multiplex methylation SNaPshot reaction described in (166) was modified to include seven CpG markers that show a methylation signal only in the target body fluids. For this study, a blood marker, cg01543184, with cross reactivity to semen was replaced with blood marker cg08792630 (176), and a semen marker cg17621389 with that produced semen-specific non-methylation was removed. The resultant multiplex included two CpGs (SE1 and SE2) for semen, two CpGs (BL1 and BL3) for blood, two CpGs (VF1 and VF2) for vaginal fluid and one CpG (SA1) for saliva; SE1, SE2, BL1, BL3, VF1, VF2, and SA1 represent cg17610929, cg26763284, cg06379435, cg08792630, cg09765089-231d, cg26079753-7d, and cg09652652-2d, respectively. The multiplex PCR was performed in 20 µL reactions containing 1 µL of bisulfite-converted DNA (10 ng), 3 U of AmpliTaq Gold® DNA polymerase, 2 µL of Gold ST*R 10× buffer (Promega, Madison, WI, USA), and PCR primers (Figure 5.5).

	Target ID	Sequence (5'→3')	Conc. (uM)	Amplicon size (bp)
SE1	cg17610929	TTG TTG ATA TGT TTT GAA TTA TTA AG	3.0	174
		ATA ACT TCC CTT ATC AAC ACC AAC	3.0	
SE2	cg26763284	TGA TTT ATA ATT ATT AGG GAG GGA AAT AG	1.0	105
		CCT AAA ACA ACC RAT TCC CAA C	2.0	
BL1	cg06379435	TTT ATT GGG GTA TTT TTA TTG GTT AG	10.0	157
		AAA ATA CAA CTT ACT CCT AAA CAC C	10.0	
BL3	cg08792630	TGT TTT AAG AGG ATG ATA AGG AA	3.0	220
		CCA CCT CAA TCC AAA CTA ACT ACA	3.0	
VF1	cg09765089-231d	TTG GTA GTT TTT GGA TTT TGG AG	3.0	137
		AAA CRT AAA ACR ACC CRA AC	24.0	
VF2	cg26079753-7d	TTT TGT GAG TGT GAG AGA TTT TTA AGA	2.0	176
		AAA ACC TCC AAA ACA AAA CCT CTA	2.0	
SA1	cg09652652-2d	GGG GAT TYG TTT YGT TAG GT	16.0	153
		CCA TTT CCC CCT TCC TAA AA	4.0	

Figure 5.5 – List of PCR primers used in this study.

PCR cycling was conducted using a Veriti™ Thermal Cycler (Applied Biosystems) under the following conditions: 95˚C for 11 min; 34 cycles of 94˚C for 20 s, 56˚C for 60 s, and 72˚C for 30 s; and a final extension at 72˚C for 7 min. Then, 5 μL of PCR products were purified with 1 μL of ExoSAP-IT (USB, Cleveland, OH, USA) by incubation at 37˚C for 45 min followed by heat inactivation at 80˚C for 15 min. Multiplex SBE reaction was performed using 1 μL of purified PCR products, SBE primers (Figure 5.6) and a SNaPshot™ kit according to the manufacturer's instructions.

The final extension products were purified with SAP-Recombinant (USB) enzyme. Large batch of 5× PCR primer and 10× SBE primer stocks were prepared and aliquoted to provide all laboratories with the same primer mixtures.

	Target ID	Sequence (5'→3')	Conc. (uM)	Length (nt)
SE1	cg17610929	(T)₆ CCG AAA CCC TCC CCA C	4.0	21
		(T)₈ CCA AAA CCC TCC CCA C	4.0	22
SE2	cg26763284	(T)₈ CGC GTA ACG ACT ATA AAA CCC TC	0.3	29
		(T)₈ CAC ATA ACA ACT ATA AAA CCC TC	1.0	31
BL1	cg06379435	(T)₁₇ CCR ATA AAA CCT CAA ACR TAA AAC	40.0	41
BL3	cg08792630	(T)₂₁ CCR TAA TAA CTT CTA CCT ATA AAT AAA CCC	6.0	51
VF1	cg09765089-231d	(T)₃₄ TCC CCA AAT AAC AAA CRA CRA AAA TC	9.0	60
VF2	cg26079753-7d	(T)₄₄ CRA TCA ACT ACT ATA AAA ACA CC	9.0	67
SA1	cg09652652-2d	(T)₄₈ CCA CGA ATA AAT AAC CAC GAT AAA AC	15.0	74

Figure 5.6 – List of SBE primers used in this study

The specified bisulfite converted DNA input for the multiplex PCR of parts 2 and 3 was 10 ng; in part 4, the input amount was not suggested. Example protocols were provided to participating laboratories, but a choice was given regarding which PCR buffers and amplicon purification methods were to be used.

Capillary electrophoresis and analysis of DNA methylation profiles

In the organizing laboratory, the extension products were analyzed using the ABI PRISM 3130*xl* Genetic Analyzer and GeneScan software 3.1 (Applied Biosystems). In participating laboratories, samples were injected and run according to conditions of choice on various types of standard genetic analyzers (Tables 5.7 and 5.8).

For the analysis of DNA methylation profiles, participating laboratories were requested to report peak heights observed from the electropherograms and to calculate percentage methylation values (0–100%) at each CpG site by dividing nucleotide G intensity (detection of unconverted methylated cytosine on the reverse strand) by nucleotide G plus nucleotide A intensity (detection of converted unmethylated cytosine on the reverse

strand). Then, body fluid typing results were reported in reference to the electrophoretic

results of SBE products in part 1.

Table 5.7 - Summary of the kits, chemicals and instrumentation used in 7 participating laboratories

DNA extraction		Lab no.
DNA extraction kit	QIAamp DNA Mini kit	1, 2, 5
	QIAamp DNA Investigator kit	3
	Qiagen Micro kit	4
	Qiagen EZ1 DNA Investigator kit	6, 7
Quantification method	Quantifiler$^®$ Duo DNA Quantification kit	1, 2, 5
	Nanodrop	3
	Qiagen Human Duo kit	4
	Qubit 2.0	6
	Picogreen	7
Bisulfite conversion		
Conversion kit	Imprint$^{™}$ DNA Modification kit	1, 2, 3, 4, 7
	Epitect Fast Bisulfite Conversion kit	5, 6
Multiplex PCR		
Thermocycler	ABI VERITI	1, 5
	GeneAmp$^®$ PCR System 9700	2, 3, 4, 7
	ABI 2720	6
PCR buffer	Gold ST*R 10x buffer	1, 2, 3, 4, 7
	AmpliTaq Gold Buffer	5, 6
Taq polymerase	3U of AmpliTaq Gold DNA Polymerase	1, 2, 3, 4, 5, 6, 7
Total PCR volume	20 ul	1, 2, 3, 4, 5, 6, 7
Post-PCR purification	2 ul/rxn of USB ExoSAP-IT	1, 3, 4, 5, 6, 7
	2-4 ul/rxn of USB ExoSAP-IT	2
SBE reaction		
Thermocycler	ABI VERITI	1, 5
	GeneAmp$^®$ PCR System 9700	2, 3, 4, 7
	ABI 2720	6
SBE kit	SNaPshot™ kit	1, 2, 3, 4, 5, 6, 7
Post-extension purification	1 ul/rxn of USB SAP-Recombinant	1, 2, 3, 5, 7
	2 ul/rxn of USB SAP-Recombinant	4
	1 ul/rxn of USB Fast AP	6

Table 5.8 - Summary of settings and instruments used to analyze the SBE products in the 7 participating laboratories

GeneScan settings		Lab. No
Instrument platform	ABI 3130xl	1, 4
	ABI 3130	2, 3, 6
	ABI PRISM 310	5, 7
GeneScan software	Genemapper ID v3.2	1, 4, 5, 7
	GeneMapper ID-X	2, 6
	Genemapper ID v3.2.1	3
Size standard	GeneScan 120 LIZ	1, 2, 3, 4, 5, 6, 7
SBE product	1 ul SBE product in 10 ul Hi-Di formamide	1, 2, 3, 4, 6, 7
	1.5 ul SBE product in 9 ul Hi-Di formamide	5
Injection time	3 sec	1, 2, 7
	5 sec	5
	10 sec	6
	15 sec	3
	22 sec	4
Collection time	480 sec	1
	1000 sec	2, 4, 6
	1200 sec	5
	1300 sec	3
	1440 sec	7

B.4 Results and Discussion

<u>Participating laboratories</u>

Seven laboratories including the organizing laboratory participated in this exercise. The participating laboratories varied in experience with regard to DNA methylation analyses. Two laboratories indicated they were beginners at DNA methylation testing; two other laboratories, advanced beginners; one laboratory, an experienced researcher; and two other laboratories, experts. The participating laboratories were also asked to complete a questionnaire describing the kits, chemicals, quantities and instrumentation they used (Tables 5.7 and 5.8). The seven laboratories used four different

kits for DNA extraction, five different quantification methods, two different kits for bisulfite conversion, two different PCR buffers for multiplex PCR, three different thermocyclers, three different genetic analyzers and three different analytical software and five different analytical settings. The diversity of instrumentation and respective analysis tools emphasizes the importance of such collaborative exercises in order to evaluate the robustness of implementation of methods in different laboratories.

Exercise tasks: parts 1 to 4

The collaborative exercise started with part 1, where purified SBE products are analyzed with capillary electrophoresis. Analysis of the final multiplex methylation SNaPshot products from various body fluid samples provided references to the following data interpretation and enabled adjustment of the genetic analyzer setting if necessary. Six of the seven laboratories reported very similar results in part 1 (Figure 5.7).

One laboratory (laboratory 4) had, on average, 4-fold higher peaks than the organizing laboratory with low non-specific peaks that seemed to be due to the use of higher injection settings. Because of failure in adjusting the genetic analyzer setting and unsuccessful enzyme treatment to purify PCR products in the subsequent parts 2–4 at laboratory 4, the results discussed in this report will only include data from the remaining six laboratories.

The methylation percentages at each CpG site were similar among the laboratories, but two laboratories using an ABI PRISM 310 genetic analyzer showed somewhat lower methylation percentages. In addition, it seems that because two types of menstrual blood of which one had the same profile as vaginal fluid (menstrual blood type 1) were

provided as references in part 1, participating laboratories had difficulty in inferring the sample origin with the profile of vaginal fluid. In this case, some laboratories reported the origin of the sample to be either vaginal fluid or menstrual blood, while others reported it to be vaginal fluid or menstrual blood. As such, a distinction between vaginal fluid and menstrual blood could not be provided for the reports regarding parts 2–4.
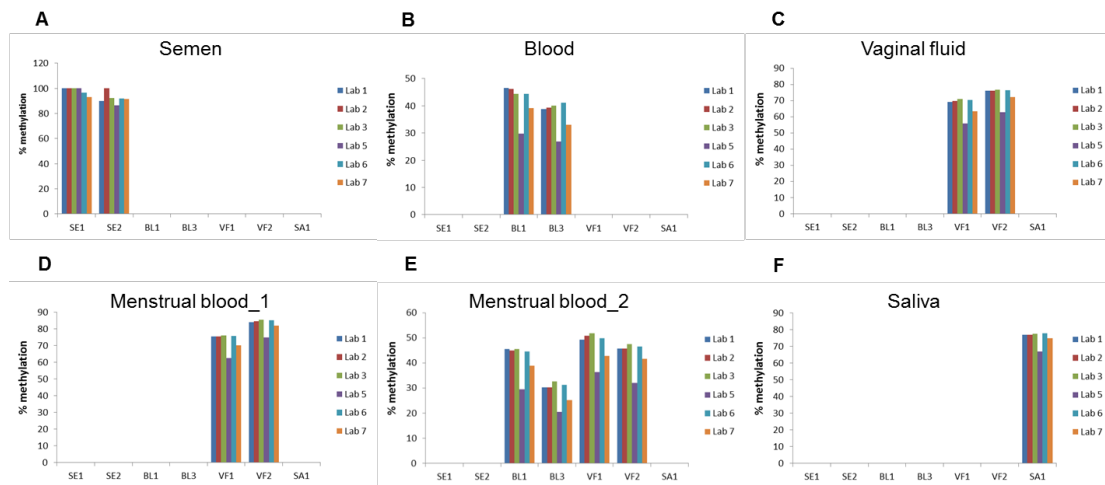
Figure 5.7 - DNA methylation profiling results of part 1 conducted with the final SBE products of the multiplex methylation SNaPshot reaction. DNA methylation was recorded as a percentage value by dividing nucleotide G intensity by nucleotide G plus nucleotide A.

In part 2, bisulfite converted genomic DNA of semen, blood, vaginal fluid, menstrual blood and saliva with known body fluid origin, and three other samples with unspecified origin of which two were single source and one was a mixture, allowed us to test the proficiency of the participating laboratories in performing the SNaPshot reaction. Four of the six laboratories reported methylation signals only at target body fluid-specific CpG

markers with correct body fluid typing results (Figure 5.8).
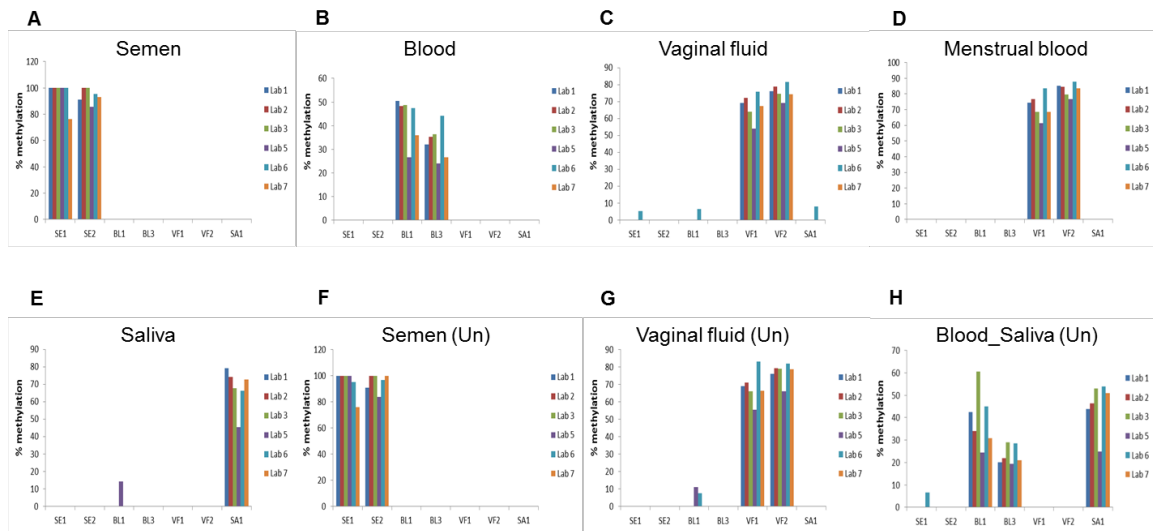
Figure 5.8. DNA methylation profiling results of part 2 conducted with bisulfite converted DNA. Part 2 results show the proficiency of each laboratory in SBE reaction. (Un) indicates samples provided with unspecified body fluid origin.

Two laboratories reported additional low level methylation signals at SE1, BL1 and SA1 markers, but these results did not affect correct interpretation of the body fluid typing results. These signals had low methylation percentages, mostly less than 15%. For example, the two lowest methylation signals from laboratory 6 were 100 rfu and 94 rfu corresponding to 6.7% and 8.1% methylation percentages at BL1 and SA1 CpG sites, respectively. Therefore, all six laboratories properly performed and implemented the multiplex SNaPshot reactions with 10 ng of bisulfite converted DNA.

For part 3, genomic DNA from semen, blood, vaginal fluid, menstrual blood and saliva and three genomic DNA samples of unspecified origin were provided to test each laboratory's proficiency in bisulfite conversion. The participating laboratories were asked to treat 100 ng of genomic DNA with a bisulfite conversion kit of choice and to perform multiplex methylation SNaPshot reactions with a 1/10 portion of the eluted bisulfite

converted DNA. Two kits were used in bisulfite conversion (Table 5.7), but no significant difference was observed between the results. As in part 2, the same four laboratories reported methylation signals only at target body fluid-specific CpG markers with correct body fluid typing results (Figure 5.9).
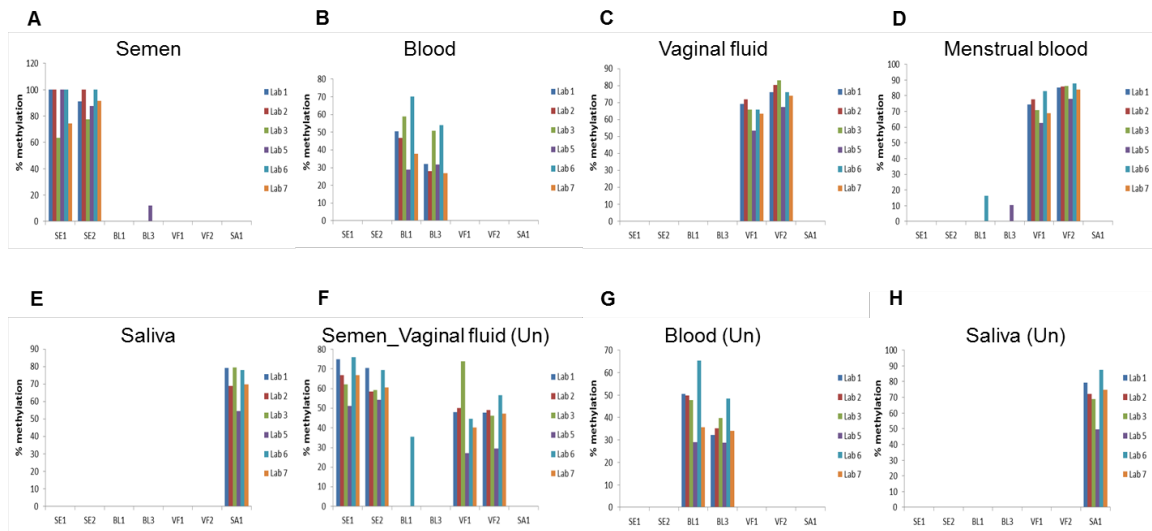
Figure 5.9. DNA methylation profiling results of part 3 conducted with genomic DNA. Part 3 results show the proficiency of each laboratory in bisulfite conversion. (Un) indicates samples provided with unspecified body fluid origin.

Two laboratories (laboratories 5 and 6) reported sporadic methylation signals at the BL1 or BL3 markers in semen, menstrual blood and in the mixture of semen and vaginal fluid. These peaks were generally low in methylation percentage, i.e. less than 15%, and did not affect the correct interpretation of body fluid typing results. However, the methylation percentages at each target CpG site showed higher variation among laboratories in part 3 when comparing to parts 1 and 2. Laboratory 3 reported no signal at the BL3 marker that was most likely a result of PCR failure at this site; however, it did not affect the identification of body fluid type. In addition, laboratory 6 reported a high methylation percentage (35.6%) at the BL1 site in the mixture of semen and vaginal fluid,

leading to a body fluid typing result of a mixture of semen and menstrual blood. In most cases, laboratories concurred on the expected interpretation of the results reported among them with genomic DNA and bisulfite converted DNA. This result suggests that DNA methylation-based body fluid identification using multiplex methylation SNaPshot and bisulfite conversion of genomic DNA could be successfully incorporated into forensic laboratory workflow.

In part 4, various body fluid samples with and without specified origin were provided. Specimens were subjected to DNA extraction, quantification, bisulfite conversion and multiplex methylation SNaPshot reaction using various methodologies according to each laboratory's preference. Three DNA extraction kits and four quantification methods were used by the six different laboratories. Sufficient amounts of DNA were obtained from most of the laboratories, but relatively small yields of DNA were obtained in one of the laboratories using an automatic DNA extraction system (laboratory 6). Various amounts of DNA were subjected to bisulfite conversion and 1/10 or 1/20 of the volume of eluted bisulfite converted DNA was amplified in a subsequent multiplex methylation SNaPshot reaction. The resultant amplified bisulfite converted DNA varied in amount from 0.8 ng to 20 ng. Of the six laboratories being evaluated, five laboratories produced acceptable results with the correct body fluid identification. The sixth laboratory (laboratory 5) failed to obtain electrophoretic results from a few specimens but produced correct results from the others. The five laboratories reported methylation signals only at target body fluid-specific CpG markers except for a menstrual

blood swab with unspecified origin (Figure 5.10).
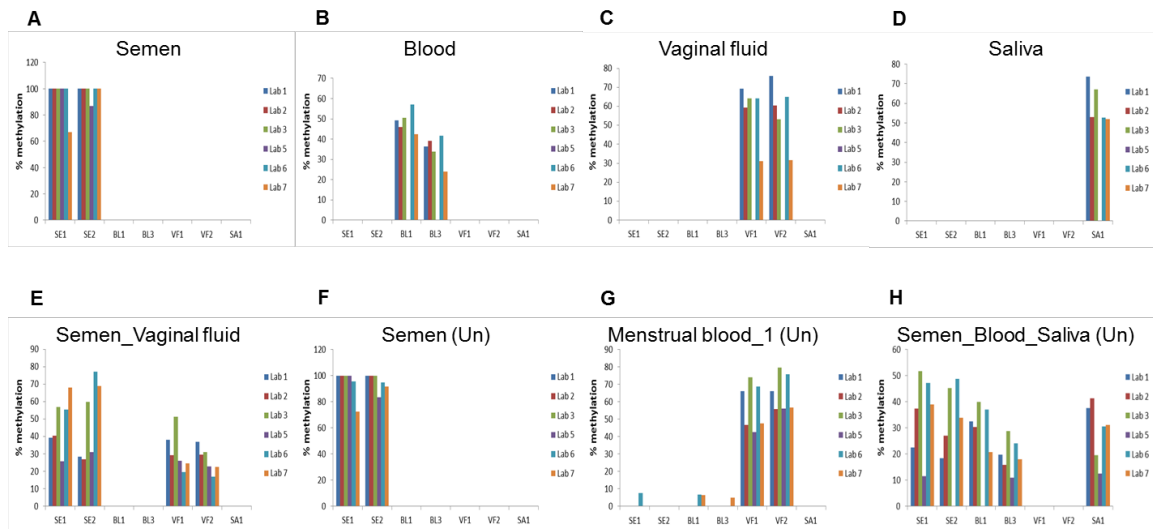
Figure 5.10. DNA methylation profiling results of part 4 with body fluid samples. Part 4 results show the proficiency of each laboratory in the whole procedure of DNA methylation-based body fluid typing raging from DNA extraction to capillary electrophoresis. Methylation percentages in a menstrual blood swab and a mixture swab of semen, blood and saliva with unspecified origin were presented in consideration of the possibility of sample switch in laboratory 5. (Un) indicates samples provided with unspecified body fluid origin.

For this sample, laboratory 6 produced unexpected methylation signals at SE1 and BL1 CpG markers, and laboratory 7 showed methylation signals at BL1 and BL3 CpG markers. However, these methylation signals were all less than 10%, and had no adverse effect on data interpretation since both laboratories reported the specimens to be either vaginal fluid or menstrual blood. Laboratory 5 did not get body fluid typing results from three samples with indicated origin. Since certain amounts of DNA were extracted and quantified from these three samples, failure seems to have occurred in the recovery of bisulfite converted DNA or in PCR amplification. This laboratory also reported "inconclusive" for a mixture swab with a methylation signal only at SE1, BL3 and SA1. However, additional low methylation signals at SE2 and BL1 (~50 rfu) (data not shown)

suggested the presence of semen, blood and saliva in this sample. In this exercise, the most experienced laboratories produced better DNA methylation profiles and interpretation results. However, even the laboratories that varied from less experienced to beginners in DNA methylation analysis produced successful body fluid typing results, providing strong evidence for the potential application of DNA methylation profiling for use in forensic body fluid typing.

B.5 Conclusion

This collaborative exercise aimed to examine the possibility of incorporating DNA methylation-based body fluid identification methods into forensic workflow. Most of the participating laboratories reported consistent results for mixtures as well as for single source samples with the use of their own laboratory equipment and different kits and chemicals of choice, which demonstrate the usefulness of this application of DNA methylation profiling in forensic body fluid typing. In this study, we found that certain laboratories reported low but detectable methylation signals at unexpected CpG sites, and there were variations in DNA methylation percentages among laboratories. Because these unexpected methylation signals were frequently at less than 10% methylation and sometimes lower than 100 rfu, we recommend that in the future an appropriate detection threshold in terms of rfu and percent of methylation, e.g., more than 100 rfu and 10% methylation, should be applied. As for the methylation variation, this is not an unexpected result given that the participating laboratories used a variety of methods and instrumentation. It is important to note that the seven CpGs used in this multiplex provide on-off signal variations in results that can alleviate the effect of methylation variation

among laboratories. Because it is not possible to differentiate vaginal fluid and menstrual blood with the present multiplex methylation SNaPshot system, the addition of new menstrual blood-specific CpG markers to the multiplex system is needed for better identification of body fluids. In addition, with the use of two or more CpG markers for each target body fluid being preferred, it is evident that there is also a need to identify more CpGs that are specific to saliva. If possible, the multiplex system should be supplemented in the future with more CpG markers to have two or more CpGs for each body fluid and consequently increase the accuracy of the identification of body fluids. While routine use of DNA methylation-based body fluid typing and identification assays may yet require further tests for sensitivity and specificity, this study demonstrates that methylation profiling can be successfully implemented in forensic laboratories once clear guidelines are set for data interpretation.

CHAPTER VI – HIGH RESOLUTION MELT ANALYSIS OF DNA METHYLATION TO DISCRIMINATE SEMEN IN BIOLOGICAL STAINS

A. Abstract

The goal of this paper was to develop a method for the detection of semen in biological stains using high resolution melt analysis and DNA methylation. To perform this task we have utilized an epigenetic locus, which targets a tissue-specific differentially methylated region for semen (tDMR). This specific locus, ZC3H12D contains methylated CpG sites that are hypo methylated in semen and hyper methylated in blood and saliva. Using this procedure, DNA from forensic stains can be isolated, processed using bisulfite modified PCR and detected by real time PCR with HRM capability. The method described in this report is robust as we were able to obtain results from samples with as little as 1 ng of genomic DNA. Samples inhibited by humic acid still produced reliable results. Furthermore, the procedure is specific and will not amplify non-bisulfite modified DNA. Because this process can be performed using real time PCR and is quantitative, it fits nicely within the workflow of current forensic DNA laboratories. As a result it should prove to be a useful technique for processing trace evidence samples for serological analysis.

B. Introduction

In cases where the presence of a suspect is expected, determining the type of body fluid present can make the distinction between innocent and criminal contact and therefore be a valuable addition to DNA genotyping. The tissue source of a DNA sample (156, 191)can be determined by analyzing DNA methylation patterns. Several reports demonstrate the existence of tDMRs (37, 58, 192) and how they can be used as a powerful tool for body fluid identification (136, 147, 173). Common methodologies for

DNA methylation analysis include DNA digestion by methylation-sensitive

endonucleases followed by PCR amplification (147) and bisulfite conversion of genomic

DNA that causes the unmethylated cytosines to be chemically converted to uracil while

the methylated cytosines are protected. After bisulfite modification, the DNA is amplified

by PCR using specific primers (193) and the amplicons can be analyzed by

pyrosequencing, which provides quantitative methylation values for each CpG site

present in the target sequence (152).

High-Resolution melt analysis (HRM) can detect DNA sequence variants based on their

different melting temperatures (158, 162, 172). PCR products can be differentiated even

if they only diverge in a single nucleotide (159) making this an optimal method to

analyze single nucleotide polymorphisms (162, 194, 195). HRM involves the

amplification of a DNA template by real-time PCR in the presence of a double strand

DNA (dsDNA) intercalating dye such as Eva Green™. The fluorescence is maximized at

the end of amplification when the largest quantity of dsDNA is present.

HRM can also be used to explore the melting differences between unmethylated and

methylated DNA after bisulfite conversion. During bisulfite conversion the unmethylated

cytosines but not the methylated ones are converted to uracils and subsequently thymine

during PCR. The amplicon resulting from amplifying unmethylated DNA has a lower GC

content when compared with the amplicon resulting from methylated DNA, thus having

lower melting temperature. If the bisulfite conversion fails, the GC content of the

amplicon is similar to methylated DNA and an overestimation of methylation occurs

(196, 197). Even though the available commercial kits for bisulfite conversion modify

99% of the DNA, appropriate controls should be in place to confirm that there is no

amplification of unmodified DNA. One way of tackling this is to design primers that anneal only with bisulfite converted DNA.

Several reports (54, 158, 197-199) have explored differences in melting temperatures making use of HRM analysis of differentially methylated DNA with success. Even though DNA methylation analysis by HRM cannot quantify individual CpGs, it provides a robust and inexpensive method to differentiate DNA with different degrees of methylation.

One challenge faced in PCR based methods is the decrease of amplification efficiency due to inhibition. DNA is often co-extracted with substances that can hinder PCR. There are two main mechanisms of PCR inhibition. One mechanism occurs when the inhibitor binds the DNA and the other mechanism is when the inhibitor hinders the catalytic activity of Taq polymerase. When the inhibitor binds to the DNA it can diminish the processivity of the DNA polymerase or prevent primers to anneal to the template DNA, thus decreasing PCR efficiency. Moreover for inhibitors binding to the DNA, a simple cleanup step may not be sufficient to remove the decrease in PCR efficiency. Humic acid is one substance known to inhibit PCR because it binds to the template DNA (86, 88).

In this work we investigated whether the methylation differences between blood, saliva and semen for the locus ZC3H12D can be identified using HRM. When pyrosequencing analysis is performed in the locus ZC3H12D, blood and saliva show hypermethylation when compared to semen (156). The successful analysis of this locus using HRM would be extremely useful in the identification of the source of DNA samples in forensic investigations.

C. Methods

DNA collection, extraction and bisulfite conversion

Ten Blood, 9 buccal swabs and 7 semen samples were collected from volunteers according to the approved IRB-13-0555 from Florida International University. Swabs were air-dried and DNA extraction was performed using the EZ1® DNA Investigator kit (Qiagen, CA) and the BioRobot® EZ1 automated purification workstation (Qiagen, CA) according to the manufacturer's specifications. Quantification was performed using the PicoGreen® method (Life Technologies, CA). One and fifty nanograms of DNA were bisulfite modified using the EpiTect® Fast DNA Bisulfite Kit (Qiagen, CA) according to manufacturer's instructions. For the sensitivity studies, samples were serial diluted to obtain 0.5 and 0.25 ng of input DNA to bisulfite conversion. In parallel the DNA samples that were modified using 1 ng input to bisulfite modification were serially diluted after modification to determine the lowest amount amplifiable by HRM. Samples that were not bisulfite modified were diluted to 0.5 ng/μL in order to obtain a 1 ng input to the HRM reaction.

DNA amplification

Amplification reactions were performed using the Epitect® HRM kit (Qiagen, CA) on a Rotor Gene 6000 real time instrument (Qiagen, CA). The kit is composed of an HRM buffer that contains Eva Green®, HotStarTaq® Plus and a dNTP mix all with unknown concentrations. Two microliters of DNA were added to the master mix composed of HRM buffer and 0.75 μM of each forward and reverse primers. Primers were designed using the online tool MethPrimer (200). The primers sequence is provided

on Table 6.1. Amplification was made by initially holding the temperature for 95 ºC for 5 minutes followed by 45-50 cycles of 95 ºC for 10 seconds, 55 ºC for 30 seconds and 72 ºC for 10 seconds. Melt analysis was performed by increasing the temperature from 65 to 85 ºC in 0.3 °C increments and detecting fluorescence in the HRM channel. Melt curve analysis was made using the Rotor-Gene 6000 Series Software version 1.7.

Table 6.1 – Sequence of the primers used in this study.

Primers	Sequence
ZC3H12D Forward	GGG TGA GGG TTT AAG GGT
ZC3H12D Reverse	CTC CCC TCA AAA CCT CAT

Humic Acid inhibition

Humic acid was added to each DNA sample at a final concentration of 1.92×10^{-3} ng/µL prior to bisulfite conversion. Both samples without (control samples) and with humic acid were bisulfite converted and amplified as described above. A master mix containing humic acid with a final concentration of 1.92×10^{-3} ng/µL was made and samples were amplified using this master mix as described above.

Statistical analysis

One-way ANOVA with a Tukey post-hoc was performed using SPSS® (Version 22, IBM) to compare the melting temperatures between the 7 semen samples, 10 blood samples and 9 saliva samples and determine if the difference observed is statistically significant. For p-values inferior to 0.05 the melting temperature was considered to be significantly different.

D. Results and Discussion

High-resolution melt can be used to discriminate between fragments of DNA that present varying degrees of methylation. HRM has the advantage of being cost effective, besides being a useful alternative/additional method to pyrosequencing protocols, since it permits the fast analysis of amplicons with sizes larger than 70 base-pairs. From the work of Madi *et al.* 2012, we note that for locus ZC3H12D, semen presents low percent of methylation (around 10%) whereas other body fluids have levels of methylation around 100%. The hypo-methylation of semen is expected to result in a melt curve with a lower melting temperature when compared to blood and saliva cells. Figures 6.1 and 6.2 illustrate that DNA from semen presents a melting temperature of 75.5 ºC with a standard deviation of 0.2 ºC, which is lower than for other body fluids (78.2 ºC and 78.1 ºC for blood and saliva, respectively). Our primers amplify a specific genomic region that is hypermethylated in saliva and blood resulting in amplicons with high GC content. Methylated cytosines are protected from bisulfite conversion and remain as cytosines. Amplification will result in amplicons with a high GC content, which results in higher T_M as higher temperatures have to be reached to break the triple hydrogen bonds that bind guanine to cytosine. In semen, the same region amplified by our primers is hypomethylated. The lack of methyl groups in the cytosines from semen DNA causes them to convert to uracil through the bisulfite reaction. Those uracils are amplified as thymines and result in an amplicon with low GC content, resulting in curves with lower T_M. The differences in T_M are visible in Figures 6.1 and 6.2. To determine the sensitivity of this method we performed serial dilutions of extracted DNA from blood, saliva and

153

semen to obtain inputs of DNA to bisulfite conversion of 1, 0.5 and 0.25 ng. Those

samples were amplified and the T_M was determined by HRM analysis.
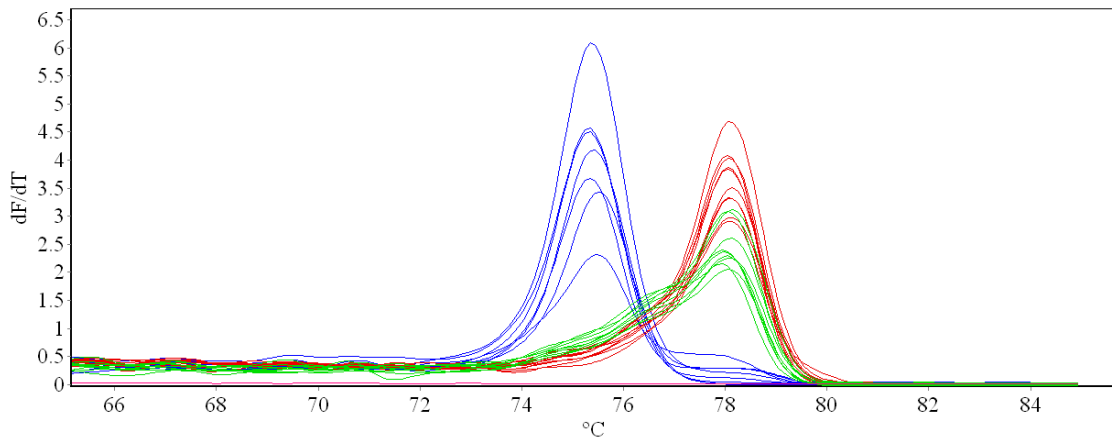
Figure 6.1 – Negative of the first derivative calculated from the change in slope resulting from a decrease in fluorescence with temperature (-dF/dT) for high-resolution melt curves, showing 7 semen samples (blue), 10 blood samples (red) and 9 saliva samples (green). A no template control (NTC, no DNA) was also included and is seen at the baseline in pink color. Body fluid samples were donated from individual volunteers. Samples were amplified from 50 ng of genomic DNA added to bisulfite conversion.

Figure 6.3 shows that for inputs of DNA lower than 1 ng, some samples fail to amplify,

even when the amplification is extended for 50 cycles. One nanogram of genomic DNA

seems to be the minimum amount to input to bisulfite conversion to guarantee

amplification using this protocol for HRM analysis.
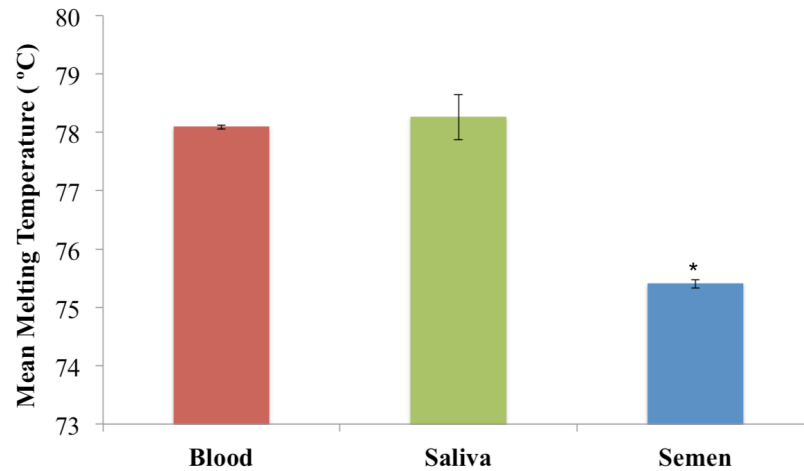
Figure 6.2 – Chart showing the mean values for melting temperatures (ºC) determined by HRM for 10 blood samples, 9 saliva samples and 7 semen samples. Analysis of variance was performed showing that the melting temperature is significantly different for semen when compared to blood and saliva (* signifies $p < 0.05$).

To further explore whether amplification for low content samples fails due to low template added to PCR or as a higher loss on the bisulfite conversion, we performed serial dilutions from the samples that were bisulfite-modified with 1 ng input to bisulfite, posterior to the conversion. Four serial dilutions of 1:2 were performed from the bisulfite-modified samples with 1 ng of input to bisulfite conversion. Figure 6.4 shows that a dilution of 1:2 is enough to impair amplification for 2 out of a total of 9 samples.
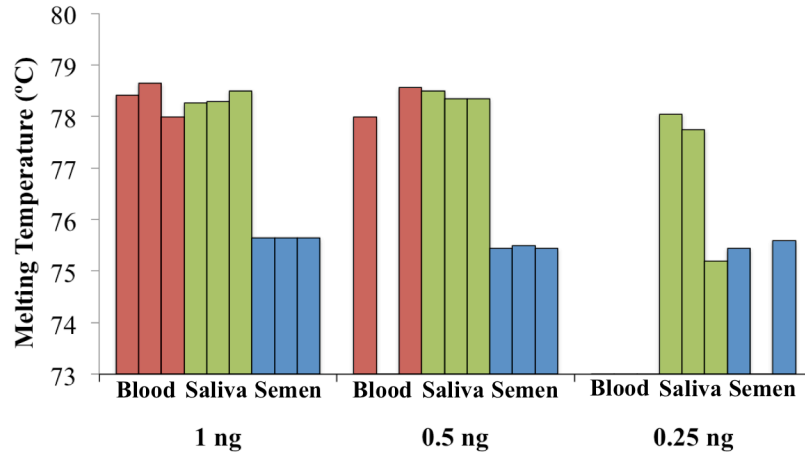
Figure 6.3 – Chart showing the melting temperatures (ºC) for 3 individual samples from 3 different volunteers of each body fluid, Blood (red), Saliva (green) and Semen (blue). The x-axis shows the different amounts of input DNA added to bisulfite modification. Each body fluid should have 3 individual bars for individual amplifications. The absence of a bar indicates that amplification failed.

As the bisulfite-modified samples get further diluted, fewer samples amplify properly as it can be seen for the 1:16 dilution only having 5 successful amplifications out of a total of 9 samples. Figure 6.4 also shows that for the dilutions of 1:2 and 1:4, when amplification is successful the expected T_M is observed for the body fluid tested. The saliva samples at 1:8 dilution show a higher standard deviation (77.50 ± 0.72) when compared to the 1:2 (78.50 ± 0.15) and 1:4 (78.25 ± 0.00) possibly due to stochastic effects on amplification and melt analysis when low amounts of DNA are used.

The results displayed in Figures 6.3 and 6.4 show that below 1 ng of input to bisulfite conversion, amplification may be impaired due to low amounts of DNA added to PCR. However, in the cases where amplification occurs below 1 ng, distinction between semen and blood and saliva may still be possible using the melting analysis. The fact that the melting temperatures do not change when low amounts of DNA are successfully

amplified shows that this method is unlikely to give false results that would otherwise
lead to a misclassification of a body fluid. Additional samples including replicates for
each dilution can further establish the sensitivity for this method.
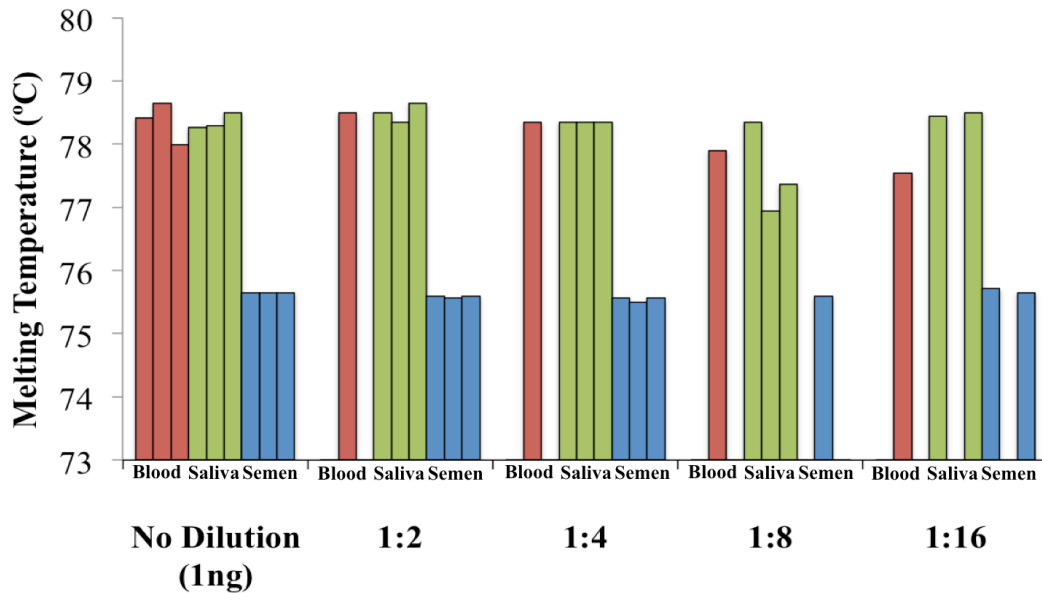
Figure 6.4 –Chart showing the melting temperatures (ºC) for 3 individual samples
from 3 different volunteers of each body fluid, Blood (red), Saliva (green) and
Semen (blue). The x-axis shows the dilution factor from samples that were bisulfite-
modified with 1 ng of DNA. For comparison the non-diluted samples (1ng) are
displayed. Each body fluid should have 3 individual bars for individual
amplifications. The absence of a bar indicates that amplification failed.

To further test the robustness of the method we explored whether the designed

primers can amplify non-bisulfite modified DNA. Differences in methylation can only

accurately be explored if the DNA is completely bisulfite modified prior to amplification

and HRM. Incomplete bisulfite modification can result in overestimation of methylation

in DNA due to lack of conversion of unmethylated cytosines to uracils. As a result, a

PCR product resulting from an unmodified template will keep all cytosines regardless of

their methylation status and will have a T_M similar to a PCR product that results from a

modified, fully methylated template (196). To determine if our primers are specific for

bisulfite-converted DNA we performed HRM for the same extracted samples of semen,

blood and saliva before (genomic DNA, gDNA) and after bisulfite conversion (bisulfite-

modified DNA, bDNA).
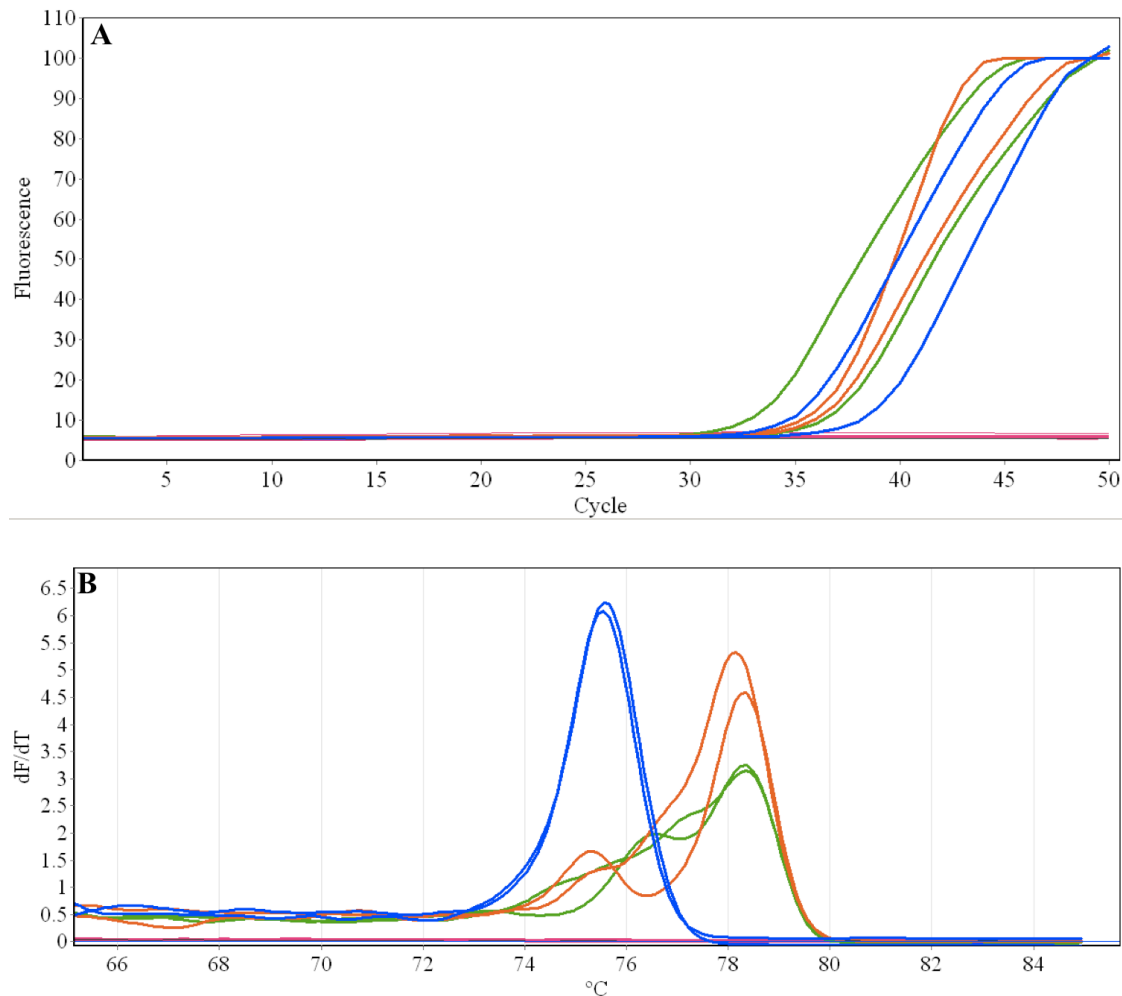
Figure 6.5 – Graphs showing samples before and after bisulfite modification. bDNA from semen samples are blue, from blood samples are in orange and from saliva samples are in green. gDNA samples are shown in pink. **(A)** shows the increase in fluorescence as a result of incorporation of the green fluorescent dye Eva Green® as the cycle number increases. DNA samples that were not bisulfite modified (gDNA) do not amplify and are in pink color at the baseline. The bDNA samples amplify showing an increase in fluorescence. **(B)** Melt curve analysis shows that the bDNA samples amplified show a melt profile consistent with Figures 6.1 and 6.2. Semen samples are blue, blood samples are in orange and saliva samples are in green.

Figure 6.5A demonstrates that gDNA samples did not amplify (pink lines) even after 50 cycles. The bDNA amplified properly (blue, orange and green lines on Figure 6.5A) and presented the expected melt curves, with a lower T_M for semen when compared to blood and saliva (Figure 6.5B).

In some circumstances samples collected from crime scenes contain substances that when co-extracted with DNA can cause inhibition of the amplification. To determine if the cleanup steps in the bisulfite conversion reaction is sufficient to remove such inhibitors, we incorporated humic acid before and after bisulfite modification. Humic acid is an inhibitor known to decrease amplification efficiency by binding to the template DNA (86). PCR efficiency is evaluated by the number of copies of DNA synthesized per cycle (or unit of time). On a graph displaying the amplification curve, such as Figure 6.5A a steeper slope on the exponential curve will represent a more efficient amplification. Efficiency is 100% when E=2 ($E=10^{slope}$) and it means that the PCR product doubles at each cycle.

Humic acid was added before bisulfite conversion to simulate co-extraction with DNA. Figure 6.5A shows that samples without humic acid (control, thin lines) and samples where the humic acid was added before bisulfite conversion (dashed lines) amplify with similar efficiency. The fact that the steepness of the curves for the control samples and those where humic acid was added before bisulfite treatment are similar allows us to conclude that the cleanup steps performed as part of the bisulfite kit seem sufficient to remove humic acid co-extracted with DNA at the concentration tested.

To determine whether the same concentration of humic acid causes changes in PCR efficiency we performed another reaction where the humic acid was added to the

PCR mix. All the samples amplified in the presence of humic acid (thick lines) show a decrease in PCR efficiency which is seen by a decrease in slope steepness on Figure 6.6A.
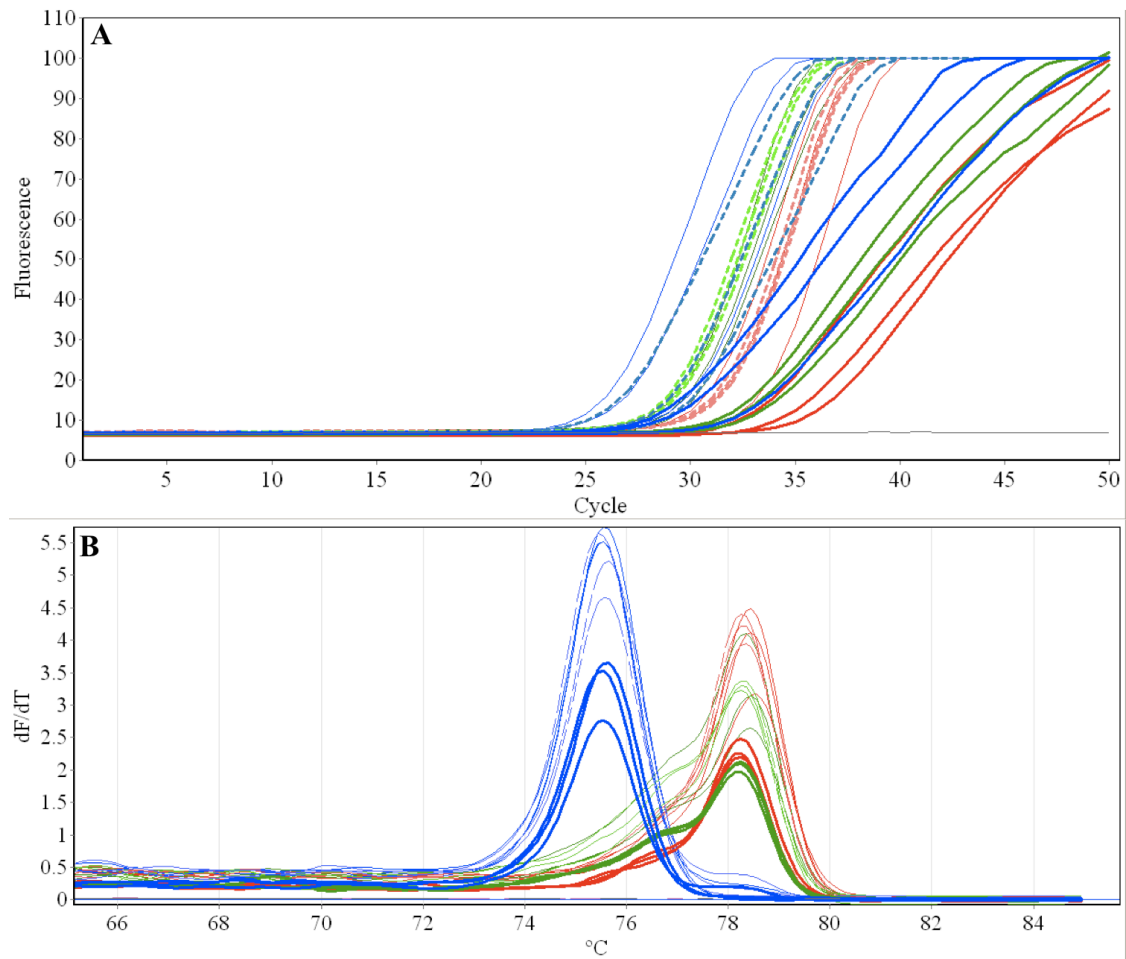
Figure 6.6 – Graphs showing amplification curves for samples without humic acid (control, thin lines), samples where humic acid was added prior to bisulfite conversion (dashed lines) and samples where humic acid was added to the PCR master mix (thick lines). DNA from blood is in red, saliva cells in green and semen in blue. (**A**) The presence of the inhibitor during amplification (thick lines) causes a decrease in the steepness of the slope of the curve when compared to samples without humic acid (thin lines) or when humic acid was added prior to bisulfite conversion (dashed lines). (**B**) The presence of humic acid does not affect melt curve since all samples have a similar T_M even when amplified in the presence of humic acid.

Even though the presence of inhibitor in the amplification seemed to decrease the PCR efficiency as seen by a less steep slope in Figure 6.6A, it did not cause a shift on the melting curve (Figure 6.6B). The T_M of DNA extracted from semen samples (blue in Figure 6.6B) still shows a lower T_M when compared to blood (red in Figure 6.6B) and saliva (green in Figure 6.6B), regardless of the presence of humic acid (thick lines). The results observed are likely due to the fact that even though amplification efficiency decreased, the number of cycles was sufficient to reach completion of PCR. Since melt analysis is done to the PCR products, as long as there is a sufficient number of amplicons at the end of amplification, the melt results will be similar. Previous studies by Thompson and colleagues (86) speculated that changes in the melt curve due to the presence of inhibitors that can bind to the PCR product can depend on the size and GC content of the amplicon. Our amplicon has a size of 91 base-pairs and a GC content of 21% for complete methylated (blood and saliva cells) and 14% for unmethylated (semen) DNA strands. Pionzio and colleagues (88) concluded that long amplicons with low GC content are more prone to display effects of inhibition. Our amplicon is considered to be short and with low GC content but shows robustness in melt curve analysis even when the concentration of inhibitor is sufficient to reduce PCR efficiency.

E. Conclusion

The HRM analysis here described allows the amplification and identification of DNA from semen even when only 1 ng of input DNA is bisulfite-converted. When tested with DNA that is not bisulfite converted, our primers failed to amplify the DNA, which proves that incomplete conversions will not result in false results which would possibly

lead to mis-identification of a body fluid. The results presented here also show that when inhibitors like humic acid are co-extracted with the DNA, the cleanup step performed as part of the bisulfite conversion process is capable of cleaning the DNA from such inhibitors, since no decrease in the amplification efficiency was observed. Furthermore when humic acid was added to the amplification step, the HRM method provided reliable results since the T_M values observed are similar to those obtained in the absence of inhibitor. HRM analysis is a promising technology to identify the tissue source of a given DNA sample in forensic laboratories.

CHAPTER VII – IDENTIFICATION OF NEW CpGs ABLE TO DISCRIMINATE

BODY FLUIDS FROM CRIME SCENES

A. Introduction

DNA typing methodologies have been developed a short time ago but nonetheless have evolved in sensitivity and accuracy. Since January 1, 2017 the CODIS (Combined DNA Index System), which was started in 1990 with a specific team at the FBI, has expanded the number of STRs from 13 to 20 to increase the power of discrimination on DNA typing (https://www.fbi.gov/services/laboratory/ biometric-analysis/codis, Sept 29).

Similarly, in the analysis of DNA methylation patterns for body fluid discrimination, having a greater number of markers per body fluid will also increase the ability to discriminate between body fluids, especially when a sample contains a mixture of tissues. Currently, the majority of markers for a body fluid present intermediate levels of methylation. For example, the marker BCAS4, which is specific for saliva, includes 4 CpGs. CpG1, which is the one presenting higher differences in methylation between saliva and other body fluids, shows 64.0 ± 7.1 % methylation in saliva and 3.9 ± 1.6 % (semen) to 39 ± 21% (nasal epithelia), with the percent for other body fluids in between those of semen and nasal epithelia (Table 5.3 and (201)). When the same marker is tested with a mixture of body fluids, the percent of DNA methylation is similar for the sample composed of 75%saliva+12.5% blood+12.5% semen and the sample composed of 75% saliva+25% blood (201) which indicates that semen could go undetected and be confused with blood.

Similarly, other research groups (167, 176) have found the need to identify new loci able to discriminate body fluids using DNA methylation patterns. The simplified way to identify potential new CpGs able to discriminate body fluids is the use of a DNA methylation bead array that interrogates 450,000 individual CpGs in the human genome. The Infinium HumanMethylation450 (Illumina, CA) quantifies DNA methylation based on two types of assays and on bisulfite-modified template. The Infinium I assay uses two probes with different colors, one that anneals to methylated DNA and another to unmethylated DNA. The Infinium II assay has one single probe that anneals to the DNA regardless of its methylation state, however the extension product will emit a color depending on whether the DNA is methylated or unmethylated.

There are an estimated 30 million CpGs in the human genome (202, 203), therefore those arrays only cover a small portion of CpGs. Additionally, the results obtained by a bead array methodology do not always correlate with PCR-based methods, due to several factors including the samples used, technical limitations of the array, and the power of statistical analysis performed (204). For example, Park and colleagues (176) identified several CpGs with potential to identify blood, saliva, semen and vaginal epithelia, according to the statistical analysis of the array data. However, looking at the pyrosequencing data for 8 CpGs selected and obtained by sequencing 20 samples of each body fluid, the results show such a high variability in the percent values of DNA methylation between samples of the same body fluid, that some samples of one body fluid could be mistaken by another body fluid. Additionally, from the 2986 CpGs identified as statistically significant to discriminate body fluids, the 8 validated by pyrosequencing were selected by the authors based on the following criteria: "1. the rank

of the p value in each body fluid, 2. hypermethylation in one body fluid, 3. a high methylation difference, and 4. practicality with regard to pyrosequencing." Despite of criteria number 3 being a high methylation difference between body fluids, all 4 CpGs selected for saliva and blood show a predicted percent of methylation of about 50-59%, which corresponds to an intermediate level. The 2 CpGs selected as potential markers for vaginal epithelia present an estimated percent of methylation somewhat higher 65-72% but with standard deviations of 13-22%. In summary, even if only considering the statistical analysis of the results obtained from the array, the selected CpGs do not seem able to show large differences in methylation for the target body fluid when compared to other body fluid. For that reason, additional studies using either arrays or whole genome methods must be performed in order to increase the number of body fluids tested for forensic purposes.

Regarding the statistical analysis performed in the same report, it was made based on the beta-value, which is the default output value given by the software that accompanies the Infinium HumanMethylation450 bead array. The beta-value is defined as the ratio of the methylated signal versus the total signal (methylated+unmethylated signals). However, the beta-value is known to be highly heteroscedastic, that is, the variance across samples with beta values close to 0 or 1 is artificially reduced. In practical terms, this means that when performing a statistical analysis based on the beta-value, the CpGs that present beta-values close to 0 or 1 can show low variability from sample to sample. To tackle this limitation, the M-value, which is defined as the log of the ratio of the methylated signal versus the unmethylated signal, is preferable since it displays an approximately constant variation across the methylation range of values (204, 205).

Park and colleagues adopted a statistical analysis of the array data by performing

Student's t-test "to evaluate the significance of differences in DNA methylation levels

between any one type of body fluid and the other types of body fluids." This use of the

Student t-test implies that the average methylation of all 4-6 samples of one body fluid

for one CpG on the array will be compared to the average methylation of all other

samples from all other body fluids in the same CpG. This approach may allow the sorting

of those CpGs that present bigger differences between body fluids since to show

significance, the methylation on the target body fluid, for that specific CpG, must be

quite different from all others, regardless of which body fluid is in the other group.

However, DNA methylation for three groups of body fluids was measured separately,

which would make the analysis of variance (ANOVA) more appropriate as a statistical

test. Additionally, ANOVA also provides an output considering the variance between the

methylation values within each group, which may result in loci with true differences in

mean methylation values (206).

Finally the use of a PCR-based method to confirm the results obtained by the

array should allow for an easy and quick distinction between body fluids, even if it

cannot quantify methylation at individual CpGs. Due to its easy use, fast turnaround and

inexpensive properties, HRM has the potential to provide the confirmation needed for

array-based methods.

Taking this into account, the goal of this work was to perform a statistical analysis

of the array data reported by Park *et al.* (2014) and screen the results obtained using

HRM to identify new potential CpG able to discriminate blood, saliva and vaginal

epithelia. Additionally it is also the goal of this work to select a set of primers able to be

amplified in a multiplex PCR reaction for body fluid identification.

This project was submitted to the National Institute of Justice on the form of a

Graduate Research Fellowship grant proposal which was awarded in 2015 (2015-R2-CX-

0012).

B. Materials and Methods

Samples

Blood, saliva, vaginal epithelia and semen samples were collected from

volunteers according to the approved IRB (IRB-16-0021) from Florida International

University. Blood samples were collected by finger prick and soaking the blood onto

sterile cotton swabs. Saliva samples were collected by asking volunteers to rub the inside

of their cheek with a sterile swab and vaginal epithelia and semen samples were collected

by volunteers on their own time and later returned to the laboratory. All swabs were dried

and stored at room temperature until DNA extraction was performed. Aliquots of semen

samples were created and frozen at -20 ºC until DNA extraction was performed. DNA

control samples were purchased and consisted of human fully methylated and

unmethylated DNA, not bisulfite modified (Zymo Research, Irvine, CA) or already

bisulfite modified (Qiagen, Valencia, CA).

DNA extraction, quantification and bisulfite modification

Each swab or liquid semen sample was placed in a 2 mL tube containing lysing

buffer G2 (Qiagen, Valencia, CA) and proteinase K, according to the manufacturer's

instructions for the EZ1 DNA Investigator kit (Qiagen, Valencia, CA). After 3 hours of incubation at 56 ºC, the tubes were dried, quickly spun and the swabs transferred to spin baskets in order to remove the liquid absorbed by them. After centrifugation, all the lysate was transferred to a 2 mL tube and Buffer MTL (Qiagen, Valencia, CA) was added. The extraction procedure was performed automatically by the Biorobot EZ1 Automated purification system (Qiagen, Valencia, CA), according to the manufacturer's specifications. DNA quantification was performed either by the PicoGreen method (90, 91) for blood samples or by real-time quantification using Alu primers (79) for all other body fluids, due to the likelihood of containing non-human DNA derived from naturally occurring microflora.

Two hundred microliters of DNA from samples or controls (Zymo Research, Irvine, CA) were bisulfite modified using the Epitect® Fast DNA Bisulfite kit (Qiagen, Valencia, CA) which converts unmethylated cytosines but not methylated ones to uracil. The bisulfite-modified DNA was eluted by adding 20 μL of elution buffer to the silica column from the kit.

Bioinformatic Analysis

The raw data for the array study reported by Park *et al.* (2014) was retrieved from Gene Expression Omnibus (GEO) accession number GSE55734. Using Microsoft Excel the M-value was calculated by performing the log of the ratio of the methylated signal versus the unmethylated signal for all 6 blood, 6 vaginal, and 4 saliva samples of over 450,000 CpGs. Statistical analysis was performed using R software and in collaboration with Dr. Vanessa-Aguiar Pulido from the BioRG group at Florida International

University. The analysis consisted in performing ANOVA and Kruskal-Wallis tests followed by post-hoc Tukey HSD and Dunn's tests, respectively. The CpG locations that showed a *p-value*<0.05 on both tests were selected. To further filter the results, CpGs were selected based on the ratio of methylated and unmethylated signals. The sorting of CpGs was performed in a two-step fashion; first by filtering locations with ratios above 2 for the target body fluid and below 0.5 for the other two body fluids; and secondly by filtering those locations with a ratio below 0.5 for the target body fluid and above 2 for the other two body fluids. If too many locations were still returned, the values of 2 and 0.5 were increased and lowered, respectively. This type of sorting returns CpGs that present more than double of signal intensity for the methylated probes when compared to the unmethylated probes (ratio $\geq$ 2) of one body fluid, while the two remaining body fluids present more than double the signal for the unmethylated probe when compared to the methylated one (ratio $\leq$ 0.5).

Primer design

The raw array data provides the identification of each CpGs with a 'cg' code attributed by Illumina (San Diego, CA) and deposited in genome browsers. For this study, the genome browser from University of California at Santa Cruz (http://genome.ucsc.edu/), assembly GRCh37/hg19 was used to search for each 'cg' code sorted from the bioinformatics analysis, to which corresponds to a specific CpG in the human genome (207). A DNA sequence consisting of 300 bp upstream and 300 bp downstream of the CpG was downloaded and used to design primers. Three different tools were used to design primers, the online tools Methprimer (200) and Bisearch (208,

209) and the PyroMark Assay Design SW version 2.0 (Qiagen, Valencia, CA). The tools work by initially assuming that all cytosines not followed by guanine are unmethylated and therefore are considered thymines by the algorithm. Then, the primers are designed based on those sequences and some user-defined parameters. Following the guidelines by Wojdacz and colleagues (210, 211), the parameters where set to allow only 1 ºC difference on the melting temperature (T_M) of forward and reverse primers, an optimal T_M of primers near 65 ºC, allowing one CpG within the primer sequence and setting 100 bp as the preferred amplicon length. Due to the fact that bisulfite modified DNA is almost a 3-letter genome sequence, the difficulty to design primers specific for one region is increased. Furthermore, some of the CpGs selected were located in areas with high density in either CpGs or thymines. Thus, the parameters for primer design were modified whenever none of the primer design tools could design any oligo sequences. The parameters for primer design were altered consecutively by first changing the preferred length of amplicon size from 100 bp to 150 bp and a maximum 300 bp. If no primer sequences were returned, a second change was made which consisted on altering the preference of having one CpG on the primer sequence to having none or more than one. If still no primer was designed, the stringency on having a specific primer T_M was lowered which allows more dissimilar T_M between forward and reverse oligos and T_M outside the vicinity of 60 ºC. Fifty different regions were tested for blood and vaginal epithelia, and 31 tested for saliva.

PCR amplification and melt curve analysis

Initially the amplification of bisulfite-modified DNA was performed using the Epitect® HRM kit (Qiagen, CA). Two microliters of bisulfite modified DNA (samples or controls) were added to the master mix according to manufacturer's instructions, with 0.5-0.75 μM of each forward and reverse primers. Ultimately a custom made master mix was optimized ('My Mix') which consists of 1X PCR buffer (Qiagen, CA), 0.125U of Hot Star Taq Plus (Qiagen, CA), 0.2 mM of each dNTP (Denville Scientific, MA), 0.5-0.75 μM of each forward and reverse primers, 1X of EvaGreen (in water, Biotium Inc., CA), 0.2 mg/mL of Bovine Serum Albumin (BSA, Sigma Aldrich, MO) and 2.5% of Dimethyl sulfoxide (Sigma Aldrich, MO) in a total volume of 20 μL in the PCR reaction which also contained 2 μL of bisulfite-modified DNA. Optimization was performed by combining different concentrations of master mix constituents and additives and determining which provided a C_T and T_M similar to those obtained by amplifying the ZC3H12D primer with the Epitect® HRM kit (Qiagen, CA). The concentrations to be analyzed were determined by the Design Expert® (version 10, Stat-Ease, Inc.). The singleplex PCR reactions were introduced in the real-time instrument which was set to an initial hold at 95 ºC for 5 min, followed by 45-55 cycles of 95 ºC for 10 seconds, 40-60 ºC (depending on T_M for the primers being amplified) for 30 seconds and 72 ºC for 10 seconds. The multiplex reactions were amplified with similar settings except the annealing step was performed at 54 ºC for 40 seconds and the extension step was modified to 20 seconds. Melt analysis for singleplex and multiplex reactions were performed by increasing the temperature from 65 to 85 ºC in 0.3 °C increments and

detecting fluorescence in the HRM channel. Melt curve analysis was made using the

Rotor-Gene 6000 Series Software version 1.7.

C. Results

Bioinformatic analysis

From the initial comparison between body fluids 30,075 individual CpGs were

returned for blood, 5,667 for saliva and 9,469 for vaginal epithelia by using both

statistical tests (ANOVA and Kruskal-Wallis), and defining the p-value<0.05. A filter

was applied to that initial list so that the ratio of methylated signal versus the

unmethylated signal for the target body fluid would be very low or very high whereas for

the other two body fluids it would be of opposite magnitude. There were several rounds

of filtering as each CpG got experimentally tested. The actual values used and the

number of CpGs filtered per set of values are shown on Table 7.1. The total number of

newly identified CpGs to be tested was aimed at 50, which was not possible for saliva

since only 31 new CpGs showed methylation differences considered relevant between

saliva and other body fluids, according to our defined parameters. Thus, 50 CpGs were

tested for blood and vaginal epithelia and 31 CpGs tested for saliva.

Experimental Results using HRM

The selected CpGs and the surrounding DNA sequence was retrieved from the

UCSC genome browser online platform (207) in order to allow for primer design. Several

of the CpGs selected (Table 7.1) had to have several sets of primers designed because the

initial set did not amplify properly or showed very inefficient amplification. For blood,

only one CpG had 2 primers designed, saliva had 6 primers designed twice and 1 primer

designed three times and vaginal epithelia had 5 primers designed twice.

In the initial screen for all primers, 1 sample of each body fluid and both

methylated and unmethylated controls were amplified.

Table 7.1 – Summary of parameters used to filter CpGs by ratio of methylated versus unmethylated probes for each target body fluid. The number of CpGs displayed correspond to those obtained from each parameter only.

Target Body Fluid	Methylated/unmethylated Target body fluid	Methylated/unmethylated Other two body fluids	# CpGs per filtering criteria
Blood	<0.1	>1	5
	<0.5	>1	24
	<0.5	>2	3
	≥2	≤0.5	17
Saliva	<0.6	>1	10
	<1	>1	1
	<0.5	>0.8	11
	<0.6	>0.8	4
	≥2	≤0.5	5
Vaginal Epithelia	<0.2	>0.5	16
	<0.1	>0.5	15
	<0.4	>1	3
	<0.5	>1	15
	<0.6	>1.9	3
	≥2	≤0.5	3

The primer sets that showed a different T_M between the target body fluid and the other

three body fluids were selected to test with additional samples. A total of 12 blood, 3

saliva and 6 vaginal epithelia CpGs were tested with multiple samples, from which 3

locations for blood, 1 for vaginal epithelia and 1 for saliva (Table 7.2) provided

consistently different T_M for 10 samples of the target body fluid when compared to 10

samples of each one of the other body fluids (Table 7.3).

Table 7.2 – List of primer sequences for the 5 markers identified in this study. Fw stands for forward primer and Rv stands for reverse primer

Marker Name	'cg' code		Sequence 5'-3'
INPP5D200_TM60	cg00438740	Fw	AAGGAGAGGGTTGGTAAGAGAGT
		Rv	CAATAAAAAAAAAAAAACTAAACTACRAAAAAC
INPP5D1500	cg02935132	Fw	TTATGTTTTTTTAGAAGTGGT
		Rv	TCTCCTTCCTAACAACTAAT
INPP5D200_2	cg22029587	Fw	TTTTTTGTTTGGTTTTTGTAATGAGGAAGT
		Rv	AAACACCCCCAAAACCCACAC
VE_8	cg08751438	Fw	TTAAATTAGGGTGTGGGTAGA
		Rv	AAACATACCAAAAAAACAAAAC
VE_15	cg25975690	Fw	GTGAAAGGAAAGAGAGAATAAATATT
		Rv	CAAAAACAAACAAAAAACCTTTC

Table 7.3 – Summary of the T_M values for the loci identified in this study as potential markers for blood and vaginal epithelia (VE). $T_M \pm$ SD for each body fluid represents an average melting temperature value and standard deviation calculated for 10 samples analyzed per body fluid, 3-4 samples analyzed for the unmethylated (UnM) and methylated control (M).

Target Body Fluid	Marker Name	T_M (ºC) ± S.D.					
		Blood	Saliva	Semen	VE	UnM	M
Blood	INPPD200_TM60	80.0 ±0.1	83.0 ±0.2	82.7 ±0.3	82.6 ±0.4	80.0 ±0.2	83.0 ±0.2
	INPPD1500	74.3 ±0.1	75.8 ±0.1	75.5 ±0.5	75.3 ±0.8	74.1 ±0.1	75.6 ±0.1
	INPPD20_2	80.0 ±0.1	81.6 ±0.7	81.6 ±1.1	80.2 ±0.5	80.0 ±0.1	82.1 ±0.1
VE	VE_8	78.1 ±0.1	78.1 ±0.1	78.1 ±0.1	76.6 ±0.2	76.5 ±0.1	78.1 ±0.1

The CpGs able to distinguish saliva (cg25975690) was discovered from the list of

potential CpGs for vaginal epithelia by showing an intermediate genotype for saliva (10

samples) when compared to the methylated control and all other body fluids (10 samples of each body fluid) and the unmethylated controls.

The genotype discrimination function on the Rotor Gene software consists of a curve showing the decrease in fluorescence as a result of the melting of DNA and release of intercalating dye molecules with the increase in temperature. Each line depicting each sample is the fluorescent signal resulting from a set of molecules of DNA melting at the same time.
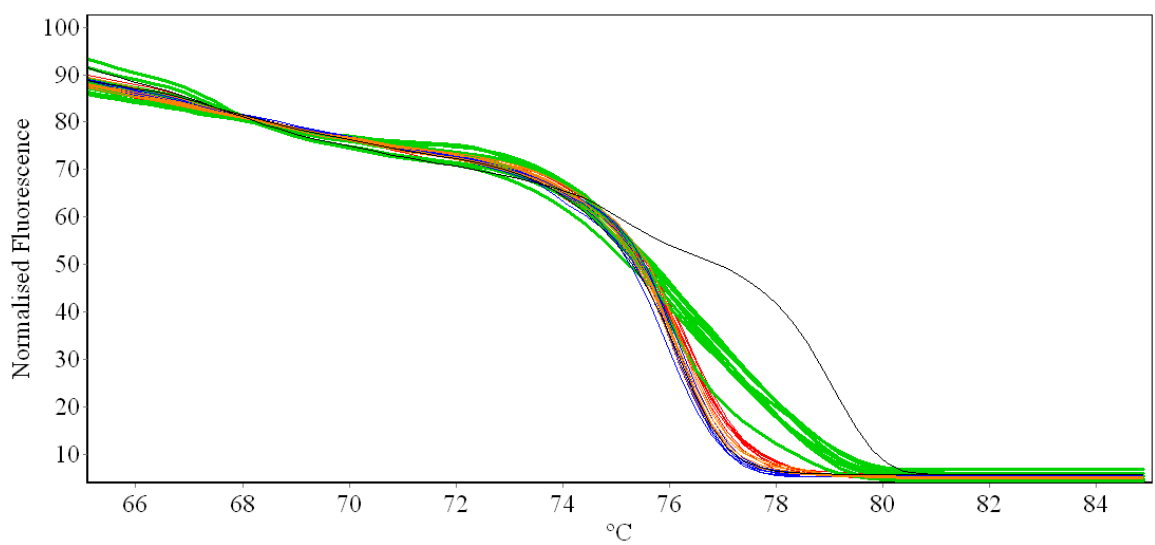
Figure 7.1 – Genotype differences between saliva (n=10), in green color and thick lines, when compared to blood (n=10) in red color, vaginal epithelia (n=10) in orange and semen (n=10) in blue. Controls used are depicted in black.

Considering the area near the T_M, a steeper drop in fluorescence indicates that all DNA molecules in each sample have the same amplicon sequence, a less steeper drop in fluorescence may indicate that the DNA molecules vary in the levels of methylation with the result that each molecule melts at slightly different rate around the same T_M (Figure 7.1).

Optimization of a PCR master mix to amplify bisulfite-modified DNA

A commercial kit, Epitect® HRM (Qiagen, CA), consisting of a pre-made master mix that contains dNTPs, magnesium, Eva Green as intercalating dye and Hot Star Taq Plus (Qiagen, CA) in a buffer solution that contains other additives was initially used. In order to design our own master mix, the Material Safety Data Sheet (MSDS) for the Epitect® HRM kit is available online and was consulted (212). The MSDS showed that the master mix on the kit contains glycerol and DMSO each with concentrations between 1 and 10%. The information regarding concentrations of magnesium, dNTPs, polymerase and eva green dye used were not available on the MSDS. The polymerase in the kit (Hot Star Taq Plus DNA polymerase, Qiagen, CA) is available for purchase as individual reagent and therefore it was also used on a newly developed master mix named 'My Mix'. The purchased polymerase includes a handbook that served as a guideline for the optimization of 'My Mix' (213). Additionally, the presence of additives was also investigated.

The software Design Expert (Stat-Ease, MN) provided a list of concentrations tested for each additive to be added to the mixture, namely BSA, betaine, DMSO and glycerol, and also provided guidelines for the concentration of dNTPs and magnesium. Each concentration was tested using control unmethylated and methylated DNA and the primer pair for ZC3H12D. The amplification efficiency and melt curve resulting from 'My Mix' and the Epitect® HRM kit were compared. The comparison allowed to determine a concentration of additives, dNTPs and magnesium for My Mix that presented similar results to those observed when the Epitect® HRM kit was used. As an example, Figure 7.2 displays the melt curves for the methylated (red) and unmethylated (green)

176

controls when amplified in the presence of 0.2 mg/mL of BSA and either 2.5% glycerol

(light green and dark red) or 2.5% DMSO (dark green and light red). The T_M for the

peaks in the figure were compared with the dashed lines that represent the T_M for the

unmethylated control (green line) and the methylated control (red line) when amplified

with the Epitect® HRM kit.
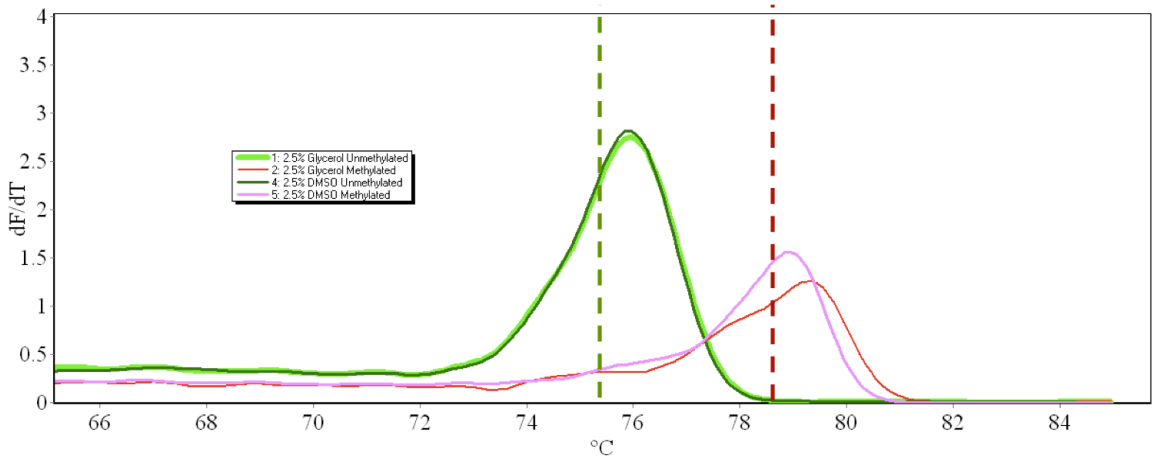
Figure 7.2 - Melt curve obtained in the presence of 2.5% DMSO or Glycerol in 'My Mix' which includes BSA (0.2 μg/μL). The dashed lines represent the melt temperature for the DNA controls amplified with the Epitect® HRM kit (Qiagen, CA) used as reference for 'My Mix', where green is the unmethylated control and red is the methylated control.

The optimized mixture consisted of 1X PCR buffer (Qiagen, CA), 0.125U of Hot

Star Taq Plus (Qiagen, CA), 0.2 mM of each dNTP, 0.5-0.75 μM of each forward and

reverse primers, 1X of EvaGreen (in water) 0.2 mg/mL of BSA and 2.5% of DMSO in a

total volume of 20 μL in the PCR reaction.

Multiplex for markers showing specific amplification of blood, semen and vaginal

epithelia

From Table 7.3, one marker per target body fluid was selected for amplification in

the same PCR reaction as the marker ZC3H12D, previously described as being

177

unmethylated for semen, but not blood or saliva. Thus, the primer set INPD200_TM60

and VE_8 were added to a master mix containing ZC3H12D. After optimization, the final

concentrations of each primer set were determined as 0.75 µM for ZC3H12D, 0.75 µM

for each INDP200_TM60 and VE_8. The average T_M of 3 samples per body fluid for

each primer when amplified with 54 ºC annealing temperature and extended times for the

annealing and extension steps of the PCR cycle is displayed in Table 7.4.

Table 7.4 – Average T_M of amplified samples with the markers INPPD200_TM60, VE_8 and ZC3H12D at 54 ºC annealing. Values displayed are of 3 samples of blood, saliva, semen and vaginal epithelia (VE) and 2 samples of each unmethylated (UnM) and methylated (M) control.

Target Body Fluid	Marker Name	'cg' code	T_M (ºC)					
			Blood	Saliva	Semen	VE	Un M	M
Blood	INPPD200_TM60	cg00438740	79.5	82.5	82.4	82.2	79.5	82.4
VE	VE_8	cg08751438	77.7	77.5	78	76.2	76.2	77.7
Semen	ZC3H12D	N/A	78.7	78.9	75.6	78.8	75.8	78.8

Figure 7.3 shows the melt curves for the multiplex amplification of all 3 primer sets in

the same PCR conditions. The T_M for ZC3H12D and VE_8 are too similar and are

therefore displayed as one single peak, making them indistinguishable.
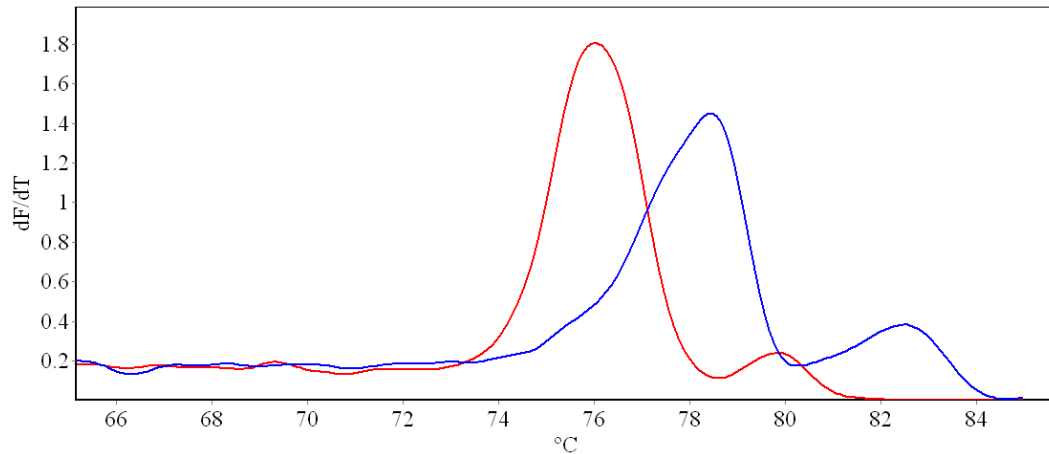
Figure 7.3 – Melt curve for an unmethylated control (red) and a methylated control (blue) amplified in a multiplex PCR reaction with INPP5D200_TM60, VE_8 and ZC3H12D, at an anneling temperature of 54 ºC.

D. Discussion

There have been continuous efforts to develop loci capable of discriminating body fluids at crime scenes using tissue-specific DNA methylation (167, 176). In an attempt to increase the number of DNA methylation markers, we have re-analyzed array data from the study of Park and colleagues from 2014. A new M-value was used to calculate the ratio between the methylated and unmethylated signal with the goal of improving the statistical analysis. Initially we obtained a higher number of potential CpGs for each target body fluid when compared to Park *et al.* (2014). We were able to obtain 30,075 CpGs for blood, 5,667 for saliva and 9,469 for vaginal epithelia, whereas Park and colleagues selected 618 for blood, 37 for saliva and 176 for vaginal epithelia. Interestingly, the number of new locations follows a similar trend when comparing body fluids. In both studies, a higher number of CpGs were selected for blood, followed by vaginal epithelia and lastly saliva.

179

Since we were targeting a simple method of analysis –HRM – we were able to screen 50 new CpGs for blood and vaginal epithelia and 31 for saliva. On the initial screening, it was easy to identify new CpGs for blood, however the same was not true for vaginal epithelia or saliva. The blood samples collected for this study and for the array study consisted of whole blood, since that is what is commonly left at crime scenes. Whole blood is composed of several cell types but only the nucleated cells provide genomic DNA (214). In comparison, saliva and vaginal epithelia are composed of more than one type of cells, including epithelial cells and immune cells (215, 216). Moreover, research has shown that DNA methylation differences are observed between the different cell types present in saliva (217). Additionally, we speculate that the immune cells existing in saliva and vaginal secretions may present higher variability in terms of DNA methylation due to its induced tolerance resulting from the presence of commensal bacteria (218). Taken together, the observed variability in cells present in saliva and vaginal epithelia may account for a more variable DNA methylation signature when compared to blood. Such variability could therefore explain why markers for blood seem to be easier to find when compared to saliva and vaginal epithelia. Additionally, the same differences in DNA methylation can also explain why a CpG initially selected as a potential marker for vaginal epithelia, showed in fact, superior results for saliva when compared to the methylation values for other body fluids. After the initial screening with one sample from each body fluid and one from each control, several markers were determined as potential vaginal secretion markers but ended up not showing consistent results when an additional 9 samples from each body fluid were analyzed.

The CpGs that showed a different T_M in blood when compared to other body fluids were cg00438740, cg02935132 and cg22029587, which were all located on a CpG island upstream of the RefSeq gene INPP5D. The gene INPP5D is expressed exclusively in hematopoietic cells, which are the precursors of blood cells (207). The marker found for vaginal cells was cg08751438, which is located on the q24.1 band of chromosome 16. It is not associated with any RefSeq gene (Table 7.3).

The CpG which showed specific methylation for saliva when compared to the other body fluids was cg25975690, located on a CpG island located upstream of the RefSeq gene HOXA9, which is a homeobox A9 coding for transcription factors that may be involved in gene expression (207). This CpG was initially selected as a vaginal epithelia-specific CpG from the statistical analysis of the bead array. However when the location was amplified with blood, saliva, semen and vaginal epithelia, this location showed differences in melting characteristics of DNA from saliva. It was interesting to notice that the difference in melting temperature for saliva samples results from a steady loss of fluorescence as the DNA molecules melt. In samples that have all their DNA strands fully methylated or fully unmethylated, the T_M should be displayed as a sharp peak on the graph of the first derivative of fluorescence over the temperature. The first derivative represents a graphical illustration of how the fluorescence decreases for each increase in temperature. If the DNA molecules present in a given sample contain different levels of methylation, with some being methylated, and others unmethylated, the decrease in fluorescence will be gradual since each group of strands will melt at a different temperature (Figure 7.1). This seems to be the case for the cg25975690 for saliva, but not for the other body fluids tested, allowing its discrimination by melt characteristics.

Regarding the optimization of a master mix to use with bisulfite-modified DNA, the initial mix was tested according to the handbook specific for the Hot Star Taq Plus DNA polymerase (Qiagen, CA). However since the polymerase was not specifically developed for application to amplify bisulfite-modified DNA, the booklet does not take into account the complexity of such DNA templates. Several additives were tested to improve the amplification and melt curve analysis step. Bovine serum albumin (BSA) is commonly used to prevent the master mix components, specifically the polymerase from adhering to the tube walls, thus increasing their availability for template amplification. Betaine and DMSO act by altering the melt characteristics of DNA either by equilibrating the T_M between GC and AT pairs, or by disrupting inter- and intra-strand re-annealing avoiding the formation of DNA secondary structures such as hairpins and others (219). According to the MSDS for the Epitect HRM® kit, which was the kit used to amplify bisulfite-modified DNA, the master mix available in the kit contains 1-10% DMSO and glycerol. Results from Farell and Alexandre, has shown that the combination of BSA and DMSO can enhance the amplification of GC-rich DNA templates (220). During experiments on the effects of DMSO and glycerol in My Mix, it was discovered that the melt peak for both the unmethylated and methylated controls show similar T_M for DMSO and glycerol, however the overall shape of the methylated peak was sharper with DMSO than with glycerol, which is why that compound was picked as an additive to be included in My Mix (Figure 7.2).

In the development of a multiplex capable of amplifying the CpGs that show specific DNA methylation patterns for blood, semen and vaginal epithelia, we obtained mixed results. On one hand, the amplification was successful since all 3 primer sets

182

amplify properly. However since two of the melt curves show peaks at similar T_M, they overlap, making analysis very difficult (Figure 7.3). The initial goal was to develop a system similar to Plexor™ (Promega), which would allow the monitoring of amplification and melt curve for each primer set individually through the use of different fluorophores for each primer set (Figure 2.3, Chapter II). The elegance of that method makes it quite attractive to use in the DNA methylation analysis of multiplex reactions. However, the Plexor™ technology is no longer available as a product that can be customized.

Other strategies were analyzed, however few offer the ability to monitor both amplification and melt curve analysis in a single step. One option was the use of a Snapback primer, which consists of a regular primer with a tail that is complementary to the target region on the amplicon, folding over the original primer and forming an hairpin (194). The melt temperature for that hairpin is smaller than the regular amplicon. One of the primers (either the one specific for vaginal epithelia or the one specific for semen) would have the extra sequence allowing the hairpin to form, thus distinguishing the melt curves based on the T_M of the hairpin and not of the major amplicon peak. Several snapback primers were designed and tested but the melting peak for the hairpin was not observed. Further experiments are continuing with the assistance of Drs. Luming Zhou and Zachary Dwight from University of Utah (221).

E. Conclusion

With the advance in technology regarding the match of a suspect with a crime scene using DNA, in certain cases, the identification of a body fluid can confirm or acquit

the suspect. The confirmatory power that DNA methylation patterns can have in such cases relies on the identification of several CpGs able to display a unique pattern of DNA methylation in one body fluid and a different pattern for all other body fluids. Thus, identifying new CpGs of interest can greatly aid the forensic discrimination of body fluids. Using a new bioinformatic analysis, we were able to screen potentially new CpGs and identify 5 new CpGs with a specific DNA methylation pattern for blood, saliva and vaginal epithelia. Moreover we were also able to show the possibility of amplifying simultaneously 3 markers which may represent an important step on the use of DNA methylation markers in forensic laboratories in a near future.

CHAPTER VIII – CONCLUDING REMARKS

The description of the chemical structure of DNA and the sequence of the human genome represented major events in the recent past of mankind. In addition, the optimization of methods like the polymerase chain reaction (PCR) made it possible to study and use the information stored in DNA for several applications. One of those applications is law enforcement. The ability to match the DNA of a suspect to the DNA found at a crime scene represents one of the most powerful piece of evidence on convictions or exonerations. However, information stored in DNA can be further used in other forensic applications. For example, patterns of DNA methylation allow the discrimination between body fluids. DNA methylation represents one type of epigenetic modification that allows differential gene expression in different tissues. The ability to identify such differential gene expression in forensic laboratories has the potential to overcome issues with current presumptive methods for body fluid identification and to promote investigative leads such as age prediction.

To successfully use DNA methylation in forensic laboratories, several parameters must be satisfied. One of those parameters has to do with the methodology and its reproducibility. DNA methylation consists of the presence of a methyl group on the 5-carbon of specific cytosines spread throughout the human genome. The extracted genomic DNA must undergo a chemical modification that uses sodium bisulfite and causes DNA fragmentation. Such bisulfite modification allows the conversion of unmethylated cytosines to uracils which can be detected with downstream analysis methods. Since methylated cytosines are protected from this modification, it is important

185

to demonstrate complete conversion during analysis. Additionally, the conversion process can result in severe fragmentation on the extracted DNA, thus careful primer design and use of controls is important for optimal results. In fact, several reports have been published in the last years showing promising results regarding the usability and validation of these methods in forensic sciences. The use of methods capable to distinguish DNA methylation for body fluid identification is preferential to those that analyze proteins or mRNA since the same extracted DNA sample used for STR analysis can be used for DNA methylation analysis, without the need to use additional evidence. In this work we demonstrate that HRM analysis can be used to distinguish body fluids that present different levels of DNA methylation in specific loci. The results here described represent our contribution to the field of forensic epigenetics.

We were able to demonstrate the reproducibility and usability of some markers discovered by our team. Namely we showed that ZC3H12D, BCAS4 and cg06379435 show differential methylation analyzed by pyrosequencing, are primate-specific, have sensitivity in a low nanogram quantity, can detect mixtures of body fluids, can withstand degradation by UV light and high temperatures and allow the discrimination of semen, saliva and blood, respectively. Moreover, we were also able to show that the ZC3H12D locus is an excellent marker for semen since it is hypomethylated when compared to saliva and blood. The marker ZC3H12D represents the ideal DNA methylation markers to be used in forensic sciences since all the CpGs amplified in the loci show levels of methylation below to 10% for semen, whereas other body fluids are highly methylated on the same CpGs. Due to the high difference in methylation when comparing body fluids and also the consistency of methylation in all of the CpGs included in the loci, this

marker can provide usable results even when using a non-quantitative method such as DNA melt analysis. The use of high-resolution melt (HRM) analysis has benefits due to the high throughput and low cost, when compared to other analysis methods.

The current number of markers available for each body fluid is low considering that mixtures of body fluids will provide intermediate values of methylation. For that reason, it is needed to identify new loci able to discriminate body fluids from crime scenes and the usability of ZC3H12D with HRM inspired us to pursue a project focused on performing bioinformatics analysis of published array data and then screen the potential CpGs using HRM. Those CpGs that showed consistent and differential methylation between the target and other body fluids on the array should also show different melting temperatures on HRM analysis. In this dissertation, results from screening work of newly identified CpGs provided new blood, saliva and vaginal epithelia markers capable of discriminating between those body fluids. Additionally some of those markers can be amplified in a multiplex PCR which is ideal for its use with crime scene samples as it permits multiple results from the same aliquot of DNA.

There is a specific need to identify markers for vaginal epithelia, since no presumptive test is commonly used in serological analysis in forensic laboratorie. To tackle that task, we examined published literature that used DNA methylation or mRNA transcripts for body fluid identification. One locus near the PFN3 gene showed mixed results in different reports. For that reason, we investigated specific CpGs on that locus to determine if they were able to distinguish vaginal epithelia from other body fluids. Through the use of pyrosequencing we established that vaginal epithelial cells show intermediate levels of methylation between the high methylation of blood and saliva and

187

the low methylation levels of semen. Moreover we also established the primate-specificity of that marker and the ability of providing results even with 1 ng of DNA.

Even though it has been known for several years that DNA methylation is the mechanism causing differential gene expression in mammals, research applying this knowledge to forensic sciences is only 7 years old. The results here presented are an important contribution to that body of knowledge and represent important progress in the development of DNA methylation differences for use in forensic sciences.

CHAPTER IX – FUTURE DIRECTIONS

The combined information on gene expression due to differential methylation established before birth, such is the case for tDMRs, or established during the lifespan of one individual, such is the case of loci that are modified with age, provide a powerful tool for forensic sciences. At this point, age prediction can only be achieved with the information of which body fluid originated the DNA sample. For that reason, it is extremely important to establish a set of several loci able to discriminate body fluid and predict age. Those loci can be amplified by multiplex PCR and analyzed in a massively parallel sequencing platform such as the MiSeq FGx (Illumina, CA), which is currently specific for forensic analysis of DNA and is the new alternative to capillary electrophoresis analysis; or on less established, but simpler platforms such as the MinIon (Oxford Nanopore, UK). The MinIon technology does not require DNA amplification, which represents a major advantage for forensics, especially for the analysis of DNA methylation patterns, since the bisulfite modification step may become unnecessary, depending on sensitivity requirments. Theoretically the MinIon has the ability to directly detect methylated cytosines, however it is such a recent technology has described the process (222).

Regardless of which technology will be used for analysis, the identification of more loci for body fluid discrimination will provide more reliable results, which is essential in forensic sciences. To solve that need, the whole human methylome, and not only a subset of it, must be analyzed using several samples from each type of body fluid commonly found in crime scenes. The large quantity of data resulting from such studies, requires

optimized bioinformatics analysis able to better identify CpGs of interest *in silico* prior to experimental screening, thus saving time and cost of analysis.

Ultimately, the use of DNA methylation patterns for body fluid identification and age prediction will have to be compatible with whatever platform is adopted in STR typing. Despite of the technology used, the findings of this study will further benefit from additional optimization to increase sensitivity as well as an increase of number of samples of each body analzyed to determine its robustness.

LIST OF REFERENCES

1. Franklin RE, Gosling RG. Molecular configuration in sodium thymonucleate. Nature. 1953;171:740-1.

2. Michelson A, Todd AR. Nucleotides part XXXII. Synthesis of a dithymidine dinucleotide containing a 3′: 5′-internucleotidic linkage. Journal of the Chemical Society (Resumed). 1955:2632-8.

3. Crick FH, Watson JD. The complementary structure of deoxyribonucleic acid. Proceedings of the Royal Society of London A: Mathematical, Physical and Engineering Sciences; The Royal Society; 1954.

4. Alberts B, Johnson A, Lewis J, Raff M, Roberts K, Walter P. Molecular Biology of the Cell. 4th Edition ed. New York: Garland Science; 2002.

5. Berg JM, Tymoczko JL, Stryer L. Biochemistry. 7th ed. Palgrave Macmillan; 2011.

6. Crick FH. The biological replication of macromolecules. Symp. Soc. Exp. Biol; ; 1958.

7. Voet D, Voet JG. Biochemistry. Fourth ed. USA: John Wiley & Sons, Inc.; 2011.

8. Watson JD, Baker TA, Bell SP, Gann A, Levine M, Losick R. Chapter 8 : The Replication of DNA. In: Wilbur B, Inglis J, editors. Molecular Biology of the Gene. 6th ed. USA: Pearson Education, Inc.; 2008. p. 195.

9. Watson J, Baker T, Bell S, Gann A, Levine M, Losick R. Part 3 - Expression of the Genome. In: Pearson Editors, editor. Molecular Biology of the Gene. 6th ed. Pearson; 2008.

10. Watson JD, Baker TA, Bell SP, Gann A, Levine M, Losick R. Chapter 12 : Mechanisms of Transcription. In: Wilbur B, Inglis J, editors. Molecular Biology of the Gene. 6th ed. USA: Pearson Education, Inc.; 2008. p. 377.

11. Watson JD, Baker TA, Bell SP, Gann A, Levine M, Losick R. Chapter 13 : RNA splicing. In: Wilbur B, Inglis J, editors. Molecular Biology of the Gene. 6th ed. USA: Pearson Education, Inc.; 2008. p. 195.

12. Berg JM, Tymoczko JL, Stryer L. RNA synthesis and processing. In: W.H. Freeman, editor. Biochemistry. 7th Edition ed. ; 2010. p. 883.

13. Watson JD, Baker TA, Bell SP, Gann A, Levine M, Losick R. Chapter 14 : Translation. In: Wilbur B, Inglis J, editors. Molecular Biology of the Gene. 6th ed. USA: Pearson Education, Inc.; 2008. p. 457.

14. Watson JD, Baker TA, Bell SP, Gann A, Levine M, Losick R. Chapter 15 : The genetic code. In: Wilbur B, Inglis J, editors. Molecular Biology of the Gene. 6th ed. USA: Pearson Education, Inc.; 2008. p. 521.

15. Berg JM, Tymoczko JL, Stryer L. Protein Synthesis. In: W.H. Freeman, editor. Biochemistry. 7th ed. ; 2010. p. 921.

16. Pollard T, Earnshaw W, Lippincott Schwartz J. Chapter 1. In: Cell Biology. 2nd ed. Philadelphia: Saunders/Elsevier; 2008. p. 3-16.

17. Pollard T, Earnshaw W, Lippincott Schwartz J. Chapter 19. In: Cell Biology. 2nd ed. Philadelphia: Saunders/Elsevier; 2008. p. 331-44.

18. Pollard T, Earnshaw W, Lippincott Schwartz J. Chapter 40. In: Cell Biology. 2nd ed. Philadelphia: Saunders/Elsevier; 2008. p. 731-46.

19. Watson JD, Baker TA, Bell SP, Gann A, Levine M, Losick R. Chapter 7 : Genome structure, chromatin, and the nucleosome. In: Wilbur B, Inglis J, editors. Molecular Biology of the Gene. 6th ed. USA: Pearson Education, Inc.; 2008. p. 135.

20. Waddington C. The epigenotype. Endeavour. 1942(1):18-20.

21. Han Y, Garcia BA. Combining genomic and proteomic approaches for epigenetics research. Epigenomics. 2013;5(4):439.

22. Allis CD, Jenuwein T. The molecular hallmarks of epigenetic control. Nature Reviews Genetics. 2016.

23. Watson JD, Baker TA, Bell SP, Gann A, Levine M, Losick R. Chapter 17 : Transcriptional regulation in eukaryoted. In: Wilbur B, Inglis J, editors. Molecular Biology of the Gene. 6th ed. USA: Pearson Education, Inc.; 2008. p. 589.

24. Mazzio EA, Soliman KF. Basic concepts of epigenetics: impact of environmental signals on gene expression. Epigenetics. 2012;7(2):119-30.

25. Zhu H, Wang G, Qian J. Transcription factors as readers and effectors of DNA methylation. Nat Rev Genet. 2016;17(9):551-65.

26. Park Y, Kuroda M. Epigenetic aspects of X-chromosome dosage compensation. Science. 2001;293(5532):1083-5.

27. Thomson JP, Skene PJ, Selfridge J, Clouaire T, Guy J, Webb S, Kerr AR, Deaton A, Andrews R, James KD, Turner DJ. CpG islands influence chromatin structure via the CpG-binding protein Cfp1. Nature. 2010;464(7291):1082-6.

28. Jiang Y, Liu S, Chen X, Cao Y, Tao Y. Genome-wide distribution of DNA methylation and DNA demethylation and related chromatin regulators in cancer. Biochimica et Biophysica Acta (BBA)-Reviews on Cancer. 2013;1835(2):155-63.

29. Ziller MJ, Müller F, Liao J, Zhang Y, Gu H, Bock C, Boyle P, Eptein BE, Lengauer T, Gnirke A. Genomic Distribution and Inter-Sample Variation of Non-CpG Methylation across Human Cell Types. PLoS Genetics. 2011 12;7(12):1-15.

30. Rivera RM, Ross JW. Epigenetics in fertilization and preimplantation embryo development. Progress in Biophysics & Molecular Biology. 2013;113(3):423-32.

31. Cedar H, Bergman Y. Programming of DNA Methylation Patterns. Annual Review of Biochemistry, Vol 81. 2012;81:97-117.

32. Jaenisch R, Bird A. Epigenetic regulation of gene expression: how the genome integrates intrinsic and environmental signals. Nat Genet. 2003;33(3s):245.

33. Suzuki MM, Bird A. DNA methylation landscapes: provocative insights from epigenomics. Nature Reviews Genetics. 2008;9(6):465-76.

34. Weber M, Schübeler D. Genomic patterns of DNA methylation: targets and function of an epigenetic mark. Curr Opin Cell Biol. 2007;19(3):273-80.

35. Quante T, Bird A. Do short, frequent DNA sequence motifs mould the epigenome? Nature Reviews Molecular Cell Biology. 2016.

36. Fernandez AF, Assenov Y, Martin-Subero JI, Balint B, Siebert R, Taniguchi H, Yamamoto H, Hidalgo M, Tan AC, Galm O, Ferrer I. A DNA methylation fingerprint of 1628 human samples. Genome Research. 2012;22(2):407-19.

37. Slieker RC, Bos SD, Goeman JJ, Bovee JVMG, Talens RP, van der Breggen R, Suchiman HE, Lameijer EW, Putter H, van den Akker EB, Zhang Y. Identification and systematic annotation of tissue-specific differentially methylated regions using the Illumina 450k array. Epigenetics & Chromatin. 2013;6:26.

38. Malygin EG, Hattman S. DNA methyltransferases: Mechanistic models derived from kinetic analysis. Critical Reviews in Biochemistry & Molecular Biology. 2012 04;47(2):97-193.

39. Chedin F. The DNMT3 Family of Mammalian De Novo DNA Methyltransferases. Modifications of Nuclear Dna and its Regulatory Proteins. 2011;101:255-85.

40. Rondelet G, Dal Maso T, Willems L, Wouters J. Structural basis for recognition of histone H3K36me3 nucleosome by human de novo DNA methyltransferases 3A and 3B. J Struct Biol. 2016 06;194(3):357-67.

41. Cheng X, Hashimoto H, Horton JR, Zhang X. Mechanisms of DNA methylation, methyl-CpG recognition, and demethylation in mammals. Handbook of epigenetics. 2010:9-24.

42. Szyf M. The elusive role of 5'-hydroxymethylcytosine. Epigenomics. 2016;8(11):1539-51.

43. Schübeler D. Function and information content of DNA methylation. Nature. 2015 01/15;517(7534):321-6.

44. Burris HH, Baccarelli AA. Environmental epigenetics: from novelty to scientific discipline. Journal of Applied Toxicology. 2014 FEB 2014;34(2):113-6.

45. Ho SM, Johnson A, Tarapore P, Janakiram V, Zhang X, Leung YK. Environmental epigenetics and its implication on disease risk and health outcomes. ILAR J. 2012;53(3-4):289-305.

46. Hou L, Zhang X, Wang D, Baccarelli A. Environmental chemical exposures and human epigenetics. Int J Epidemiol. 2011:dyr154.

47. Boyd-Kirkup JD, Green CD, Wu G, Wang D, Han JJ. Epigenomics and the regulation of aging. Epigenomics. 2013;5(2):205-27.

48. Bocklandt S, Lin W, Sehl ME, Sanchez FJ, Sinsheimer JS, Horvath S, Vilain E. Epigenetic Predictor of Age. Plos One. 2011;6(6):e14821.

49. Fraga M, Ballestar E, Paz M, Ropero S, Setien F, Ballestart M, Heine-Suñer D, Cigudosa JC, Urioste M, Benitez J, Boix-Chornet M. Epigenetic differences arise during the lifetime of monozygotic twins. Proc Natl Acad Sci U S A. 2005;102(30):10604-9.

50. Tan Q, Christiansen L, Thomassen M, Kruse TA, Christensen K. Twins for epigenetic studies of human aging and development. Ageing Research Reviews. 2013;12(1):182-7.

51. Teschendorff AE, West J, Beck S. Age-associated epigenetic drift: implications, and a case of epigenetic thrift? Hum Mol Genet. 2013;22:R7-R15.

52. Teschendorff AE, Menon U, Gentry-Maharaj A, Ramus SJ, Weisenberger DJ, Shen H, Campan M, Noushmehr H, Bell CG, Maxwell AP, Savage DA. Age-dependent DNA methylation of genes that are suppressed in stem cells is a hallmark of cancer. Genome Res. 2010;20(4):440-6.

53. Youngson NA, Whitelaw E. Transgenerational epigenetic effects. Annu.Rev.Genomics Hum.Genet. 2008;9:233-57.

54. Stewart L, Evans N, Bexon KJ, van der Meer DJ, Williams GA. Differentiating between monozygotic twins through DNA methylation-specific high-resolution melt curve analysis. Anal Biochem. 2015 5/1;476(0):36-9.

55. Vidaki A, Ballard D, Aliferi A, Miller TH, Barron LP, Court DS. DNA methylation-based forensic age prediction using artificial neural networks and next generation sequencing. Forensic Science International: Genetics. 2017;28:225-36.

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

57. Alghanim H, Antunes J, Silva, Deborah Soares Bispo Santos, Alho CS, Balamurugan K, McCord B. Detection and evaluation of DNA methylation markers found at SCGN and KLF14 loci to estimate human age. Forensic Science International: Genetics. 2017;31:81-8.

58. 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.

59. Jeffreys A, Wilson V, Thein S. Hypervariable Minisatellite Regions in Human Dna. Nature. 1985;314(6006):67-73.

60. Butler J. Chapter 3 - Historical Methods. In: Fundamentals of Forensic DNA Typing. 1st ed. USA: Elsevier, Inc.; 2010. p. 43.

61. Budowle B, Waye JS, Shutler GG, Baechtel FS. Hae III—A suitable restriction endonuclease for restriction fragment length polymorphism analysis of biological evidence samples. Journal of Forensic Science. 1990;35(3):530-6.

62. Lincoln PJ, Thomson J. Forensic DNA profiling protocols. Springer Science & Business Media; 1998.

63. Evett IW, Gill P. A discussion of the robustness of methods for assessing the evidential value of DNA single locus profiles in crime investigations. Electrophoresis. 1991;12(2-3):226-30.

64. Mullis K, Faloona F, Scharf S, Saiki R, Horn G, Erlich H. Specific enzymatic amplification of DNA in vitro: the polymerase chain reaction. Cold Spring Harb Symp Quant Biol. 1986;51 Pt 1:263-73.

65. Comey CT, Budowle B. Validation studies on the analysis of the HLA DQα locus using the polymerase chain reaction. Journal of Forensic Science. 1991;36(6):1633-48.

66. Duncan GT, Balamurugan K, Budowle B, Smerick J, Tracey M. Microvariation at the human D1S80 locus. Int J Legal Med. 1997;110(3):150-4.

67. Butler J. Chapter 2 - Basics of DNA biology and genetics. In: Fundamentals of Forensic DNA Typing. 1st ed. USA: Elsevier, Inc.; 2010. p. 19-41.

68. Butler J. Chapter 4 - Sample Collection, Storage and Characterization. In: Fundamentals of Forensic DNA Typing. 1st ed. USA: Elsevier, Inc.; 2010. p. 79-97.

69. Bond JW. Value of DNA evidence in detecting crime. J Forensic Sci. 2007;52(1):128-36.

70. Benecke M. Forensic DNA samples - collection and handling. In: Fuchs J, Podda M, editors. Encyclopedia of diagnostic genomics and proteomics. New York: Taylor & Francis, Inc.; 2005. p. 500.

71. Kobilinsky L. Recovery and Stability of DNA in Samples of Forensic Science Significance. Forensic Sci Rev. 1992 Jun;4(1):67-87.

72. Butler J. Chapter 5 - DNA extraction. In: Fundamentals of Forensic DNA Typing. 1st ed. USA: Elsevier, Inc.; 2010. p. 99-109.

73. Cattaneo C, Gelsthorpe K, Sokol R. DNA Extraction Methods in Forensic Analysis. Encyclopedia of Analytical Chemistry. 2000.

74. Comey CT, Koons BW, Presley KW, Smerick JB, Sobieralski CA, Stanley DM, et al. DNA extraction strategies for amplified fragment length polymorphism analysis. Journal of Forensic Science. 1994;39(5):1254-69.

75. Marmur J. A procedure for the isolation of deoxyribonucleic acid from micro-organisms. J Mol Biol. 1961;3(2):208IN1-218.

76. Vandeventer PE, Lin JS, Zwang TJ, Nadim A, Johal MS, Niemz A. Multiphasic DNA adsorption to silica surfaces under varying buffer, pH, and ionic strength conditions. The Journal of Physical Chemistry B. 2012;116(19):5661-70.

77. Vandeventer PE, Mejia J, Nadim A, Johal MS, Niemz A. DNA adsorption to and elution from silica surfaces: influence of amino acid buffers. The Journal of Physical Chemistry B. 2013;117(37):10742-9.

78. Nguyen TH, Elimelech M. Plasmid DNA adsorption on silica: kinetics and conformational changes in monovalent and divalent salts. Biomacromolecules. 2007;8(1):24-32.

79. Butler J. Chapter 6 - DNA quantitation. In: Fundamentals of Forensic DNA Typing. 1st ed. USA: Elsevier, Inc.; 2010. p. 111-24.

80. Quantifiler Duo® DNA quantification kit [Internet].; 2008. Available from: https://www.thermofisher.com/us/en/home/industrial/forensics/human-identification/forensic-dna-analysis/dna-quantification-solutions/quantifiler-duo-dna-quantification-kit.html.

81. PowerQuant® System [Internet].; 2017. Available from: https://www.promega.com/products/genetic-identity/genetic-identity-workflow/human-specific-dna-quantitation/powerquant-system/?catNum=PQ5002.

82. Plexor® HY System for the Applied Biosystems 7500 and 7500 FAST Real-Time PCR Systems [Internet].; 2016 [updated April 2016]. Available from: https://www.promega.com/-/media/files/resources/protocols/technical-manuals/0/plexor-hy-system-for-the-applied-biosystems-7500-and-7500-fast-real-time-pcr-systems-protocol.pdf.

83. Swango KL, Hudlow WR, Timken MD, Buoncristiani MR. Developmental validation of a multiplex qPCR assay for assessing the quantity and quality of nuclear DNA in forensic samples. Forensic Sci Int. 2007;170(1):35-45.

84. Higuchi R, Fockler C, Dollinger G, Watson R. Kinetic PCR analysis: real-time monitoring of DNA amplification reactions. Nat Biotechnol. 1993;11(9):1026-30.

85. Quantifiler® HP and Trio DNA Quantification Kits User Guide [Internet].: ThermoFisher Scientific; [updated October 1; 2015]. Available from: https://tools.thermofisher.com/content/sfs/manuals/4485354.pdf.

86. Thompson RE, Duncan G, McCord BR. An Investigation of PCR Inhibition Using Plexor (R)-Based Quantitative PCR and Short Tandem Repeat Amplification. J Forensic Sci. 2014;59(6):1517-29.

87. Moreno LI, McCord BR. The use of direct analysis in real time (DART) to assess the levels of inhibitors co-extracted with DNA and the associated impact in quantification and amplification. Electrophoresis. 2016;37(21):2807-16.

88. Pionzio AM, McCord BR. The effect of internal control sequence and length on the response to PCR inhibition in real-time PCR quantitation. Forensic Science International-Genetics. 2014;9:55-60.

89. Nicklas JA, Buel E. Development of an Alu-based, real-time PCR method for quantitation of human DNA in forensic samples. J Forensic Sci. 2003;48(5):936-44.

90. Blotta I, Prestinaci F, Mirante S, Cantafora A. Quantitative assay of total dsDNA with PicoGreen reagent and real-time fluorescent detection. Annali dell'Istituto superiore di sanità. 2005;41(1):119.

91. Dragan A, Casas-Finet J, Bishop E, Strouse R, Schenerman M, Geddes C. Characterization of PicoGreen interaction with dsDNA and the origin of its fluorescence enhancement upon binding. Biophys J. 2010;99(9):3010-9.

92. Mandrekar MN, Erickson AM, Kopp K, Krenke BE, Mandrekar PV, Nelson R, et al. Development of a human DNA quantitation system. Croat Med J. 2001;42(3):336-9.

93. Opel KL, Chung D, McCord BR. A study of PCR inhibition mechanisms using real time PCR. J Forensic Sci. 2010;55(1):25-33.

94. Lawyer FC, Stoffel S, Saiki RK, Myambo K, Drummond R, Gelfand DH. Isolation, characterization, and expression in Escherichia coli of the DNA polymerase gene from Thermus aquaticus. J Biol Chem. 1989 Apr 15;264(11):6427-37.

95. Butler J. Chapter 7 - DNA amplification (The Polymerase Chain Reaction). In: Fundamentals of Forensic DNA Typing. 1st ed. USA: Elsevier, Inc.; 2010. p. 125-46.

96. Henegariu O, Heerema NA, Dlouhy SR, Vance GH, Vogt PH. Multiplex PCR: critical parameters and step-by-step protocol. BioTechniques. 1997;23(3):504-11.

97. Walsh PS, Erlich HA, Higuchi R. Preferential PCR amplification of alleles: mechanisms and solutions. PCR Methods Appl. 1992;1(4):241-50.

98. Martin R. Gel electrophoresis: nucleic acids. Garland Pub; 1996.

99. Butler J. Chapter 10 - STR Genotyping and Data Interpretation. In: Fundamentals of Forensic DNA Typing. 1st ed. USA: Elsevier, Inc.; 2010. p. 205-27.

100. McCord B. Troubleshooting capillary electrophoresis systems. Profiles in DNA. 2003;6(2):10-2.

101. Butler JM, Buel E, Crivellente F, McCord BR. Forensic DNA typing by capillary electrophoresis using the ABI Prism 310 and 3100 genetic analyzers for STR analysis. Electrophoresis. 2004;25(10-11):1397-412.

102. Butler J. Chapter 9 - Fundamentals of DNA separation and detection. In: Fundamentals of Forensic DNA Typing. 1st ed. USA: Elsevier, Inc.; 2010. p. 175-203.

103. Butler J. Chapter 8 - Short Tandem Repeat Markers. In: Fundamentals of Forensic DNA Typing. 1st ed. USA: Elsevier, Inc.; 2010. p. 147-73.

104. Barber M, Parkin B. Sequence analysis and allelic designation of the two short tandem repeat loci D18S51 and D8S1179. Int J Legal Med. 1996;109(2):62-5.

105. Wenz H, Robertson JM, Menchen S, Oaks F, Demorest DM, Scheibler D, et al. High-precision genotyping by denaturing capillary electrophoresis. Genome Res. 1998 Jan;8(1):69-80.

106. Butler JM, Mccord B. Workshop on troubleshooting common laboratory problems. October 16th; http://strbase.nist.gov/training.htm; 2008.

107. SWGDAM Guidelines for the Collection and Serological Examination of Biological Evidence [Internet]; [updated January 15; 2015]. Available from: http://www.swgdam.org/publications.

108. Scientific Working Group on DNA Analysis Methods (SWGDAM) [Internet]; 2015. Available from: http://www.swgdam.org/.

109. Scientific Working Group on DNA analysis Methods (SWGDAM) [Internet]. Available from: www.swgdam.org.

110. Ballantyne J. Serology: overview. Encyclopedia of Forensic Sciences. 2000:1322-31.

111. Gaensslen RE. Sourcebook in forensic serology, immunology, and biochemistry. US Department of Justice, National Institute of Justice; 1983.

112. Hanson EK, Lubenow H, Ballantyne J. Identification of forensically relevant body fluids using a panel of differentially expressed microRNAs. Anal Biochem. 2009 APR 15;387(2):303-14.

113. Jones E. The identification of semen and other body fluids. Forensic Science Handbook, Prentice Hall, Upper Saddle River, NJ. 2005:329-82.

114. Cordier A, Budowle B, Gehrig C, Schmidt L, Hochmeister MN, Thali M, et al. Validation studies of an immunochromatographic 1-step test for the forensic identification of human blood. Journal of Forensic Science. 1999;44(3):597-602.

115. Horjan I, Barbaric L, Mrsic G. Applicability of three commercially available kits for forensic identification of blood stains. Journal of Forensic and Legal Medicine. 2016 2;38:101-5.

116. RSID™-Blood [Internet]. Available from: http://www.ifi-test.com/rsidtm-blood/.

117. Schweers BA, Old J, Boonlayangoor P, Reich KA. Developmental validation of a novel lateral flow strip test for rapid identification of human blood (Rapid Stain Identification™-Blood). Forensic Science International: Genetics. 2008;2(3):243-7.

118. Muro CK, Lednev IK. Identification of individual red blood cells by Raman microspectroscopy for forensic purposes: in search of a limit of detection. Analytical and Bioanalytical Chemistry. 2016:1-7.

119. Virkler K, Lednev IK. Analysis of body fluids for forensic purposes: from laboratory testing to non-destructive rapid confirmatory identification at a crime scene. Forensic Sci Int. 2009;188(1):1-17.

120. Saliva Identification Test [Internet]. Available from: https://www.abacusdiagnostics.com/saliva.htm.

121. Myers JR, Adkins WK. Comparison of modern techniques for saliva screening. J Forensic Sci. 2008;53(4):862-7.

122. Pang B, Cheung BK. Applicability of two commercially available kits for forensic identification of saliva stains. J Forensic Sci. 2008;53(5):1117-22.

123. Old J, Schweers B, Boonlayangoor PW, Reich K. Developmental validation studies of RSID^{TM} - Saliva lateral flow immunochromatographic strip test for the forensic detection of saliva. Hillside, IL: Independent Forensics; 2010.

124. Boward ES, Wilson SL. A comparison of ABAcard® p30 and RSID™-Semen test kits for forensic semen identification. Journal of forensic and legal medicine. 2013;20(8):1126-30.

125. Harbison SA, Fleming RI. Forensic body fluid identification: state of the art. Research and Reports in Forensic Medical Science. 2016;6:11-23.

126. Juusola J, Ballantyne J. Messenger RNA profiling: a prototype method to supplant conventional methods for body fluid identification. Forensic Sci Int. 2003;135(2):85-96.

127. van den Berge M, Carracedo A, Gomes I, Graham E, Haas C, Hjort B, Hoff-Olsen P, Maroñas O, Mevag B, Morling N, Niederstatter H. A collaborative European exercise on mRNA-based body fluid/skin typing and interpretation of DNA and RNA results. Forensic Science International: Genetics. 2014;10:40-8.

128. Čikoš Š, Bukovská A, Koppel J. Relative quantification of mRNA: comparison of methods currently used for real-time PCR data analysis. BMC molecular biology. 2007;8(1):113.

129. Hanson EK, Ballantyne J. Rapid and inexpensive body fluid identification by RNA profiling-based multiplex High Resolution Melt (HRM) analysis. F1000Res. 2013;2:281,281.v2. eCollection 2013.

130. Juusola J, Ballantyne J. Multiplex mRNA profiling for the identification of body fluids. Forensic Sci Int. 2005;152(1):1-12.

131. Roeder AD, Haas C. mRNA profiling using a minimum of five mRNA markers per body fluid and a novel scoring method for body fluid identification. Int J Legal Med. 2013;127(4):707-21.

132. Bauer M, Kraus A, Patzelt D. Detection of epithelial cells in dried blood stains by reverse transcriptase-polymerase chain reaction. Journal of Forensic Science. 1999;44(6):1232-6.

133. Juusola J, Ballantyne J. mRNA profiling for body fluid identification by multiplex quantitative RT-PCR. J Forensic Sci. 2007;52(6):1252-62.

134. Nussbaumer C, Gharehbaghi-Schnell E, Korschineck I. Messenger RNA profiling: A novel method for body fluid identification by Real-Time PCR. Forensic Sci Int. 2006;157(2-3):181-6.

135. Park S, Park S, Kim J, Kang T, Park J, Woo K, et al. Genome-wide mRNA profiling and multiplex quantitative RT-PCR for forensic body fluid identification. Forensic Science International: Genetics. 2013;7(1):143-50.

136. An JH, Shin K, Yang WI, Lee HY. Body fluid identification in forensics. Bmb Reports. 2012;45(10):545-53.

137. Bauer M, Patzelt D. A method for simultaneous RNA and DNA isolation from dried blood and semen stains. Forensic Sci Int. 2003;136(1):76-8.

138. Alvarez M, Juusola J, Ballantyne J. An mRNA and DNA co-isolation method for forensic casework samples. Anal Biochem. 2004;335(2):289-98.

139. Probst J, Brechtel S, Scheel B, Hoerr I, Jung G, Rammensee H, et al. Characterization of the ribonuclease activity on the skin surface. Genetic vaccines and therapy. 2006;4(1):4.

140. Haas C, Klesser B, Maake C, Bär W, Kratzer A. mRNA profiling for body fluid identification by reverse transcription endpoint PCR and realtime PCR. Forensic Science International: Genetics. 2009;3(2):80-8.

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

142. Courts C, Madea B. Micro-RNA - A potential for forensic science? Forensic Sci Int. 2010;203(1-3):106-11.

143. Courts C, Madea B. Specific Micro-RNA Signatures for the Detection of Saliva and Blood in Forensic Body-fluid Identification. J Forensic Sci. 2011;56(6):1464-70.

144. Kurdyukov S, Bullock M. DNA methylation analysis: choosing the right method. Biology. 2016;5(1):3.

145. Rauch T, Pfeifer G. Chapter 9 - Methods for assessing genome-wide DNA methylation. In: Collins L, Schönfeld B, Chen X, editors. Handbook of Epigenetics: The New Molecular and Medical Genetics. Elsevier, New York; 2011. p. 135.

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

147. Frumkin D, Wasserstrom A, Budowle B, Davidson A. DNA methylation-based forensic tissue identification. Forensic Science International-Genetics. 2011;5(5):517-24.

148. Roach H, Hashimoto K. PCR-based methods to determine DNA methylation status at specific CpG sites using methylation-sensitive restriction enzymes In: Hughes S, Moody A, editors. PCR : Methods Express. Scion Publishing; 2007. p. 279.

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

150. Shapiro R, Servis RE, Welcher M. Reactions of uracil and cytosine derivatives with sodium bisulfite. J Am Chem Soc. 1970;92(2):422-4.

151. Hayatsu H, Wataya Y, Kai K. Addition of sodium bisulfite to uracil and to cytosine. J Am Chem Soc. 1970;92(3):724-6.

152. Tost J, Gut IG. DNA methylation analysis by pyrosequencing. Nature Protocols. 2007;2(9):2265-75.

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

154. Antunes J, Balamurugan K, Duncan G, McCord B. Tissue-Specific DNA Methylation Patterns in Forensic Samples Detected by Pyrosequencing®. In: Lehmann U, Tost J, editors. Methods in Molecular Biology. Springer New York; 2015. p. 397-409.

155. Pyromark Q24 Software User Guide [Internet].: Qiagen; 2017. Available from: https://www.qiagen.com/us/resources/resourcedetail?id=a06f1196-2bd0-40af-87d5-45c80c285b48&lang=en.

156. 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.

157. 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).

158. Balic M, Pichler M, Strutz J, Heitzer E, Ausch C, Samonigg H, et al. High Quality Assessment of DNA Methylation in Archival Tissues from Colorectal Cancer Patients Using Quantitative High-Resolution Melting Analysis. Journal of Molecular Diagnostics. 2009;11(2):102-8.

159. Nicklas JA, Noreault-Conti T, Buel E. Development of a Real-Time Method to Detect DNA Degradation in Forensic Samples*. J Forensic Sci. 2012;57(2):466-71.

160. Interpreting melt curves: An indicator, not a diagnosis [Internet].: Integrated DNA Technologies. Available from: http://www.idtdna.com/pages/decoded/decoded-articles/core-concepts/decoded/2014/01/20/interpreting-melt-curves-an-indicator-not-a-diagnosis.

161. Luo T, Jiang L, Sun W, Fu G, Mei J, Gao Q. Multiplex real-time PCR melting curve assay to detect drug-resistant mutations of Mycobacterium tuberculosis. J Clin Microbiol. 2011;49(9):3132.

162. Vossen RHAM, Aten E, Roos A, den Dunnen JT. High-Resolution Melting Analysis (HRMA)-More than just sequence variant screening. Hum Mutat. 2009;30(6):860-6.

163. Wojdacz TK. Methylation-sensitive high-resolution melting in the context of legislative requirements for validation of analytical procedures for diagnostic applications. Expert Review of Molecular Diagnostics. 2012;12(1):39-47.

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

165. DNA Fragment Analysis by Capillary Electrophoresis [Internet].: Thermo Fisher Scientific, inc.; 2014. Available from: https://www.thermofisher.com/content/dam/LifeTech/global/Forms/PDF/fragment-analysis-chemistry-guide.pdf.

166. Lee HY, An JH, Jung S, Oh YN, Lee EY, Choi A, et al. Genome-wide methylation profiling and a multiplex construction for the identification of body fluids using epigenetic markers. Forensic Science International: Genetics. 2015 7;17:17-24.

167. Lee HY, Jung S, Lee EH, Yang WI, Shin K. DNA methylation profiling for a confirmatory test for blood, saliva, semen, vaginal fluid and menstrual blood. Forensic Science International: Genetics. 2016;24:75-82.

168. Frumkin D, Wasserstrom A, Davidson A, Grafit A. Authentication of forensic DNA samples. Forensic Science International-Genetics. 2010;4(2):95-103.

169. Zubakov,D.,Kayser,M. Forensic DNA application. In: Primorac,D.,Schnfield,M., editor. Forensic DNA applications: An Interdisciplinary Perspective. Boca Raton: CRC Press; 2014. p. 389,390, 404.

170. Wang Z, Zhang J, Luo H, Ye Y, Yan J, Hou Y. Screening and confirmation of microRNA markers for forensic body fluid identification. Forensic Sci Int -Genet. 2013;7(1):116-23.

171. Lindenbergh A, de Pagter M, Ramdayal G, Visser M, Zubakov D, Kayser M, et al. A multiplex (m)RNA-profiling system for the forensic identification of body fluids and contact traces. Forensic Science International-Genetics. 2012;6(5):565-77.

172. Hanson E, Ballantyne J. Rapid and inexpensive body fluid identification by RNA profiling-based multiplex High Resolution Melt (HRM) analysis. F1000 Research. 2014;2:281.

173. Choi A, Shin K, Yang WI, Lee HY. Body fluid identification by integrated analysis of DNA methylation and body fluid-specific microbial DNA. Int J Legal Med. 2014 JAN;128(1):33-41.

174. Antunes,J.,Silva,DSBS.,Kuppareddi,B.,Duncan,G.,Alho,CS.,McCord,B. High Resolution Melt analysis of DNA methylation to discriminate semen in biological stains. Analytical Biochemistry. 2016;494:40-45.

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

176. Park J, Kwon O, Kim JH, Yoo H, Lee H, Woo K, et al. Identification of body fluid-specific DNA methylation markers for use in forensic science. Forensic Science International-Genetics. 2014;13:147-53.

177. Frascione N, Thorogate R, Daniel B, Jickells S. Detection and identification of body fluid stains using antibody-nanoparticle conjugates. Analyst. 2012;137(2):508-12.

178. Miranda TB, Jones PA. DNA methylation: the nuts and bolts of repression. J Cell Physiol. 2007;213(2):384-90.

179. Simmons M, Snustad D. Fundamentos de Genética. Guanabara Koogan, Rio de Janeiro. 2001:332-6.

180. Wang J, McCord B. The application of magnetic bead hybridization for the recovery and STR amplification of degraded and inhibited forensic DNA. Electrophoresis. 2011;32(13):1631-8.

181. Chaitanya L, van Oven M, Weiler N, Harteveld J, Wirken L, Sijen T, et al. Developmental validation of mitochondrial DNA genotyping assays for adept matrilineal inference of biogeographic ancestry at a continental level. Forensic Science International: Genetics. 2014;11:39-51.

182. Pellacani D, Kestoras D, Droop A, Frame F, Berry P, Lawrence M, et al. DNA hypermethylation in prostate cancer is a consequence of aberrant epithelial differentiation and hyperproliferation. Cell Death & Differentiation. 2014;21(5):761-73.

183. 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;39(11):e77.

184. Sijen T. Molecular approaches for forensic cell type identification: On mRNA, miRNA, DNA methylation and microbial markers. Forensic Science International: Genetics. 2015;18:21-32.

185. Haas C, Hanson E, Bär W, Banemann R, Bento A, Berti A, et al. mRNA profiling for the identification of blood—results of a collaborative EDNAP exercise. Forensic Science International: Genetics. 2011;5(1):21-6.

186. Haas C, Hanson E, Anjos M, Bär W, Banemann R, Berti A, et al. RNA/DNA co-analysis from blood stains—results of a second collaborative EDNAP exercise. Forensic Science International: Genetics. 2012;6(1):70-80.

187. Haas C, Hanson E, Anjos M, Banemann R, Berti A, Borges E, et al. RNA/DNA co-analysis from human saliva and semen stains–results of a third collaborative EDNAP exercise. Forensic Science International: Genetics. 2013;7(2):230-9.

188. Haas C, Hanson E, Banemann R, Bento A, Berti A, Carracedo Á, et al. RNA/DNA co-analysis from human skin and contact traces–results of a sixth collaborative EDNAP exercise. Forensic Science International: Genetics. 2015;16:139-47.

189. Haas C, Hanson E, Anjos M, Ballantyne K, Banemann R, Bhoelai B, et al. RNA/DNA co-analysis from human menstrual blood and vaginal secretion stains: results of a fourth and fifth collaborative EDNAP exercise. Forensic Science International: Genetics. 2014;8(1):203-12.

190. Forat S, Huettel B, Reinhardt R, Fimmers R, Haidl G, Denschlag D, et al. Correction: Methylation Markers for the Identification of Body Fluids and Tissues from Forensic Trace Evidence. PloS one. 2016;11(5):e0156472.

191. Balamurugan K, Bombardi R, Duncan G, McCord B. Identification of spermatozoa by tissue-specific differential DNA methylation using bisulfite modification and pyrosequencing. Electrophoresis. 2014;35(21-22):3079-86.

192. Song F, Mahmood S, Ghosh S, Liang P, Smiraglia DJ, Nagase H, et al. Tissue specific differentially methylated regions (TDMR): Changes in DNA methylation during development. Genomics. 2009;93(2):130-9.

193. Sepulveda AR, Jones D, Ogino S, Samowitz W, Gulley ML, Edwards R, et al. CpG Methylation Analysis‚ÄîCurrent Status of Clinical Assays and Potential Applications in Molecular Diagnostics: A Report of the Association for Molecular Pathology. The Journal of Molecular Diagnostics : JMD. 2009 04/09;11(4):266-78.

194. Wittwer CT. High-resolution DNA melting analysis: advancements and limitations. Hum Mutat. 2009;30(6):857-9.

195. Erali M, Voelkerding KV, Wittwer CT. High resolution melting applications for clinical laboratory medicine. Exp Mol Pathol. 2008 8;85(1):50-8.

196. Kristensen LS, Hansen LL. PCR-Based Methods for Detecting Single-Locus DNA Methylation Biomarkers in Cancer Diagnostics, Prognostics, and Response to Treatment. Clinical Chemistry. 2009 August 01;55(8):1471-83.

197. Kristensen LS, Mikeska T, Krypuy M, Dobrovic A. Sensitive Melting Analysis after Real Time- Methylation Specific PCR (SMART-MSP): high-throughput and probe-free quantitative DNA methylation detection. Nucleic Acids Res. 2008 02/19;36(7):e42-.

198. Tse MY, Ashbury JE, Zwingerman N, King WD, Taylor SAM, Pang SC. A refined, rapid and reproducible high resolution melt (HRM)-based method suitable for quantification of global LINE-1 repetitive element methylation. BMC Research Notes. 2011 12/28;4:565-.

199. Malentacchi F, Forni G, Vinci S, Orlando C. Quantitative evaluation of DNA methylation by optimization of a differential-high resolution melt analysis protocol. Nucleic Acids Res. 2009 04/27;37(12):e86-.

200. Li L, Dahiya R. MethPrimer: designing primers for methylation PCRs. . Bioinformatics. 2002;18(11):1427-31.

201. Silva DS, Antunes J, Balamurugan K, Duncan G, Alho CS, McCord B. Developmental validation studies of epigenetic DNA methylation markers for the detection of blood, semen and saliva samples. Forensic Science International: Genetics. 2016;23:55-63.

202. Lövkvist C, Dodd IB, Sneppen K, Haerter JO. DNA methylation in human epigenomes depends on local topology of CpG sites. Nucleic Acids Res. 2016;44(11):5123-32.

203. Cocozza S, Akhtar MM, Miele G, Monticelli A. CpG islands undermethylation in human genomic regions under selective pressure. PloS one. 2011;6(8):e23156.

204. Dedeurwaerder S, Defrance M, Bizet M, Calonne E, Bontempi G, Fuks F. A comprehensive overview of Infinium HumanMethylation450 data processing. Briefings in bioinformatics. 2013;15(6):929-41.

205. Du P, Zhang X, Huang C, Jafari N, Kibbe W, Hou L, et al. Comparison of Beta-value and M-value methods for quantifying methylation levels by microarray analysis. BMC Bioinformatics. 2010;11:587.

206. Kim H. Analysis of variance (ANOVA) comparing means of more than two groups. Restorative dentistry & endodontics. 2014;39(1):74-7.

207. Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, et al. The human genome browser at UCSC. Genome Res. 2002 Jun;12(6):996-1006.

208. Aranyi T., Varadi A., Simon I. and Tusnady G.E. The BiSearch Web Server. BMC Bioinformatics. 2006:431.

209. Tusnady GE, Simon I, Varadi A, Aranyi T. BiSearch: primer-design and search tool for PCR on bisulfite-treated genomes. Nucleic Acids Res. 2005;33.

210. Wojdacz Tomasz, Hansen Lise, Dobrovic Alexander. A new approach to primer design for the control of PCR bias in methylation studies. BMC Research Notes. 2008;1(1):54.

211. Wojdacz TK, Borgbo T, Hansen LL. Primer design versus PCR bias in methylation independent PCR amplifications. Epigenetics. 2009 MAY 16 2009;4(4):231-4.

212. Epitect HRM PCR kit SDS [Internet].: Qiagen. Available from: https://sds.qiagen.com/ehswww/QIAGENwww/result/result_htmllist.jsp?P_LANGU=E&P_SYS=4&P_SSN=42571&C001=MSDS&C700=KIT&C007=PROD&C100=QIA&C002=US&C013=59445.

213. HotStarTaq PCR Handbook [Internet].; 2017. Available from: https://www.qiagen.com/us/resources/resourcedetail?id=53e71936-cf64-4a9b-a9d2-fccee4868aa7&lang=en.

214. Waite LL, Weaver B, Day K, Li X, Roberts K, Gibson AW, et al. Estimation of Cell-Type Composition Including T and B Cell Subtypes for Whole Blood Methylation Microarray Data. Front Genet. 2016 Feb 18;7:23.

215. Proctor GB. The physiology of salivary secretion. Periodontol 2000. 2016;70(1):11-25.

216. Kelly K. Chapter 179 : Tests on Vaginal Discharge. In: Walker H, Hall W, Hurst J, editors. Clinical methods: the history, physical, and laboratory examinations. American Medical Association; 1990.

217. Eipel M, Mayer F, Arent T, Ferreira MR, Birkhofer C, Gerstenmaier U, et al. Epigenetic age predictions based on buccal swabs are more precise in combination with cell type-specific DNA methylation signatures. Aging (Albany NY). 2016 May;8(5):1034-48.

218. Belkaid Y, Hand TW. Role of the microbiota in immunity and inflammation. Cell. 2014;157(1):121-41.

219. Jensen MA, Fukushima M, Davis RW. DMSO and betaine greatly improve amplification of GC-rich constructs in de novo synthesis. PLoS One. 2010;5(6):e11024.

220. Farell EM, Alexandre G. Bovine serum albumin further enhances the effects of organic solvents on increased yield of polymerase chain reaction of GC-rich templates. BMC research notes. 2012;5(1):257.

221. Zhou L, Errigo RJ, Lu H, Poritz MA, Seipp MT, Wittwer CT. Snapback primer genotyping with saturating DNA dye and melting analysis. Clin Chem. 2008 Oct;54(10):1648-56.

222. Simpson JT, Workman RE, Zuzarte P, David M, Dursi L, Timp W. Detecting DNA cytosine methylation using nanopore sequencing. nature methods. 2017;14(4):407-10.

APPENDICES

Appendix 1

DNA extraction using biorobot EZ1 Advanced (Qiagen, CA)

Materials:

EZ1 DNA Investigator kit Qiagen Cat#952034
Buffer MTL Qiagen Cat#19112

Protocol:

1. With sterile scissors cut the swab into an appropriate tube. For liquid samples (semen) use 150 μL semen per extraction

2. Add 490 μL buffer G2 and 10 μL of Proteinase K

3. Vortex and incubate for more than 2 hours at 56 ºC. During incubation vortex the tube occasionally to improve cell lysis (every 30-45 min)

4. Using a spin basket remove the liquid from the swab

5. All the 500 μL of lysate should be in the 2 mL sample tubes provided with the DNA Investigator kit (Qiagen, CA)

6. Add 400 μL of Buffer MTL to each sample tube with lysate. Vortex

7. Load the sample tubes in the proper location of the EZ1 Advanced BioRobot (follow instructions manual)

8. Load all tubes and tips to the machine as well as the reagent cartridges

9. Run the Large Volume protocol with 50 μL final volume elution with TE buffer

10. Run the program. Remove the samples and freeze at -20 ºC for large periods of time or at 4 ºC if quantification is performed on the same day

DNA Quantification using PicoGreen dye (Blood samples only)

Materials:

PicoGreenDNA reagent Life Technologies Cat#P7581
OR
SybrGreen 10,000X in DMSO Invitrogen Cat# S7563

Protocol:

1. Prepare DNA standards from 50 ng/µL down to 0.5 ng/µL concentration by serial dilutions of the stock.

2. Add 30 µL of molecular grade Millipore water to enough 0.5 mL tubes for all standards, No Template Control and the samples

3. Add 2 µL of each standard or sample to the appropriate tubes and 2 µL water to the control. Make a 1:10 and/or 1:2 dilution of each unknown sample to make sure the calculated value is within the range for the standard curve.

4. In the dark add 32 µL of a 1:200 dilution of PicoGreen solution in DMSO to each tube

5. Vortex, quick spin to remove drops from the inside of lids and incubate for at least 5 minutes in the dark

6. Transfer 20 µL aliquots of each tube to 3 different 0.2 mL (PCR) tubes

7. Load tubes in the real-time PCR machine with the standard of highest concentration on position 1 (per manual instruction)

8. Run the protocol with the settings for 20 μL of sample and hold for 2 min. at

   50 ºC followed by 10 cycles of 15 seconds at 50 ºC with fluorescence reading

   after each cycle on FAM (470 nm source and 510 nm detector at gain 5)

9. Determine concentration of the sample based on the standard curve

ALTERNATIVELY: If PicoGreen is not available, SybrGreen can be used if diluted in
the following manner:
     Working stock is 1:200 dilution in DMSO:
       2 μL SybrGreen + 398 μL DMSO
     Quantification stock:
       2 μL working stock SybrGreen + 48 μL TE buffer 1X

DNA Quantification using Alu markers (samples from all body fluids)

Materials:

  Buffer 10X with 15 mM MgCl2 Applied Biosystems
  MgCl2 25 mM Applied Biosystems
  Alu primers at 100 uM
  BSA Sigma tube at 20 mg/mL concentration
  dNTPs mix at 2.5 mM each
  Triton X cat# T9284-100 from Sigma, 10% dilution
  Sybr Green Life Technologies 10,000. Dilute 1:200 in DMSO
  RampTaq Denville 5U/uL

Protocol:

1. make the following master mix for the number of reactions you will run, plus 1:

Water	15.1 μL
Buffer 10X	2.3 μL
dNTPs	1.9 μL
MgCl2	1.4 μL
Alu Forward primer	0.2 μL
Alu Reverse primer	0.2 μL
Triton X	0.2 μL
BSA	0.2 μL
Sybr Green	0.2 μL
Ramp Taq	0.4 μL

2. Take out 18 uL and add 2 uL of your DNA/ standards/ water (for NTC)

Bisulfite Conversion

Materials:

Epitect® Fast Bisulfite kit (200) Qiagen Cat#59826

Protocol:

1. Samples and all reagents need to be at room temperature

2. According to each sample's concentration pipette enough volume to have 400 ng of DNA to a 0.2 mL (PCR) tube

3. Add molecular grade water up to 40 μL of volume

4. Add 85 μL of bisulfite solution

5. Add 15 μL of DNA protect buffer

6. Vortex and quick spin on the centrifuge to remove drops from inside lid

7. Place tubes in a thermal cycler and run the program 5 min at 95 ºC, 20 min at 60 ºC, 5 min at 95 ºC, 20 min at 60 ºC and hold indefinite at 20 ºC

8. Briefly centrifuge the tubes and transfer the solution to 1.5 mL tubes

9. Mix carrier RNA solution with buffer BL at the ratio of 3.5 μL carrier RNA to 350 μL Buffer BL. Multiply each volume for the number of samples

10. Add 310 μL of the mixture buffer BL with carrier RNA to each tube. Vortex and quick spin in the centrifuge

11. Add 250 μL of ethanol (96-100%) to each sample. Vortex for 15 seconds and quick spin in the centrifuge

12. Transfer the content of each tube to labeled MinElute® DNA spin columns

13. Centrifuge the columns for 1 minute at 12000 x g or maximum speed. Discard the liquid on bottom of tubes (flow through) and place the columns back on the collection tubes

14. Invert Buffer BW several times before use and add 500 μL of Buffer BW to each spin column. Centrifuge for 1 minute at 12000 x g or maximum speed, discard the flow through and place the columns back on the collection tubes

15. Add 500 μL of Buffer BD to each tube; close the tubes and incubate for 15 minutes at room temperature; make sure Buffer BD is not open for longer than required

16. Centrifuge the columns for 1 minute at 12000 x g or maximum speed. Discard the flow through and place the spin columns back in the collection tubes

17. Add 500 μL of Buffer BW to each spin column and centrifuge for 1 minute at 12000 x g or maximum speed. Remove flow through and place spin columns back in the collection tubes

18. Repeat step 17

19. Add 250 μL of ethanol (96-100%) to each spin column and centrifuge for 1 minute at 12000 x g or maximum speed

20. Place the spin columns in new 2 mL collection tubes. Centrifuge for 1 minute at 12000 x g or maximum speed to remove any remaining liquid.

21. Incubate the spin columns for 5 minutes at 60 ºC with the lids open to evaporate any residual ethanol

22. Place the spin columns into clean labeled 1.5 mL tubes (not provided with the kit)

23. Add 20 µL of Buffer EB into the center of the membrane on the spin column.

Incubate for 1 minute at room temperature and centrifuge for 1 minute at

12000 x g or maximum speed.

24. Repeat step 23

25. Store the converted DNA at -20 ºC

NOTE: the concentration of DNA on step 2 and the volume of elution buffer on step 23 were the ones used to screen new genome locations.

Polymerase Chain Reaction for Pyrosequencing

Materials:
PyroMark® PCR Kit (200) Qiagen Cat#978703
QiagenPyroMark®CpG Assays Qiagen
BCAS4: Cat#PMC0002408
FGF7: Cat#PMC0004067
ZC3H12D: Cat#PM00124145
PFN3A: Cat# PMC0084567

Protocol:

1. Thaw the Pyromark PCR Master Mix, the Coral Load Concentrate, Molecular

grade water aliquot, the aliquot of the Pyromark assay primers and the

bisulfite-modified DNA to be amplified

2. To determine the number of reactions, add the number of DNA samples with

2 for No Template Controls (NTC, no DNA is added, just water) plus one (to

account for pipetting loss)

3. Prepare a master mix by adding 12.5 µL of Pyromark® PCR Master Mix, 2.5

µL of Coral Load Concentrate, 2.5 µL of Pyromark assay primers, 5.5 µL of

water per reaction. To determine the volumes to add, multiply each volume

given per the number of reactions determined in step 2

4. Aliquot 23 μL of the mix prepared in step 3 to the appropriate number of 0.2 mL (PCR) tubes

5. Add 2 μL of molecular grade water to the NTC tubes and 2 μL of DNA to the remaining tubes.

6. Centrifuge the tubes and load them in a thermal cycler. Program the thermal cycler for an initial step of 15 minutes at 95 ºC followed by 45 cycles of 30 seconds at 94 ºC, 30 seconds at 55 ºC and 30 seconds at 72 ºC, with a 10 minutes final extension at 72 ºC and a last hold at 4 ºC for infinite time. For custom primers the annealing temperature is usually set to Tm - 5ºC.

7. When the PCR is complete, remove the tubes and store at – 20 ºC

Confirmation of PCR products by running an Agarose gel

Material:

Agarose molecular biology grade
TBE buffer 1X containing 10.8 g Tris base, 5.5g Boric acid, 0.93g disodium ethylenediaminetetraacetate ($Na_2C_{10}H_{14}O_8N_2.H_2O$) all dissolved in 1 liter of deionized water.
Ethidium bromide
Low molecular weight DNA ladder (New England Biolabs, cat#N3233S)

Protocol:

1. Prepare a 2% agarose gel in 1X Tris-Borate-EDTA buffer

2. Cool the gel and add ethidium bromide before pouring the gel in the tray

3. Let the gel solidify for at least 20 minutes

4. Take the comb out and place the tray with the gel on the electrophoresis apparatus

5. Load the ladder with the loading buffer and 2-5 μL each PCR product to the appropriate wells on the gel. The PCR product doesn't need loading buffer since the Coral Load Concentrate has that function

6. Turn the electrophoresis power ON at a proper current for the size of the gel

7. Take a picture of the gel under an ultraviolet light and check for the bands around 200 bp size and no bands for the NTC

Pyrosequencing on the Q24 Advanced system

Materials:
PyroMark Q24 Advanced System Qiagen Cat#9002270
PyroMark Q24 Advanced Software Qiagen Cat#9022779
PyroMark Q24 Advanced CpG Reagents (4 X 24) Qiagen Cat#970922
PyroMark Denaturation Solution (500 mL) Qiagen Cat#979007
PyroMark Wash Buffer (concentrated, 200 mL) Qiagen Cat# 979008
PyroMark Q24 Cartridge (3) Qiagen Cat#979202
PyroMark Q24 Plate (100) Qiagen Cat#979201
PyroMark Q24 Plate Holder Qiagen Cat#979205
PryoMark Q24 Vacuum Prep Troughs Qiagen Cat#979206
PyroMark Q24 Vacuum Workstation (110V) Qiagen Cat#9001516
PyroMark Vacuum Prep Filter Probe (100) Qiagen Cat#979010
Streptavidin Sepharose High Performance Beads (5 mL) GE Healthcare   Cat#17-5113-01

Protocol:

1. Turn the Pyrosequencer ON at least 30 minutes before using

2. Thaw the PCR products,an aliquot of biotinylated PCR primers (this aliquot is only used for pyrosequencing and should never be used to amplify to avoid contamination from PCR products to pre-amplification DNA) and the sequencing primer. Equilibrate at room temperature the diluted washing

216

buffer, the denaturation buffer, the annealing solution, the binding solution, the streptavidin-coated beads and the Pyromark Gold reagents. Place the metal tray for the Q24 plate on a heating plate set at 80ºC

3. Prepare the beads mixture by adding 40μL of binding buffer, 29μL of molecular grade water, and 1 μL of beads per reaction. Multiply each volume per the number of samples plus one for pipetting loss. Add 70 μL of this mixture to 22 wells of a 24-well PCR plate. Leave the last two wells empty for negative controls and do not put beads mixture in these wells.

4. Add 10 μL of PCR product in the corresponding well of the 24-well PCR plate. Close the wells with a strip cap and mix at 1400 rpm for at least 15 minutes. Remove from mixing only when ready to use the vacuum station

5. Add 20 μL of 1x sequencing primer in annealing buffer to 22 wells on the Pyromark® Q24- sequencing plate. The last two wells are reserved for the primer controls. Add the sequencing primer in one well, and add a mixture of biotinylated PCR primer (2 μL) and sequencing primer (20 μL) to the other well

6. Place the Pyromark® Q-24 plate on the vacuum station and fill all the trays of the station with the corresponding solutions

7. Prepare the template for the plate on the Pyrosequencing software. Check the volumes to add to the cartridge for each nucleotide, enzyme mix and substrate mix. Add the volumes to the cartridge and place it on the Q24 pyrosequencer

8. Turn the vacuum switch in the suction probe to 'ON' and place it on the prime tray with distilled water aspirating approximately 70 mL

9. Remove the 24-well PCR plate from agitation, take the lids off and place in the vacuum station. Place the suction probe into the PCR plate wells with the vacuum position 'ON' carefully capturing all the solution from the wells for about 15 seconds

10. Wash the beads/samples with 70% Ethanol for 5 seconds

11. Continue to the tray with Denaturing solution, flush for 5 seconds.

12. Aspirate for 10 seconds on the Washing Buffer tray and tilt the suction probe to beyond vertical 90º for a few seconds

13. Place the vacuum suction probe on top of the Pyromark® Q24 plate without touching the liquid. Turn off vacuum and then lower the handle on the Pyromark® Q24- sequencing plate. Shake the suction probes gently from side to side for 30 seconds to 1 minute to release the beads on the sequencing plate containing sequencing primers

14. Rinse the vacuum suction probe on the rinse tray containing distilled water by agitating for 10 seconds

15. Aspirate distilled water on the prime tray with aspirating 70 mL of water

16. Move the suction probe beyond 90º vertical for a few seconds and disconnect the vaccum

17. Place the sequencing plate on the metal tray and heat at 80 ºC for 5 minutes

18. Within a 30 seconds interval place the sequencing plate on the Q24 Pyrosequencer, close the lid and start the run

19. When the run is complete discard the Q24 plate, and the unused contents of the cartridge by inverting it and rinse the cartridge with deionized water. Fill

up each well of the cartridge with water and by applying pressure on the top

with your fingers force the water to pass the small needles on the bottom.

Repeat this for at least 3 times each well. Let it dry at room temperature

20. Remove and rinse all trays with deionized water from the vacuum station and

    dry at room temperature

Data analysis on the Q24 Software

1. The Pyromark Q24 software will perform the methylation percent analysis at

   each variable CpG sites and the data is displayed as a pyrogram

2. If any yellow or red warning is displayed due to a height variation on a

   homopolymer, deselect that peak as a reference peak

3. Get the percent methylation value for each variable position and make an

   excel spreadsheet with values for all tissues tested for each primer set

High-Resolution Melt Amplification and Analysis

NOTE: The DNA is extracted and bisulfite modified as described above. This step is
performed instead of the Pyromark® PCR amplification and pyrosequencing

Reagents:

- Hot Star Taq Plus (Qiagen, CA) containing Hot Start Polymerase and 10X Buffer
  (without Coral Load)
- Solution of 5 mM each dNTP
- Primers forward and reverse at 10 μM concentration
- Eva Green 20X dye solution in water (Biotium, CA)
- BSA solution at 10 mg/mL (1:2 dilution from Sigma Aldrich stock)
- DMSO 100%

- Water, molecular grade

Procedure:

1) Thaw all reagents at room temperature, except for the polymerase

2) Prepare a master mix for the number of samples, plus one to account for pipetting.

   The master mix for one reaction is as follows:

   a. 2 µL of 10X buffer (Qiagen, CA)

   b. 0.8 µL dNTP (5 mM)

   c. 1.5 µL forward primer (10 µM)*

   d. 1.5 µL reverse primer (10 µM)*

   e. 1 µL Eva Green dye (20X)

   f. 0.4 µL of BSA (10 m/mL)

   g. 0.5 µL DMSO (100%)

   h. 0.5 µL Hot StartTaq Plus polymerase

   i. 9.8 µL water

   * primer ZC3H12D is amplified with only 1µL of each forward and reverse.
   In that case the amount of water is adjusted to 10.8 µL

3) Pipette 18 µL of the master mix prepared to each individual 0.2 mL tube

4) Add 2 µL of bisulfite-modified DNA

5) Close the tubes, insert in the RotorGene and start the protocol
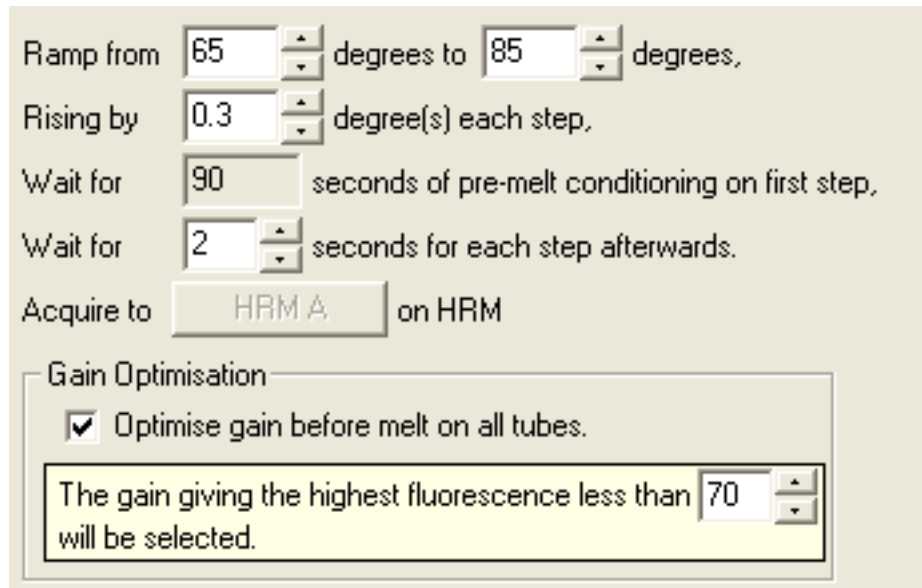
RotorGene protocol:

PCR cycle:
- Set a Hold step that runs for 5 minutes at 95 ºC (this step will activate the HotStart Polymerase)

- Set the instrument to run 45 cycles. If needed more cycles can be added to make sure that all the amplification curves reach a plateau.
  o 10 seconds at 95 ºC
  o 30 seconds at <<annealing temperature of the primer¥>>
  o 10 seconds at 72 ºC, acquiring on green and HRM channels

  ¥ - Table with the list of annealing temperatures for the primers ran with 10 samples each for this report

Marker	Annealing Temperature (ºC)
INPP5D200_TM60	55
ZC3H12D	55
VE_8	50
VE_15 (saliva marker)	50
INPP5D1500	45
INPP5D200_2	55

- Set a Melt step according to the settings:

Ramp from 65 degrees to 85 degrees,

Rising by 0.3 degree(s) each step,

Wait for 90 seconds of pre-melt conditioning on first step,

Wait for 2 seconds for each step afterwards.

Acquire to HRM A on HRM

Gain Optimisation
☑ Optimise gain before melt on all tubes.

The gain giving the highest fluorescence less than 70 will be selected.

VITA

JOANA ANTUNES

Born, Covilhã, Portugal

1999-2005	B.A., Biochemistry Universidade da Beira Interior Covilhã, Portugal
2007- 2009	M.S., Biochemistry Universidade da Beira Interior, Covilhã, Portugal & Anacor Pharmaceuticals, Palo Alto, CA
2009-2011	Research Technician Instituto Gulbenkian de Ciência Oeiras, Portugal
2011 -2017	Doctoral Candidate Florida International University Miami, Florida
2013-2015	National Institute of Justice Grant Award 2012-DN-BX-K018
2013	Third place oral presentation Graduate Student scholarly forum Florida International University, Miami, Florida
2013	National Institute for Standards and Technology travel grant to present at Annual Conference of International Association for Identification Providence, Rhode Island
2015	First place oral presentation Graduate Student scholarly forum Florida International University, Miami, Florida
2015	First place poster presentation Statewide Graduate Student Research Symposium Orlando, Florida
2016-2017	National Institute of Justice Graduate Fellowship Award 2015-R2-CX-0012

PUBLICATIONS

Antunes J, Silva DSBS, Balamurugan K, Duncan G, Alho CS, McCord B. (2016). *Forensic discrimination of vaginal epithelia by DNA methylation analysis using pyrosequencing*. Electrophoresis, 37(21): 2751-2758

Antunes J, Silva DSBS, Balamurugan K, Duncan G, Alho CS, McCord B. (2016) *High Resolution Melt analysis of DNA methylation to discriminate semen in biological stains*. Analytical Biochemistry, 494:40-5

Antunes J, Balamurugan K, Duncan G, McCord B. (2015) *Tissue-specific DNA methylation patterns in forensic samples detected by Pyrosequencing* Methods in Molecular Biology, 1315:397-409

Freund YR, Akama T, Alley MRK, Antunes J, Dong C, Jarnagin K, Kimura R, Nieman JA, Maples KR, Plattner JJ, Rock F, Sharma R, Singh R, Sanders V, Zhou Y. (2012) *Boron-based phosphodiesterase inhibitors show novel binding of boron to PDE4 bimetal center*. FEBS Letters, 586: 3410-3414

Hussain A, Antunes J, Silva DS, Alho C, 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

Jung S, Cho S, Antunes J, Gomes I, Uchimoto ML, Oh YN, Di Giacomo L, Schneider PM, Park MS, van der Meer D, Williams G, McCord B, Ahn H, Choi DH, Lee Y, Lee SD, Lee HY. (2016) *A collaborative exercise on DNA methylation-based body fluid typing*. Electrophoresis, 37:2759-2766

Silva DSBS, Antunes J, Balamurugan K, Duncan G, Alho CS, McCord B.(2016) *Developmental validation studies of epigenetic DNA methylation markers for the detection of blood, semen and saliva samples*. Forensic Science International: Genetics, 23:55-63

Silva DSBS, Antunes J, Balamurugan K, Duncan G, Alho CS, McCord B.(2015) *Evaluation of DNA methylation markers and their potential to predict human aging*. Electrophoresis, 36(15):1775-80