11-9-2020

Development of a DNA Methylation Multiplex Assay for Body Fluid Identification and Age Determination

Quentin Gauthier
qgaut001@fiu.edu

FLORIDA INTERNATIONAL UNIVERSITY

Miami, Florida

DEVELOPMENT OF A DNA METHYLATION MULTIPLEX ASSAY FOR BODY FLUID IDENTIFICATION AND AGE DETERMINATION

A dissertation submitted in partial fulfillment of

the requirements for the degree of

DOCTOR OF PHILOSOPHY

in

CHEMISTRY

by

Quentin Thibault Gauthier

2020

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

This dissertation, written by Quentin Thibault Gauthier, and entitled Development of a DNA Methylation Multiplex Assay for Body Fluid Identification and Age Determination, 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

_____
Jeffrey Joens

_____
Jeffrey Wells

_____
Bryan Young

_____
George Duncan

_____
Bruce McCord, Major Professor

Date of Defense: November 9, 2020

The dissertation of Quentin Thibault Gauthier 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, 2020

ORCID 0000-0001-8541-2810

DEDICATION

I dedicate this work to my parents, Jean-Marie et Brigitte Gauthier. Your love and support throughout the years has meant the world to me. You have given me every opportunity in life to succeed and encouraged me to pursue my dreams. I could not be the person I am today without you both. I will be forever grateful to you.

ACKNOWLEDGMENTS

I would like to thank everyone in my life that has helped to make this moment possible. Each and every one of you has helped me to realize my dreams in one way or another. My time at Florida International University has helped me to grow as a researcher and a person; this work has shown me my strengths and weaknesses and taught me to how become a better researcher.

First, I would like to thank Dr. McCord for his tireless support and optimism for the work that I have done. He has shown tremendous interest not only in my work, but also in me. He has supported me at every step of the way for this dissertation; guiding me when needed, challenging me when needed. I have become the scientist I am today because of him.

I also thank the rest of my dissertation committee – Dr. Duncan, Dr. Liu, Dr. Wells, Dr. Young, and Dr. Joens. With your support, I have gained the confidence as a researcher to pursue my scientific interests and to push forward with my work. With your criticism, I have revisited my results, updated my understanding, and made my work more robust and impactful. Each and every interaction with all of you has been to my benefit as a student, scientist, and person. Thank you all so much for nurturing my love for the work that I did.

I want to thank my lab mates throughout the years. In particular, I want to thank Joana Antunes, Hussain Alghanim and Sohee Cho. The three of you helped me when I first and took time from your own research goals to help teach me everything you knew about this field and set me up to be an independent researcher. To everyone else in the McCord Research group, thank you for helping to keep me sane, listening to me vent, and helping me out whenever I got in a bind.

Christina Burns, thank you for your eternal optimism and being the most accessible friendly ear that I had during my whole time at FIU. Your level-headedness helped guide me through many stressful moments and your friendly nature helped me to persevere.

To Elizabeth Plant, my 7th grade science teacher, thank you for igniting my love for science that has led me to where I am today.

I would like to thank my parents and my brothers, Jean-Marie Gauthier, Brigitte Gauthier, Nicolas Gauthier, and Alexis Gauthier for always reaching out to me and expressing your love and support. Being so far from family for so long at a time has been difficult, but you have all always made sure to let me know that you are thinking of me and supporting me in your own ways. Thank you and I love you.

Thank you to my closest friends – Ashley Kimble, Harrison Redd, Chiara Deriu, Jenna Chenevert Aijala, Mike Redd and Luke Pirtle. Thank you for your support and kind words. Thank you for helping to keep me focused when necessary, and for distracting me when necessary. Thank you for the late nights watching garbage horror movies, gaming sessions, and drinks with friends. Thank you for forcing me to get some actual sunlight, and for experiencing all that life has to offer.

Additionally, I would like to thank the National Institute of Justice for funding my research; Qiagen for the technical support in my work, especially Mary Jones Dukes and Mark Guilliano; the University Graduate School for awarding me the Dissertation Year Fellowship; the International Forensic Research Institute for the IFRI Student Development Award and finally Florida International University for facilitating my work and continuously being Worlds Ahead.

ABSTRACT OF THE DISSERTATION

DEVELOPMENT OF A DNA METHYLATION MULTIPLEX ASSAY FOR BODY FLUID IDENTIFICATION AND AGE DETERMINATION

by

Quentin Thibault Gauthier

Florida International University, 2020

Miami, Florida

Professor Bruce McCord, Major Professor

For forensic laboratories, the determination of body fluid origin of samples collected at a crime scene are typically presumptive and often destructive. However, given that in certain cases the presence of DNA is not in dispute and rather where the DNA came from is of primary concern, new methodologies are needed. Epigenetic modifications, such as DNA methylation, affect gene expression in every cell of every mammal. These DNA methylation patterns typically are observed as the addition of a methyl group on the 5' carbon of a cytosine followed by guanine (CpG). Methylation patterns have been observed to change in response to the needs of the cell as well as to external stimulus.

The investigation of DNA methylation patterns for forensic applications is a relatively new field, with the first publication in 2010. Since then, enormous growth in knowledge and technology has allowed for new and sensitive applications. Two of the primary branches of DNA methylation analysis for forensic applications are body fluid identification and age determination. In our study, we designed, optimized and validated a body fluid identification multiplex capable of identifying saliva, blood, vaginal epithelia, and semen samples via pyrosequencing. The multiplex assay gives results consistent with

the literature and the interpretation of the results can be automated by classification modeling which reduces human error. The results of the multiplex represent the first multiplex assay via pyrosequencing for body fluid identification. Lastly, the construction of a Targeted Methyl Sequencing assay for body fluid identification and age determination using next generation sequencing was explored in order to push this branch of research into the future of forensic DNA methylation analysis.

As the cost of next generation sequencing begins to come down, it is important that work begins now to ensure that the tools for tomorrow's forensic DNA analyst exist. It is our hope that the results of the targeted methyl sequencing assay serve as a starting point for an exciting future for forensic laboratories across the world.

TABLE OF CONTENTS

LIST OF TABLES

TABLE PAGE

LIST OF FIGURES

FIGURES PAGE

ABBREVIATIONS

5-mC	5-methylcytosine
A	Adenine
aDMRs	Aging-specific differentially methylated regions
ANOVA	Analysis of variance
AIC	Akaike information criterion
AMP	Adenosine monophosphate
ATP	Adenosine triphosphate
APS	Adenosine 5’ phosphosulfate
BER	Base Excision Repair
BIC	Bayesian Information Criterion
BLRT	Bootstrap Likelihood Ratio Difference Test
bp	Base pairs
C	Cytosine
cDMRs	Cancer-specific differentially methylated regions
CpGi	CpG islands
DNA	Deoxyribonucleic Acid
DNMT	DNA methyltransferase
dNTP	Deoxynucleoside triphosphate
dNTPs	Deoxyribonucleotides
dsDNA	Double-stranded DNA
G	Guanine
HRM	High resolution melt
hv	Light energy
iDMRs	Imprinting-specific differentially methylated regions
LPA	Latent Profile Analysis

MPS	Massively Parallel Sequencing
miRNA	Micro RNA
mRNA	Messenger RNA
ms-SNuPE	Methylation-sensitive single nucleotide primer extension
MVPs	Methylation variable positions
NGS	Next Generation Sequencing
ORF	Open Reading Frame
PCR	Polymerase Chain Reaction
PP_i	Pyrophosphate
rDMR	Reprogramming-specific differentially methylated regions
RFU	Relative fluorescence units
RLU	Relative light units
RNA	Ribonucleic acid
SBE	Single base extension
ssDNA	Single-stranded DNA
STRs	Short tandem repeats
T	Thymine
Taq	*Thermus aquaticus*
tDMRs	Tissue-specific differentially methylated regions
TET	Ten-eleven translocation
T_M	Melting Temperature
TE	Tris-Ethylenediaminetetraacetic acid
tRNA	Transfer RNA
TSS	Transcription start site
U	Uracil
UMI	Unique molecular identifiers
VMRs	Variably methylated regions

CHAPTER I – INTRODUCTION AND LITERATURE REVIEW

The application of deoxyribonucleic acid (DNA) analysis for criminal investigations has exploded over the past thirty years. The development of Short Tandem Repeat (STR) kits in the 1990s allowed investigators to have develop profiles that are so specific to an individual, that forensic DNA analysis has earned the distinction of the gold standard in all fields of forensic sciences.[1,2] Since the 1990s, forensic DNA analysis advancements have focused on increasing the number of STRs in an assay, decreasing the amount of DNA needed for analysis, and improving the instrumentation. But in spite of these advancements, there remains large swaths of information useful to investigators that are not currently obtained from standard DNA analysis. This is because information that can be developed from gene expression markers such as epigenetic methylation, has not been applied in forensic analysis.

The term epigenetics was first used by Conrad Waddington in 1942 to describe the "casual interactions between genes and their products, which bring the phenotype into being".[3] Essentially, epigenetics refers to all genome modifications that cause variation in gene expression across cells, that are independent of DNA sequence differences.[4] This means that gene function and phenotypic outcomes across different cells are caused by a wide gamut of biological mechanisms independent of any heritable changes in DNA sequence.[5] Epigenetic modifications include chromatin condensation, post translational modification of histones, differential expression of messenger RNA (mRNA) and modifications to the DNA itself, such as DNA methylation. Each of these modifications play vital roles in gene expression without causing any change to the genome itself.[6] To better understand the mechanisms behind gene expression, the fundamentals of the

structure of genetic information and why different genes need to be expressed at different times and levels must be examined.

A. The molecular structure of DNA

The establishment of the structure of DNA is a result of the tireless efforts of Rosalind Franklin and Maurice Wilkins who developed X-ray diffraction data for the molecular structure of DNA, Alexander Todd who provided an understanding of the DNA phosphodiester bond, and James Watson and Francis Crick who created a model of the structure of double-stranded helical DNA.[7–9] In its simplest form, DNA is composed of two linear strands composed of three main components: a phosphodiester backbone, deoxyribose sugars, and nitrogenous bases. These nitrogenous bases – adenine (A), cytosine (C), guanine (G) and thymine (T) – are located in the interior portion of the double helix and induces the double helical structure of DNA with complementary hydrogen bonding between the nitrogenous bases or nucleotides (Figure 1.1). The double helix form is the result of various chemical forces acting upon the molecule of DNA and all of its various pieces. The covalent bonds found in the nucleotides are the strongest chemical forces present in the entire structure of DNA.

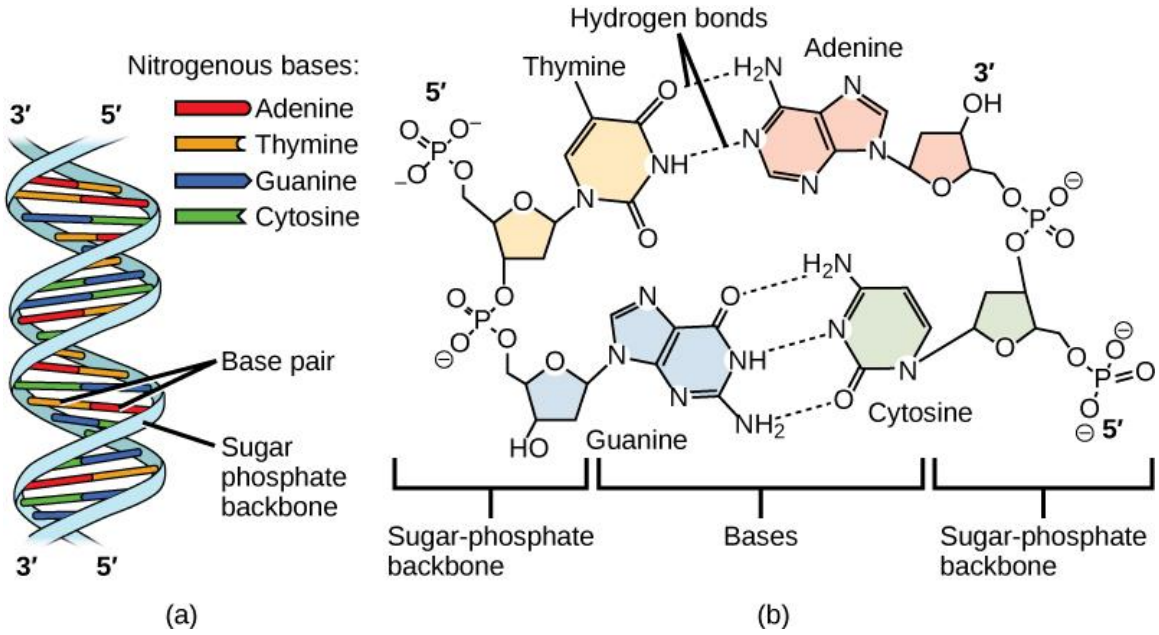

Figure 1.1 – The structure of double helix DNA showing (a) the double helix structure and (b) the four nucleotides shown in antiparallel strands with two hydrogen bonds linking adenine to thymine and three hydrogen bonds linking cytosine to guanine. Reproduced with permission from Molnar and Gair, 2012.[10]

For this reason, the nucleotides are extremely stable and not subjected to random modifications – any modification to the nitrogenous base is quite intentional. Much of the stability of the double-stranded DNA structure is imparted by the hydrogen bonds between nucleotides on complementary strands and Van der Waals forces from adjacent nucleotides within the same strand.[11] Hydrogen bonds are formed between hydrogen donors and acceptors, such as the nitrogen and oxygen atoms found in the four nucleotides. The hydrogen bonding seen in DNA consists of two hydrogen bonds found between an adenine and a thymine, and three hydrogen bonds found between cytosine and guanine.[10] These hydrogen bonds are the primary reason for the large amount of energy needed to dissociate double-stranded DNA into single-stranded DNA and explains why DNA with a larger guanine-cytosine (GC) content requires a higher temperature to completely dissociate than DNA with a relatively low GC content.[12] Beyond the simple yet elegant

structure of DNA is the vast volumes of information that are stored in the endless sequence of nucleotides.

B. Genomic Information and the Central Dogma

Deoxyribonucleic acid itself does not carry out cellular functions. Genomic or nuclear DNA (gDNA) is located in the nucleus of a cell and is incapable of directly carrying out the reactions that take place across the various organelles found in the cytoplasm of the cell. Rather, DNA carries the information needed to generate each and every protein needed within the cell.[13] Francis Crick first proposed the central dogma of biology in 1958 – that genetic information written in the DNA is transcribed in the nucleus to ribonucleic acid (RNA) which then exits the nucleus to be translated to proteins.[14] The central dogma infers three specific characteristics. Firstly, genetic information stored in DNA can be duplicated within the nucleus to form two identical copies of the DNA. Second, the flow of information is unidirectional – DNA to RNA to Proteins. And third, RNA is transcribed in the nucleus, exits the nucleus, and is then translated into a protein in the cytoplasm.[15]

For the proper continuation of the cell's life, replication of DNA within the nucleus must take place. Replication requires a suite of enzymes which include DNA topoisomerase, helicases, DNA polymerases, DNA ligases, and primases. In addition, there are sliding clamp proteins and single-strand binding proteins that impart a level of stability to the strand of DNA that is being replicated.[10] The topoisomerase unwinds the double strand DNA at specific sites along the strand of DNA at the origin of replication which occurs approximately every 35 kilobases (kb). The helicase then binds to the single-strand DNA (ssDNA) in order to make room for the primase to create RNA primers that will bind

to the template strand of DNA.[16] These RNA primers then allow for a DNA polymerase to clamp around the strand of DNA and begin replication. The DNA polymerases incorporate deoxynucleoside triphosphates (dNTPs) along the new strand of DNA with the incorporated nucleotide being complimentary to the template strand of DNA. The available 3'-hydroxyl (3'-OH) group of the deoxyribose sugar in the incorporated nucleotide is then able to attack the alpha-phosphoryl group of the next dNTP that is to be incorporated. This incorporation event creates a pyrophosphate (PP_i) which can later be used to form more dNTPs, most commonly Adenosine Triphosphate.[17] This process of incorporating new dNTPs continues until the polymerase reaches a terminator region, which signals the end of a particular gene, or, in the case of cell maintenance, the full strand of DNA has been replicated in order to be passed to the subsequent generations of the cell during cellular division.[18]

Transcription, the second process of the central dogma, is the process of transcribing information stored in DNA to RNA. The process is largely similar to DNA replication with similar enzymes carrying out the actual process of transcription, but for three key differences. First, RNA is composed of the nucleotides adenine, cytosine, guanine, and uracil (U), but not thymine. Secondly, the transcription process is started at transcription start sites (TSS) with the nucleotides located before the TSS referred to as being upstream and denoted with a negative number relative to their distance to the TSS and nucleotides after the TSS referred to as being downstream and denoted with a positive number relative to the TSS.[19] The third difference is that RNA exists as a single stranded molecule as soon as transcription is complete. The newly formed strand of RNA (pre-mRNA) undergoes various post-transcriptional modifications (addition of poly-A tail,

addition of a 5'-cap and splicing, before emerging from the nucleus as a fully formed mRNA.[16]

The final process of the central dogma is translation, reading the code found within the mRNA to synthesize proteins that carry out any number of cellular functions. Upon exiting the nucleus, mRNA is fully captured by the two subunits of the ribosome (60S and 40S). The larger subunit is responsible for forming the peptide bond between amino acids while the smaller subunit is responsible for reading the mRNA strand to determine the corresponding amino acid. Upon assembly, the ribosome scans the RNA until it finds the start codon (AUG). Once found, the ribosomes shift along the RNA every three nucleotides which is referred to as the open reading frame (ORF). Within the ORF, codons are read by the 40S subunit to recruit the proper transfer RNA (tRNA) holding the next amino acid. The amino acid is incorporated onto the chain of amino acids by the 60S subunit and the process continues.[11] Because each codon is composed of three nucleotides, there are 64 possible combinations. Among these combinations, AUG codes for the start of translation and the incorporation of methionine and UGA, UAA and UAG all code for the stop of translation. The remaining 60 combinations code for the remaining 19 amino acids which allows for multiple combinations leading to the same amino acid. The redundancy allows for the cell to create most proteins without defect even if there is some error committed during the transcription process.[20] The proteins created within a cell will be relevant to the type of cell. For example, salivary alpha-amylase is an enzyme that breaks down starches into maltose and glucose.[21] As the name implies, this protein is primarily found in saliva, as well as the pancreas, and the regulation of this gene's expression would prioritize the creation of this protein in saliva, but not other tissue types, such as blood.[21] The ability to

ensure that proteins are only created where they will be useful is the primary purpose of the various mechanisms that dictate gene expression.

C. Gene Expression and Mechanisms

As previously discussed, there are various mechanisms that can be used to differentially express specific genes found within the human genome. Given that the human genome is approximately 3,200,000,000 bases (3.2 Gb) long and each cell contains a full copy of the genome, it is critically important that cells have a way to dictate which genes will be expressed, and therefore which proteins will be produced, so that resources are not wasted.[13] Some of the most commonly researched epigenetic modifications that dictate gene expression are chromatin condensation, post translational histone modification, changes in mRNA resulting from the transcription process, or modifications to the DNA itself, such as DNA methylation. Because DNA codes for the genetic information necessary for chromatin condensation and mRNA synthesis, DNA methylation is intrinsically connected to the modifications of chromatin and RNA, and is therefore one of the primary drivers for cell-specific gene expression.[22]

Several epigenetic mechanisms can result in the relaxation of chromatin condensation that will influence the level of DNA transcription. Precisely 147 bp of DNA wrap around histones, and the tighter the DNA:histone association, the less available DNA will be for transcription.[23] The proximity of DNA to histones in a chromatin fiber is influenced by the presence of various modifications to the histone tails, which are primarily comprise of arginine and lysine residues.[20] Modifications to histone tails occur post-translationally and include methylation, acetylation, phosphorylation, propionylation,

butyrylation, ubiquitlation, sumoylation, and citrullination. These modifications are referred to as histone marks, and combinations of histone marks and combinations of histone marks create a histone code which specifies a particular gene expression event.[24]

Much research on gene expression has focused on the methylation of lysine (K) residues of the tail on histone H3. More specifically, the tri-methylation events on lysine 9 of histone 3 (H3K9me3) affect gene repression and tri-methylation events on lysine 4 of histone 3 (H3K4me3) impact gene activation.[25] Beyond post translational modifications to the histone, there are also structural variants to the histone itself. Small variations in amino acid sequence have given rise to the histone H2 variants H2A.X and H2A.Z. These two variants function differently – H2A.X, when phosphorylated, denotes a double strand DNA break while H2A.Z shows a negative correlation with DNA methylation at transcription start sites – but both can result in differential expression of various regions of DNA. The fact that additional, less well understand mechanisms exist for the removal and addition of these histone variants suggests that their existence is an intended form of gene expression in organisms.[22]

As for RNA, two mechanisms are known to dictate gene expression at the posttrascriptional level: RNA secondary structure binding preventing transcription and small regulatory RNA binding to mRNA. The secondary structure of RNA can be caused when factors bind to mRNA transcripts preventing the mRNA to unfold into a linear strand available for translation. An example of this process includes the feedback loop regulating iron content within a cell. The primary protein regulating iron content is ferritin and the 5'-end of the mRNA for this protein must be free of iron regulatory protein (IRP) in order for translation to occur. Iron has a much higher affinity for IRP, meaning that when excess

levels of iron exist in the cell, IRP switches binding to iron rather than the ferritin-coding mRNA, leading to higher levels of ferritin which then reduces the level of iron in the cell. Once iron levels are reduced, the IRP reverts back to mRNA binding and ferritin expression is reduced once again.[26]

The portion of the human genome that codes for actual proteins is only 48 Mb, while large swaths of the human genome code for intergenic DNA (2,000 Mb) or introns, untranslated regions, and pseudogenes (1,152 Mb). This includes regulatory regions that code for micro RNA (miRNA) which, in conjunction with mRNA, allow for the regulation of gene expression.[19] These miRNA are complementary to mRNA and lead to the recruitment of proteins from the Argonaute family which have the ability to completely degrade mRNA transcripts.[26] This form of gene expression control has also been linked to back-signaling to DNA, leading to a repressed chromatin expression for that region of the DNA. This mechanism has also been found to be capable of being passed down through multiple generations of cell division.[22]

DNA methylation represents one of the primary focuses of epigenetic research given the widespread effects of this DNA modification, and the available pathways for research via chemical modifications and instrumental analysis. In humans, methylation occurs almost exclusively on the 5' carbon of a cytosine forming a 5-methylcytosine (5-mC) (Figure 1.2).[27] Although it has been demonstrated that 5-mC can be followed by adenine (CpA) cytosine (CpC) and thymine (CpT), the primary form of DNA methylation occurs by 5-mC followed by guanine (CpG)[15] This form, CpG, has been associated with so many portions of gene expression, that it has been unofficially dubbed the 5th nucleotide.[28] Methylated cytosines are estimated to comprise approximately 4-6% of all

cytosines in the human genome, with CpGs accounting for more than half of the methylated cytosines.[29,30] Regions upstream of transcription start sites, promoter regions, often have higher concentrations of CpG dinucleotides, giving the moniker CpG islands (CpGi), and have been associated with gene expression. The existence of these CpG islands, predominantly in promoter regions, helped to spur the research examining DNA methylations role in gene expression.[31]

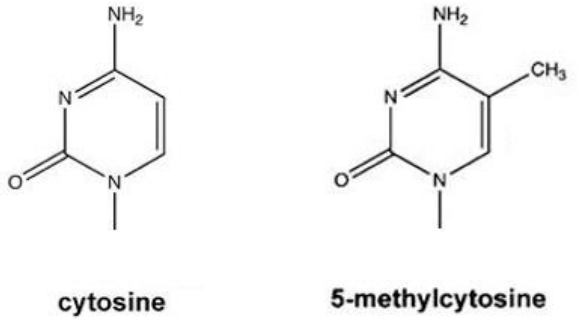

Figure 1.2 – DNA methylation of a cytosine residue consists of the addition of a methyl group to the 5' position. Reproduced with permission from Genereux, Johnson, and Burden et al., 2008.[32]

D. DNA Methylation – Mechanisms, Functions, Influences

Cytosine methylation is regulated primarily through a family of proteins called DNA methyltransferases (DNMTs), which includes DNMT1, DNMT2, and DNMT3.[33,34] These proteins are charged with the process of transferring a methyl group from S-adenyl-methionine (SAM) to the fifth carbon of a cytosine residue resulting in a methylcytosine.[35] The mechanism of DNMTs during embryonic development, while poorly understand thus far, are critically important to the establishment of methylation patterns across the genome in the embryo. Specifically, levels of DNMT3A expression are highest in germ cells, while DNMT3B are highest in the period of early development after fertilization.[36] Various

histone modifications in conjunction with the DNMT3 family of proteins are largely attributed to the creation of the methylome during the fertilization of a new line of cells.[37] After the methylome is established, it becomes the responsibility of the other DNMT proteins to maintain methylation patterns across the genome for proper gene expression. During the cell cycle, after DNA duplication but before cell division, DNMT1 is charged with establishing the methylation pattern on the newly formed strand of DNA to compliment the methylation found on the template strand of DNA. Recruitment of DNMT1 for this task is caused by a higher affinity for DNMT1 to hemimethylated DNA, DNA that is methylated only on the template strand while the newly synthesized strand is completely unmethylated. Additionally, DNMT1 contains a motif able to bind to the sliding clamps associated with DNA replication which allows for an almost immediate establishment of the methylome in the newly synthesized strand of DNA.[38] While maintenance of the methylome is the responsibility of DNMT1, *in vitro* experiments of successive cell divisions where DNMT3A and 3B have been knocked out showed a gradual decrease in global methylation across successive generations.[37] These results suggest that the stability of the methylome is imparted by a careful combination of DNMTs and other factors.

Just as important as the methylation of cytosine is the demethylation of cytosine. During mammalian development, the male genome is actively demethylated while the female genome is passively demethylated.[39] This demonstrates that proper development of embryos, including differentiations based on gender, is reliant in part on the process of demethylation. Passive demethylation is the process by which DNMT1 does not recognize, bind, or properly function when presented with a hemimethylated strand of newly synthesized DNA. If the cell divides a second time before this error can be corrected, the

change will become permanent unless acted upon by other forces, either biological or environmental.[34] Active demethylation utilizes a family of three enzymes – ten-eleven translocation (TET) – that are associated with embryonic development, meiosis, stem-cell reprogramming and maintenance of the DNA methylation patterns that control gene expression.[40] These TET proteins have been associated with the oxidation of 5-mC to 5-hydroxymethlcytosine (5-hmC), 5-formylcytosine (5-fC), and 5-carboxylcytosine (5-caC). And while the exact function of these nucleotides has not yet been determined, the relatively high levels that can be observed in human DNA samples would suggest that they are not merely intermediates.[41] One theory suggests that the concentration of 5-hmC influences the recruitment of further TET enzymes and Base Excision Repair (BER) enzymes to increase the oxidation state of the 5-hmC to 5-caC so that the BER enzymes will treat the oxidized cytosine as damaged DNA requiring repair, at which point the modified cytosine will be replaced with an unmodified cytosine, effectively stripping the methylation status at that CpG site until acted upon by a DNMT.[41]

The predominant function of DNA methylation has been to inhibit DNA transcription, and therefore regulate gene expression (Figure 1.3).[6] It should be noted, however, that some gene activation has been observed with increased levels of gene expression in certain biological feedback loops.[42] The mechanism of DNA methylation affecting transcription is thought to occur in three primary fashions. First, the presence of methyl groups on cytosine can physically block the binding of transcription factors which will prevent transcription from starting.[43] Second, the recruitment of methyl-CpG binding proteins can create a more pronounced physical block to the association of transcription factors to the template strand during transcription.[44]

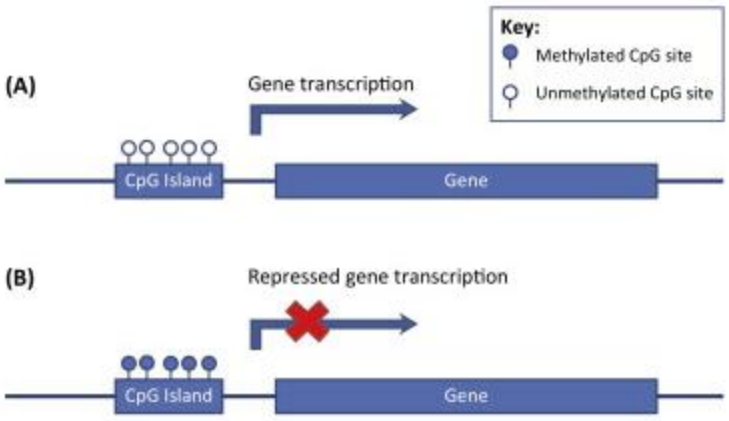

Figure 1.3 – The degree of methylation in CpG islands upstream of a gene is the predominant fashion in which gene expression is controlled at the DNA level. Reproduced with permission from Nikolova and Hariri, 2015.[45]

The third predominant mechanism of DNA methylation affecting gene expression is chromatin packaging. Given that most methylated CpGs are found to cluster in the center of nucleosomes, the chromatin goes into a condensed state. Additionally, chromatin modeling proteins rely on the presence of 5-methylcytosine to shape the structure of the chromatin.[43] This can help to explain why regions of the genome are silenced by default throughout the lifetime of an organism, such as the second copy of the X chromosome in females, and only become active when an external force acts upon the cell.

Research into DNA methylation has classified various subsets of DNA methylation using the patterns that are observed in organisms with respect to changes in the methylation values. DNA methylation can be subdivided into three main categories: methylation variable positions (MVPs), variably methylated regions (VMRs) and differentially methylated regions (DMRs).[46] Methylation variable positions consist of regions containing a single CpG site that is either methylated or unmethylated while variably methylated regions contain multiple CpG sites expressing a variable level of methylation dependent

on the need for gene expression.[47] Finally, DMRs consist of regions containing multiple CpG sites with a methylation pattern that differs dependent on several factors. As such, DMRs have been assigned to a series of subcategories including: tissue-specific DMR (tDMR), aging-specific DMR (aDMR), imprinting-specific (iDMR), reprogramming-specific DMR (rDMR) and cancer-specific DMR (cDMR).[48]

There are a host of factors that influence the levels of methylation that can be seen in a cell. These factors include nutrition, diet and exercise, alcohol and tobacco consumption, drug use, aging, pollution, exposure to carcinogens, etc.[33] There are some proposed pathways for how these factors can affect DNA methylation including increased folate consumption in diets leading to increased SAM, the methyl donor used by DNMTs, which helps to maintain proper methylation levels in an organism.[49] In addition, physical exercise has been associated with the hypomethylation of genes associated with inflammatory response, leading to quicker recuperation compared to individuals that do not exercise.[49,50] Exposure to environmental factors, such as pollution and lifestyle choices, can also have significant effects on the methylome. Environmental factors can affect early embryogenesis, when the establishment and replication of the methylome is most critical, leading to embryonic programing of disorders that manifest later subsequently in ontegeny.[51] Exposure to environmental factors capable of affecting DNA methylation later on in life are less likely to result in wide-spread effects on the methylome, but are unfortunately associated with tissue-specific carcinogenesis.[52] When the methylome of monozygotic twins that have not spent much of their life together or who have very different lifestyles was examined, the DNA methylation levels between the two siblings showed enough difference to be able to easily differentiate between the two, despite the

twins having identical genomes.[53] Interestingly, the NASA Twins Study was able to detect subtle changes in the methylation levels in cluster of differentiation 4 (CD4) and cluster of differentiation 8 (CD8) Cells within a pair of monozygotic twins after one twin spent just under one year in space, suggesting that these environmental changes can occur relatively quickly. The methylation values in the astronaut who travelled to space returned to levels comparable to both before spaceflight and his twin within six months of his return to Earth.[54]

The change in DNA methylation as a function of biological age has also been of extreme interest within the scientific community. Aging is generally understood to be the accumulation of mutations in the genome or a decreased adaptability to the environmental factors that affect the human body. This damage can be attributed to the accumulation of reactive oxygen species within the cell or alterations to DNA repair mechanisms through diseases like cancer, possibly due to a loss of function of BER enzymes.[55] Aging has also been associated with global hypomethylation, possibly caused by the loss of function from DNMT1, but with some localized hypermethylation in regions coding for transposable repetitive elements.[56] This decrease in global methylation was observed when comparing the CpG methylation of newborns to centenarians.[57] Although the exact mechanism by which age affects DNA methylation is not clear, many theorize that changes to methylation levels as a result of aging are mostly associated with stochastic effects and the previously described environmental factors.[58] The differences seen in the methylomes of monozygotic twins generally increase with biological age. This epigenetic drift is most likely caused by stochastic errors in the establishment of the methylome after DNA replication that becomes cemented in the subsequent generation of cells.[53] The information contained in these age-

associated DNA methylation studies has prompted a slew of research to predict the biological age of an individual using only a sample of their DNA.[59–61] One of the constants in the studies, however, has been the need to establish the tissue that the DNA originated from. This is the result of much larger variations in predicted age when evaluating multiple tissue types from the same individual versus the same tissue type from individuals of varying ages.[62] Luckily, tissue identification via DNA methylation analysis has also been a much studied subject yielding assays capable of differentiating from a number of commonly encountered tissue types.[63–65] As it pertains to forensic investigations, the ability to identify tissue types from tDMRs offers a more direct confirmatory test than the commonly used serology tests currently used in forensic laboratories. As for biological age prediction, actively employed forensic DNA analysis methods currently offer no viable solution, and such an assay would provide an incredibly valuable compliment to existing forensic assays that permit the determination of phenotypic traits such as biogeographical origin, hair color, and eye color.[66]

CHAPTER II – CONTEMPORARY FORENSIC DNA ANALYSIS

With the discovery of short tandem repeats (STRs) and development of the polymerase chain reaction (PCR) method in the 1980s, forensic DNA analysis has developed a strong foundation for the establishment methods to identify unknown individuals in criminal investigations. In 1985, Kary Mullis invented the protocol for PCR, which largely mimics the mechanism of DNA replication in a nucleus, but with the ability to choose specific targets of the genome that will be replicated.[67] In PCR, a sample of DNA is added to a reaction mixture containing a DNA polymerase, primers targeting specific regions of the human genome, dNTPs, and enzymatic co-factors such as magnesium ions. The reaction mixture, cycled across several temperatures to induce dissociation of dsDNA and allow for primers to create a new strand of DNA, results in millions of copies of the target DNA. These targets, STRs, were identified in the 1980s, but were not employed by forensic scientists until 1995 when the United Kingdom Forensic Science established a six STR assay.[68] Short Tandem Repeats are non-coding regions of the genome the exhibit repetitions of two to seven base pairs, repeating as many as twenty times. In 1997, the FBI laboratory established the 13 core STRs for the USA database, and numerous companies quickly came to the market with panels of 15 or more STRs capable of powers of discrimination in excess of one in a trillion.[68] The steps associated with forensic DNA analysis have matured over the past 25 years, focusing primarily decreasing the amount of DNA needed for analysis, decreasing the amount of time for analysis, increasing the discriminatory power, and advancements in automation meant to reduce the chance of human error.

A. Sample Collection

The very first step in forensic DNA analysis is the proper collection of DNA samples that avoids contamination and aids in preservation of the sample. As the advancements in the sensitivity of DNA analysis have increased, the variety of samples that can be collected for analysis have increased, which has emphasized the need for proper training in collection and storage of the samples themselves. At the scene of a crime, the first responders may likely be law enforcement or emergency medical professionals whose first priorities are not necessarily the preservation of evidence. For this reason, it is important that the crime scene technicians that follow up later have a good understanding of the crime scene, where DNA evidence is likely to exist, and to collect it in the best way possible. In most cases, the evidence is going to be collected using a cotton swab that can be stored in a sterile package, and transported to laboratories for analysis.[68] However, DNA samples can be recovered from a multitude of items found at crime scenes, including clothing, cigarette butts, and dining utensils. Personnel collecting these samples should be wearing gloves and using other personal protective equipment that can help to prevent contamination. Given that current methodologies can reliably produce DNA profiles with as low as 100pg of DNA, any level of contamination at the crime scene and at the time of sample collection can have impacts on the whole process.

Alternatively, reference samples can be collected for comparison to the samples found at crime scenes. These reference samples can come from suspects or potential innocent donors, like family members and police personnel, for exclusionary purposes. Additionally, reference samples are collected for convicted felons, and all arrestees in some jurisdictions, and stored in databases, such as the Combined DNA Index System (CODIS)

at the national level, for future comparisons. These reference samples will mostly be collected as a buccal swab by rubbing a cotton swab on the inside of the cheek. Since these samples of DNA are relatively fresh, there will be a very large amount of DNA in the sample, and the risks of contamination and degradation of the sample are not nearly as high.[69]

Upon arrival to DNA analysis laboratories, it is critical that samples are stored properly so that sample degradation via hydrolysis caused by humidity, crosslinking by ultraviolet (UV) radiation, and enzymatic degradation resulting from DNases are minimized. Samples should be kept frozen or refrigerated but can also be stored at room temperature depending on the medium used for sample collection. For example FTA® paper is a popular storage medium which impregnates chemicals in the paper which lyse the cell and stabilizes the DNA within the filter paper for protection.[70] Regardless of storage method, prompt analysis of the sample is always ideal, which starts with proper determination of tissue type and extraction of the DNA from the collected sample.

B. Serology

The samples collected from crime scenes can come from a number of different tissue types. Understanding which tissue type was collected from the crime scene aids forensic DNA analysts in determining the proper methods for DNA extraction and can help establish the relevancy of that particular piece of evidence. For example, demonstrating that the DNA found on a piece of evidence is semen, rather than saliva, can establish the context of how a crime occurred. To achieve tissue determination, forensic DNA analysts perform a number of serological tests that are designed to either presumptively identify the

presence or absence of a number of body fluids, or to confirm the identity of a specific body fluid in the sample. One drawback of confirmatory tests, however, is that they are often destructive and therefore do not allow additional analysis of this portion of the evidence. These serological tests often take advantage of changes in color, fluorescence, or other physical properties that change in the presence of a chemical test.[71] The detection of saliva, blood, semen, and vaginal epithelial cells, the primary body fluids found in evidence and examined in this dissertation, will be examined further.

A common presumptive test for the detection of saliva is the Phadebas test which relies on the detection of enzymatic activity of alpha-amylase, an enzyme used to break down starches into sugars.[72] Unfortunately the test is not human specific, and because structural variants of the amylase enzyme are secreted in the pancreas, there is a chance of false positives with the Phadebas test when testing urine.[73] Confirmatory tests for saliva include the commercially available RSID™-Saliva and SALIgAE® from Independent Forensics and Abacus Diagnostics Incorporated, respectively. The RSID™-Saliva is a lateral flow immunoassay containing two monoclonal antibodies that are specific to alpha-amylase. It can achieve sensitivities as low as 0.01 μL of saliva.[74] A positive response for this test involves the migration of the antibodies along the strip, eventually causing a colored stripe to appear on the strip, similar to a pregnancy test. SALIgAE® functions similarly, detecting the presence of the enzyme directly, but in the form of a liquid that can be sprayed on to surfaces or swabs. Unfortunately, this method requires approximately 10 μL of saliva for reliable results, which is not always the case for forensic samples. Although both of these tests are commonly used, they still suffer the issue of not being human-

specific, and can produce false positives for amylase originating from the pancreas, and in some cases from breast milk.[75]

Blood is presumptively identified with mainly with two tests: Kastle Meyer and Luminol. These tests focus on the presence of hemoglobin in blood. The Kastle Meyer test relies on hemoglobin to act as a catalyst for the oxidation of phenolphthalein in the presence of hydrogen peroxide. The reaction causes phenolphthalein to switch from a colorless liquid to a light pink liquid.[72] Luminol relies on the iron present in hemoglobin to react with the compounds found in luminol: 3-amino-phthalhydrazide, sodium carbonate, and sodium perborate. The end result of this reaction is the fluorescence of blood samples which can easily be visualized with UV light.[76] The confirmatory tests for blood include the Takayama test, ABA card HemaTrace, and RSID^{TM}-Blood. The Takayama test utilizes dextrose, sodium hydroxide and pyridine to create small crystals in the presence of blood which can be observed under a microscope.[76] The ABA card HemaTrace and RSID^{TM}-Blood, like the RSID^{TM}-Saliva, are lateral flow assays that contain antibodies and produce stripes along the paper strip. The ABA card HemaTrace specifically targets human hemoglobin with moderate sensitivity, though with some false positives with saliva reported.[77] The RSID^{TM}-Blood test targets GlycophorinA, a protein found on the cell membrane of erythrocytes, rather than hemoglobin.[78] The common theme of these three confirmatory tests, however, is that they are destructive to the portion of the sample that was tested, and therefore do not represent the ideal scenario for forensic DNA analysts.

Semen, the primary body fluid used for the establishment of sex crimes, can be presumptively identified using tests for acid phosphatase and prostate specific antigen (PSA). Acid phosphatase is an enzyme present at approximately 400 times higher in

seminal fluid than other body fluids and is tested by combining a sample with sodium alpha-naphtylphosphate and Fast Blue B, resulting in a dark purple color change.[79] The PSA test is for the detection of p30, a protein that is most commonly found in semen, even when the semen originated from an azoospermic male.[80] Confirmation of semen is primarily accomplished by a process called Christmas tree staining. The staining method colors the sperm heads red and the sperm tails green. Under a microscope, the observation of sperm cells is quite easy and confirms that the samples contain sperm.[81] Unfortunately, if the individual is azoospermic, the result will be negative, even though semen may be present. Once again, there is a later flow assay available in the form of RSID™-Semen. This assay is specific for semenogelin, a protein produced in the seminal vesicules and is a component of semen.[82] However, this assay can suffer from false negatives in the presence of mixtures, which is a fairly common problem with collected presumed semen samples.[83]

Presumptive and confirmatory tests for vaginal epithelial samples are scarce. The acid-Schiff test can be used to stain glycogenated epithelial cells can be used, but the level of glycogenation is significantly affected by menstrual cycle, and false positives from the mouth and urethra of males have been observed.[84] Another presumptive test detects the presence of isoenzyme 4 and 5 of lactate dehydrogenase, but this test is completely non-specific to vaginal epithelia.[85] The lack of testing for vaginal epithelia has opened the door for many innovative research studies to try to address this problem.

Upon identification of the body fluid that has been collected, an extraction and purification technique can be selected to optimize the recovery of DNA from the sample.

C. DNA Extraction

As previously described, DNA is contained in the nucleus of cells regardless of the origin of that cell. And while the evidence collected at crime scenes can vary wildly, the DNA evidence is usually a body fluid, such as saliva, blood, vaginal epithelia, or semen, and extraction methods focus on two primary portions: cell disruption and sample purification. Cell disruption is primarily achieved through the use of detergents, like sodium dodecylsulfate (SDS), and proteinase K. The detergents cause the cell membrane to break down, while the enzyme, when used in combination with elevated temperatures, can break down the proteins that are found in most cells. After lysis, the resulting sample consists of free-floating DNA, cellular debris, and various constituents of the cytoplasm.[86] Purification of the sample is next and can be achieved in a number of ways. The following two methods are the ones that were employed throughout the course of this dissertation research.

The first method is Phenol-Chloroform-Isoamyl alcohol (PCIA) extraction, commonly referred to as Organic Extraction. The PCIA method, relying on the principles of affinity for DNA and cellular components to separate across organic-aqueous mixtures, has been in used since the beginning of DNA analysis because of its reliability and cost effectiveness, though it can suffer significantly from human error if part of the organic layer is transferred to the final product which can cause PCR inhibition.[87] In the PCIA method, a phenol-chloroform-isoamyl alcohol mixture (25:24:1 v/v) is combined with the lysis product containing DNA and cellular components. After thorough mixing and centrifugation, the DNA is in the aqueous phase while cellular components have migrated to the organic phase. The aqueous phase can then carefully be removed and purified via

ethanol precipitation or specific filter papers. In some protocols PCIA is used multiple times to ensure absolute purity, however this can also lead to a loss of some DNA in the sample due to repeated pipetting steps. Additionally, sample loss can occur when pipetting the aqueous phase out of the mixture since it is extremely important that no phenol is accidentally included in the final DNA extract.[88]

The second purification method relies on solid-phase extraction, referred in this application as silica bead purification. This method relies on the use of chaotropic salts, such as guanidinium chloride at a low pH, added to the lysis product to induce further protein denaturation and disruption of the hydrogen bonds and Van der Waals forces acting to impart stability to the DNA.[89] Increasing the pH slightly causes the DNA to adsorb to the surface of the silica-coated magnetic beads which can be immobilized by a magnetic on the side wall of the sample tube. With the DNA selectively immobilized in the sample tube, repeated washes of the sample can achieve purification of the sample. Once purified, an alkaline elution buffer is added to the sample tube to dissociate the DNA from the silica-coated magnetic beads.[90] The purified sample is transferred to a new sterile tube that can be stored for future uses.

After extraction, samples can be stored at -20 ºC. Frozen extracts have shelf lives lasting years and can be thawed and tested numerous times before freeze-thaw cycles begin to cause mechanical fragmentation of the DNA and affect the quality of downstream analysis.[91] With the DNA extracted, the next priority is confirming how much DNA has been isolated in this sample so that informed decisions about future analysis can take place.

D. DNA Quantification

Quantification of DNA allows for informed decisions along the rest of the DNA analysis workflow. Many of the forensic assays on the market today require no more than 1ng of DNA and amounts higher than that can actually cause issues during PCR or while analyzing the amplified product via capillary electrophoresis. There are a number of commercial kits designed for the quantification of DNA samples in a forensic laboratory setting. These kits have been designed and optimized in a number of ways for forensic DNA analysis. Firstly, these commercial kits have been designed to quantify only human DNA, and not bacterial DNA that may have been collected alongside the original body fluid or the DNA from bacteria and yeast that naturally exist in body fluids such as saliva and vaginal fluid.[92] Another design point in these commercial kits is the evaluation of the quality of the DNA via the inclusion of a DNA targets that are very small and very large in size. The ratio of the signal for these two targets can give insight regarding the level of fragmentation that the DNA sample has suffered as a consequence of various environmental factors.[93] Lastly, the inclusion of an internal control for the detection of PCR inhibitors has been developed. Stated simply, if PCR inhibitors exist in the DNA sample, they will negatively affect the amplification of the internal control, and the DNA analyst must make further decisions regarding how to proceed with the sample.[94] The advancements in DNA quantification have greatly improved the ability of forensic laboratories to handle and screen more samples for downstream analysis but has also resulted in a cost per sample that can be prohibitively expensive for research applications. For this reason, there were three main protocols that were used throughout this project.

All three protocols rely on the use of a standard curve that relates the level of fluorescence to the concentration of DNA in the sample. The first method, Alu quantification, relies on the process of real-time PCR and an intercalating dye, SybrGreen.[95] Real-time PCR relies on the use of a thermal cycler that can observe increases in fluorescence in real-time, typically after each complete cycle of PCR. In this method, a series of standards with known concentrations of human DNA are run to establish a normalized response of DNA concentration to fluorescence, Figure 2.1. Extracted samples are run alongside the standards and the standard curve from the known samples are used to quantify the DNA in the unknown samples. SybrGreen, the dye in this reaction, binds to dsDNA. After each cycle of PCR, there will be more and more DNA available in the sample for the dye to bind due, and therefore a larger signal of fluorescence will be observed. The PCR cycle at which the fluorescence starts to increase exponentially is call the cycle threshold (C_T) and is used to create the standard curve. The standards with more DNA require less cycles to reach the C_T than the standards with less DNA. With this information, a graph is generated that plots the C_T versus the log of the concentration of the standards and a linear regression formula is calculated to allow for the concentration of the unknown samples to be determined.[96] The AluQuant method uses primate-specific primers that target multiple Alu repeats that are found throughout the human genome. This method is highly sensitive as a single cell can contain multiple copies of the Alu sequence.

The other two methods used are the PicoGreen method and the use of the commercial Qubit™ system. These two methods work in largely similar fashions. Both methods use an intercalating dye, but do not require any amplification of the DNA. For this reason, the methods are not considered to be human specific, but considering the known

origins of the samples, this is not of much concern. The intercalation of this fluorescent dye is proportional to the amount of genomic DNA present in a sample and the use of standards with known concentration is once again employed to be able to infer the concentration of the DNA in the unknown samples. The primary difference in these two methods is that the Qubit™ system is proprietary to Invitrogen, a Thermo Fisher Scientific company, and uses only two standards at the minimum and maximum limits of detection, while the PicoGreen method is non-proprietary and uses a series of standards along a range of concentrations in order to quantify unknown samples.

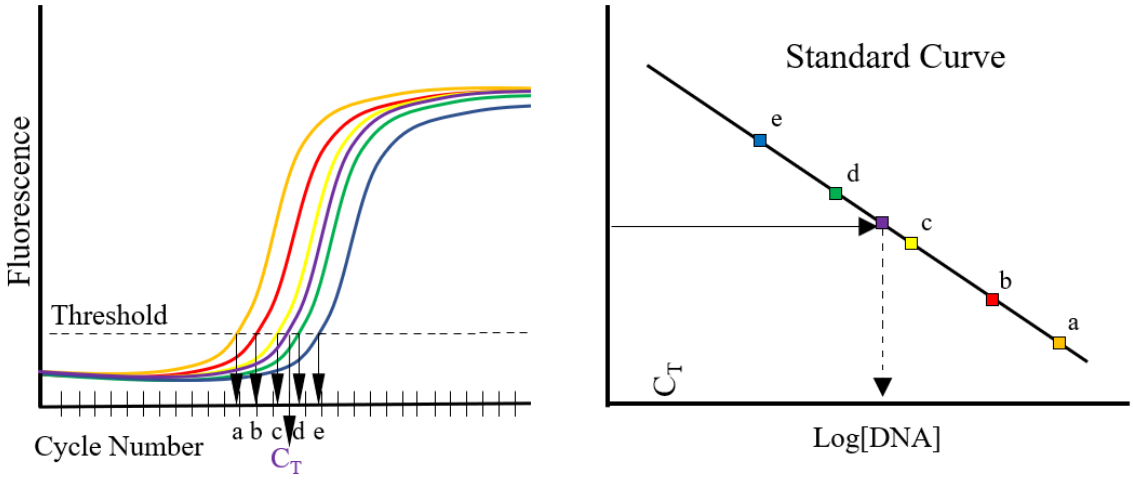

Figure 2.1 – Representation of a real-time PCR reaction. As standards and unknown samples undergo PCR, their fluorescence is individually recorded. After crossing the C_T, the data is graphed to show the known concentration of the standards versus the observed C_T values. This can then be used to infer the concentration of DNA in the unknown samples. Inspired by Butler, 2009.[68]

E. Polymerase Chain Reaction

The use of PCR has been the foundation of all advancements in forensic DNA analysis. This reaction largely mimics the process of DNA replication in the cell with the primary difference being that the primers that dictate which region of DNA is to be targeted for amplification is specific to the application that the scientist desires. For example, the

latest forensic STR multiplex assays target 24 loci that are capable of reaching probability of exclusion as low as one in a septillion.[97] The act of simultaneous multiplex amplification of several regions of DNA in the same reaction lowers cost and increases the specificity of the result. Increasing the number of loci analyzed also assists with paternity testing and the analysis of degraded DNA.

The components of the reaction mixture include the DNA template, DNA polymerase, primers, dNTPs, magnesium and a buffer that stabilizes the reaction at a specific pH across a large range of temperatures. The DNA polymerase used can depend on which commercial PCR kit has been purchased, but nearly all of them are functional derivatives of the thermo-stable polymerase found in the *Thermus aquaticus* (Taq) bacteria.[98] The use of magnesium ions in the reaction mixture ensure functionality of the Taq polymerase as it is a co-factor for amplification.[99] The primers that are used in STR kits are typically oligonucleotide sequences approximately 25bp in length, complimentary to either side of the flanking regions of the DNA target and contain a fluorophore for detection during analysis. The primers in a multiplex should ideally have similar melting temperatures, as dictated by length and GC content, so that each target has approximately equal products at the end of the reaction.[100] The melting temperature is defined as the temperature at which 50% of dsDNA or DNA template:primer complexes will dissociate, making the DNA in the reaction single-stranded and available for complexation with primers. Like in DNA replication, the DNA polymerase uses the 3'-OH of the primer to attach the next dNTP complimentary to the DNA template and this process continues until the template ends.

The PCR reaction is carried out across three primary steps: denaturation, annealing, and extension. When a PCR reaction mixture is prepared, it is placed in a thermal cycler that will typically start at 95 ºC in order to activate the Hot Start polymerase, a type of polymerase that is inactive at room temperature so that amplification does not begin until the reaction is ready.[101] After the activation step, the cycling of denaturation, annealing, and extension will begin with the specific number of cycles used following manufacturer recommendation and validation studies of the kit in that laboratory, typically 28-32 cycles. Denaturation occurs at approximately 94 ºC and serves to melt the DNA so that it is fully single-stranded. Next, the annealing step serves to allow for the primers to bind to the template DNA. The specific temperature used in this step is dependent on the melting temperature of the primers. Next, elongation occurs at 72 ºC which is the ideal temperature for Taq polymerase activity. The polymerase will elongate the new strand of DNA. In the first few cycles of PCR, the primary template for amplification will be the genomic DNA, which is much longer in length than the target regions. Eventually, the newly synthesized DNA will be the template, at which point the polymerase will only synthesize new strands of DNA to a specific point. By the end of PCR, the vast majority of the DNA present will be newly synthesized strands of DNA of one singular length. This specific length of amplification product will allow for efficient separation and detection of DNA during capillary electrophoresis.[102]

F. Capillary Electrophoresis

The end result of amplification of the multiplex STR kits is a tube containing DNA targets of various sizes and labeled with a variety of fluorophores. To analyze the DNA,

the targets need to be separated by size and detected. The main method to achieve this in forensic laboratories is capillary electrophoresis. Electrophoresis is ideal because it allows for PCR products to be separated on the basis of size by applying an electric potential across two electrodes. Because DNA is negatively charged, it will naturally migrate toward the positively charged electrode.[68] When moving through a sieving buffer, the smaller PCR products will migrate more quickly through the system, while larger products will take more time to pass by the detector. Another factor that aids the migration of DNA through the capillary is the use of formamide, a denaturant, which forces DNA into a single stranded state. This is caused by a decrease in hydration, which dissociates the hydrogen bonds between nucleotides.[103]

Although electrophoresis in the early days was carried out using agarose and polyacrylamide gels, capillary electrophoresis has completely overtaken those methods because of its far superior throughput, resolution, detection of multiple different fluorescence wavelengths, and data capture software that eases the interpretation of results. The basics of a CE instrument include: a capillary, or capillary array of up to 16 capillaries, a sieving polymer inside the capillary that aids in separation of DNA fragments based on size, electrodes on either end of the capillary in buffer reservoirs that induce migration of DNA, a laser able to excite the fluorophores that are found on each PCR product, a detector that can record the level of fluorescence that was observed, and a computer to coordinate control of the instrument and record all of the data into an easily interpreted form, called an electropherogram. The electropherogram displays peaks with widths and heights corresponding to the amount of PCR product versus the amplicon size which is determined

by the migration times of a size standard that is added to each PCR product before electrophoresis.[104]

Higher throughput in the CE system is achieved by the use of capillary arrays and the ability to dynamically control the voltage applied to the electrodes which allows for preconcentration of each sample prior to migration through the capillary. This ensures that each sample is analyzed in the same manner.[105] Capillary Electrophoresis systems can achieve single base pair resolution, similar to that of polyacrylamide gels. However, because of its high voltage and enhanced heat dissipation, CE offers the ability to analyze samples much more quickly than polyacrylamide gels, and thus speeds up the time to analyze each sample.[106] The various advancements in capillary electrophoresis instrumentation throughout the years have greatly increased the ability of forensic DNA analysts to process a much larger number of samples with ever increasing levels of accuracy. However, in the past few years, a newer technology has been developed that has the ability to completely overtake capillary electrophoresis and completely reshape the way that forensic DNA analysis is conducted.

G. Massively Parallel Sequencing

Massively Parallel Sequencing (MPS) is an umbrella term for a variety of technologies from different companies that have a common goal: the mass collection of data from genetic material on a scale that completely dwarfs all previous methods. Although there are many similarities and differences amongst the sequencing platforms offered by the largest companies e.g., Thermo Fisher Scientific, QIAGEN, and Illumina, the chemistry behind each approach offers comparable end results. For the purposes of this

dissertation, the Illumina sequencing platform, specifically the MiSeq platform, will be discussed. Illumina's approach to next generation sequencing involves the generation of millions of clonally amplified copies of single-stranded DNA captured on a flow cell followed by a sequencing reaction that employs reversible dye terminator nucleotides to facilitate the incorporation of a single nucleotide at a time. After nucleotide incorporation, an image of the flow cell is taken, recording the fluorescence of every single captured strand of DNA, and then the terminator is cleaved off, allowing for the next nucleotide to be incorporated.[107,108]

This process is one of several forms of sequencing by synthesis (SBS). The name is derived from the fact that the sequence data being recorded is from a new strand of DNA actively being synthesized [109]. After the introduction of their first sequencer, the Genome Analyzer, in 2006, Illumina has produced a number of different sequencers that are focused on multiple areas. The MiSeq was introduced in 2011 and was geared towards research laboratories, rather than clinical laboratories, and even came with a variant, the MiSeq FGx, focused towards forensic laboratories with a forensic DNA analysis panel that permits far greater multiplex sizes than CE-based STR kits.[110,111] The MiSeq platform utilizes a flow cell with a single lane containing embedded oligonucleotides that capture specific sequences of DNA that have been added to the DNA targeted for sequencing. The sequencing reaction on this instrument can take up to 56 hours and outputs up to 15 gigabytes of data in the form of DNA fragment reads with lengths up to 300 base pairs.

Prior to the sequencing reaction, DNA samples need to be prepared in a specific manner, called library preparation, in order to be properly captured on the flow cell. First, DNA samples are fragmented enzymatically using non-specific endonuclease mixes. This

makes the next step, adaptor ligation, capable of reaching as much of the genome as possible. Adaptor ligation is the process of adding specific oligonucleotides to both ends of the fragmented DNA. These adaptors contain a sequence that is complimentary to the oligonucleotides embedded in the flow cell. At this point, most protocols call for either target enrichment of the DNA or for universal PCR. Target enrichment allows for regions of the DNA to be amplified using specific primers while universal PCR uses primers that bind to the adaptors themselves, resulting in equal amplification of the whole genome.[112] The advantages of target enrichment are that specific regions of the genome can be interrogated without generating large amounts of superfluous data. This concept will be explored further in Chapter 3 of this dissertation. Universal PCR is typically employed when sequencing the whole genome is the purpose. The final portion of library preparation is quantifying each sample library to dilute the sample to a level of DNA that is appropriate for the sequencing platform and flow cell. Overloading a flow cell with too much DNA can lead to larger fragments not being captured on the flow cell as efficiently when the sample is flowed across the flow cell prior to sequencing.[113]

When the library is loaded on to the flow cell, the embedded oligonucleotides capture DNA fragments containing adaptors. Prior to sequencing, DNA templates are amplified in the flow cell by bridge amplification (Figure 2.2). Bridge amplification relies on the captured strand of DNA arching and finding an oligonucleotide complimentary to the adaptor on the floating end of the DNA. The oligonucleotide:DNA complex acts as a starting point for a polymerase to generate a new strand of DNA that is anchored at the second oligonucleotide. This results in a doubling of the amount of DNA captured on the flow cell. This cycle is repeated in a process called cluster generation until the flow cell is

completely saturated. At this point, the clusters are denatured, a sequencing primer is added to the flow cell, and sequencing by synthesis can begin.

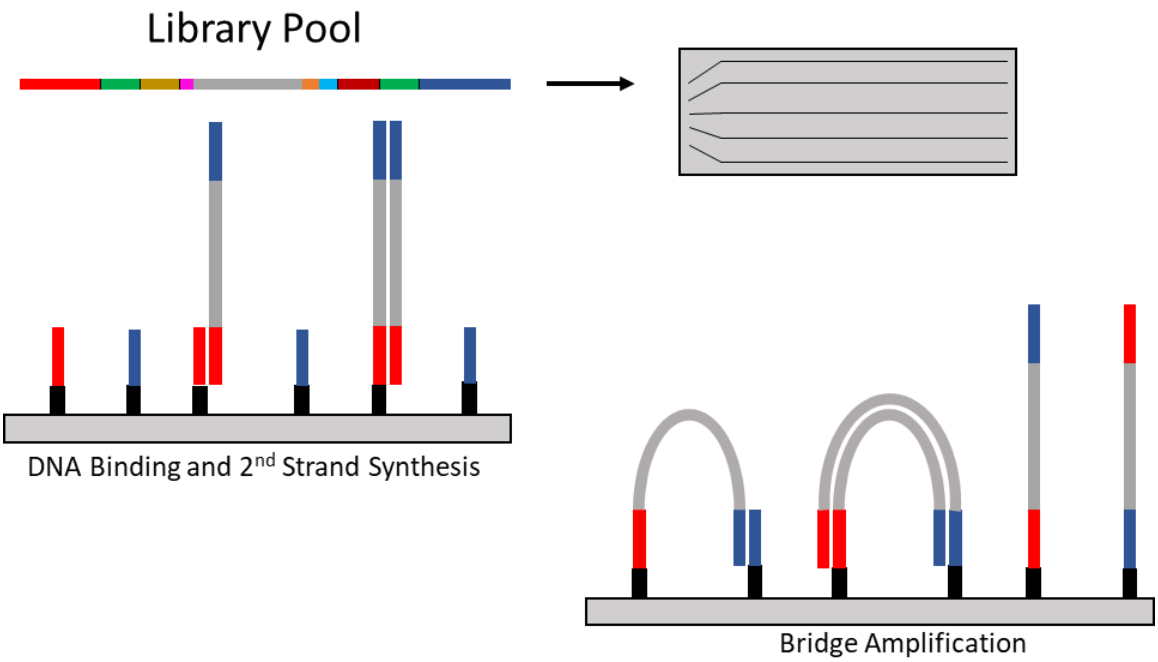

Figure 2.2 – Representation of bridge amplification and cluster generation. This process further increases the number of DNA fragments that are going to be available for sequencing. Additionally, this process allows for the DNA fragments to be sequenced in both directions. Inspired by Broad Institute.[114]

Sequencing starts with the addition of a sequencing primer that is complimentary to the adaptor sequences. Polymerases incorporate one of four different fluorescent reversible dye terminator nucleotides on to the new strand of DNA. Unincorporated nucleotides are washed away from the surface of the flow cell, a laser excites the fluorescent labels and an image is captured. Dispensation of successive chemical reagents unblock the dye terminators so that another nucleotide can be incorporated and cleave the fluorophore from that same nucleotide, so its signal is not recorded twice. At this point the

next cycle of nucleotide dispensation can occur over the surface of the flow cell (Figure 2.3).

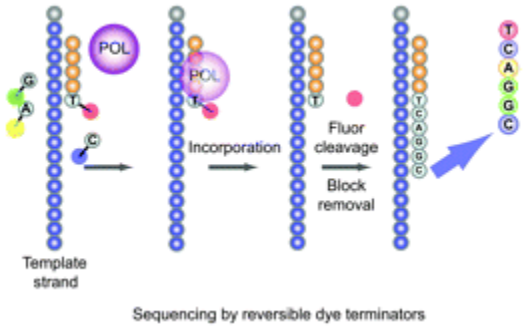

Figure 2.3 – Sequencing by synthesis reaction in Illumina sequencers. A polymerase binds to the template DNA:primer complex and begins incorporating one of four reversible dye terminators. The fluorescence is captured and then the fluorophore and block are removed from the nucleotide so that the next incorporation cycle can proceed. Reproduced with permission from Voelkerding and Dames, 2009.[115]

Each of the captured images are converted in software to a file format that translates the fluorescent wavelength recorded at each individual strand of DNA into a string of nucleotides that correspond to the strand of DNA that was sequenced. Advantages of newer sequencers from Illumina, including the MiSeq, are the ability to sequence both ends of the template DNA molecule. These paired-end sequencing reads provide positional information that aids the software to assemble fragments of DNA together and to align the consensus sequence to a known genome.[116] Another advantage is the concept of multiplex sequencing, which allows for DNA from multiple individuals to be sequenced simultaneously. This is achieved by incorporating unique molecular identifiers, UMI, or Indexes, during the adaptor ligation portion of library preparation. At the end of library preparation, multiple sample libraries can be pooled together in equimolar concentration and loaded on to the flow cell as a single sample. During data analysis, the index can be

read by the software and automatic demultiplexing of the samples aids in interpreting the results of the sequencing run.[117]

CHAPTER III – METHODS USED FOR DNA METHYLATION ANALYSIS

With the number of potential applications increasing for DNA methylation analysis, particularly in forensics, there have been large strides in innovation for new and more accurate procedures for interrogating CpGs in the genome. Although there are a number of methods currently available, the most commonly employed techniques rely on the differential chemical modification of cytosine residues in order to differentiate between methylated cytosines and unmethylated cytosines. For implementation into forensic laboratories, this approach is ideal as it can rely on the DNA that has already been extracted and quantified for traditional forensic DNA typing, using a portion of that extract to run concurrently with the other processes. Although the conversion of the DNA is not easily reversible, the original sample is not destroyed in this approach and analysts can testify that the results of the DNA methylation analysis is directly related to the results of the STR analysis.

A. Bisulfite Conversion of Methylated DNA

The vast majority of DNA methylation analysis techniques use PCR to amplify either specific targets or the whole genome. But, as previously discussed, DNA methylation is a post-replication process, as DNA polymerases do not differentiate between cytosine and 5-mC when synthesizing a new strand of DNA.[118] For this reason, a method to differentiate between methylated and unmethylated cytosine is necessary. In 1970 the reaction of cytosine residues, and their derivatives, in the presence of sodium bisulfite was first described.[119] This reaction is comprised of three steps (Figure 3.1). First, sodium bisulfite is added to a DNA sample resulting in a nucleophilic attack on the double bond

of cytosine residue between carbons 5 and 6. The presence of a methyl group on carbon 5 of 5-mC prevents this reaction from taking place on methylated cytosines. Next, in the presence of heat and a lowered pH, the cytosine sulfonate derivative undergoes hydrolytic deamination, becoming a uracil sulfonate derivative. The methylated cytosine, lacking a sulfite moiety, is left unchanged. Finally, raising the pH of the solution results in desulphonation, leaving an unmodified uracil where the unmethylated cytosine once was.[120]

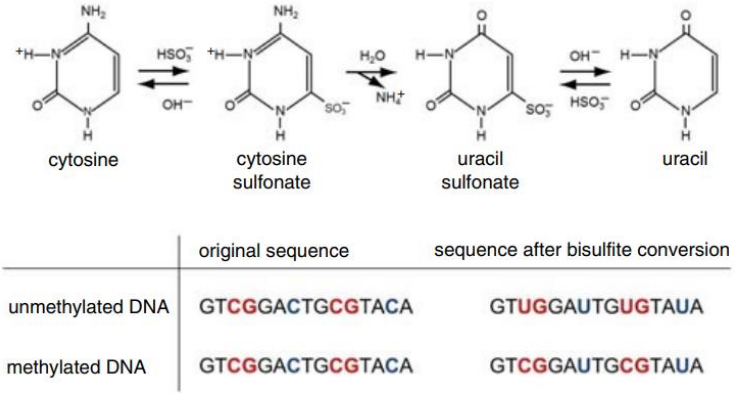

Figure 3.1 – Bisulfite modification of unmethylated cytosine. The resistance of methylated cytosine to nucleophilic attack by bisulfite allows for the differentiation of methylated and unmethylated cytosines in subsequent analyses. Reproduced with permission from Kristensen, Treppendahl, and Grønbæk, 2013.[121]

From this point, the bisulfite modified DNA can undergo PCR where the methylated cytosines will be paired with guanine, while the uracil will be paired with adenine and, in subsequent cycles, thymine. The resulting PCR products will be identical in length and sequence except at CpG sites, where methylated cytosine will be seen as just a cytosine and the unmethylated cytosine will be seen as a thymine. This differentiation is the foundation of the DNA methylation approaches that will be detailed below including High Resolution Melt (HRM) analysis, Methylation sensitive Single Nucleotide Primer

Extension (Ms-SNuPE), Matrix-assisted Laser Desorption/Ionization-Time-of-Flight Mass Spectrometry (MALDI-TOF MS), Pyrosequencing and Targeted Methyl Sequencing. Each method offers a number of pros and cons that forensic laboratories can evaluate for their own goals. Throughout this dissertation, pyrosequencing and targeted methyl sequencing are the main employed methods.

B. High Resolution Melt analysis

High Resolution Melt analysis represents one of the easiest and most cost-effective methods for DNA methylation analysis that forensic laboratories can utilize. The use of this method for methylation analysis, first described in 2007, relies on the difference in GC content of amplicons following PCR.[122] After bisulfite conversion, DNA samples can be amplified with unlabeled primers targeting a single CpG or several CpGs in close proximity to each other. The products of this reaction will consist of two types of PCR amplicons. The DNA molecules that were originally unmethylated will consist of thymine, base paired to adenine, at each CpG site. DNA molecules that were methylated will have a cytosine, base paired to guanine, at each CpG site. As described in Chapter I, the two hydrogen bonds between adenine and thymine require less energy to dissociate than the three hydrogen bonds between cytosine and guanine. This translates to a higher melting temperature for methylated strands of DNA than the unmethylated strands (Figure 3.2).

The method is performed in a single tube process in which a targeted region of the genome is amplified by a polymerase along with a dsDNA intercalating dye, like EvaGreen. The PCR must be carried out in a real-time thermal cycler with melt analysis capability. Immediately after amplification of the target region is complete, the

fluorescence of the PCR product is measured, and then the temperature in the instrument is gradually increased in 0.1 ºC intervals. After each increase in temperature, the fluorescence is measured. PCR products with lower GC content will melt at a lower temperature, and therefore the fluorescent signal will decrease. To visualize the data, the analysis software connected to the thermal cycler can automatically generate a graph of the negative first derivative in fluorescence per temperature vs the temperature (Figure 3.2).[123]

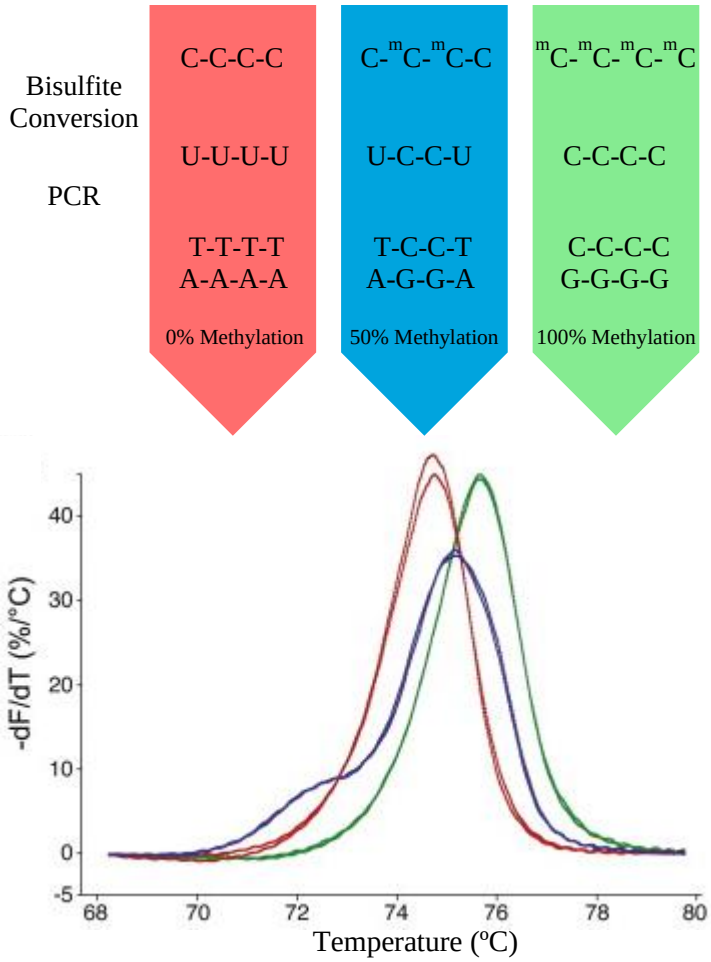

Figure 3.2 – Schematic representation of High Resolution Melt analysis for one DNA target with three levels of methylation. The melt peaks are higher for DNA strands with higher methylation/GC content. Reproduced with permission and inspired by Erali, Voelkerding, and Wittwer 2008.[124]

High Resolution Melt analysis provides a quick and relatively inexpensive way to probe the methylation of a particular region of DNA. The real-time PCR instrument is often already available in forensic laboratories, the method is nondestructive, and the benchtop techniques are nearly indistinguishable from routine forensic DNA sample preparation. Additionally, the PCR reaction requires only a forward and reverse unlabeled primer and a cheap intercalating dye. This means that the adoption of new and better optimized assays

within a laboratory can be accomplished without the high costs associated with many commercial kits.[125]

There are, however, a number of disadvantages. The technique is indirectly analyzing methylation status by observing the fluorescence of melting dsDNA. This means that the precise percent methylation at each CpG site is unknown. Indeed, the result for each sample is the melting temperature for the total number of CpGs that may be present in the amplified region. Furthermore, if the target region for differentiating two body fluids differs in methylation status by only 10%, the distance between the two peaks may not be large enough for reliable differentiation of the tissue types.[126] Although multiplexing is possible, there are limitations to the number of amplicons that can be analyzed in a single reaction, and therefore efficient identification of body fluid types would likely require several reactions, and therefore higher volumes of sample.[127]

C. Methylation sensitive Single Nucleotide Primer Extension

Methylation sensitive Single Nucleotide Primer Extension (Ms-SNuPE) is often referred to as a SNaPshot assay, and the method represents the methylation analysis method most similar to methods currently used in forensic laboratories, similar to Sanger sequencing.[128] Bisulfite modified DNA is amplified with unlabeled primers targeting a specific region for CpGs. The PCR product is then purified using Shrimp Alkaline Phosphatase (SAP) and Exonuclease I to degrade unincorporated dNTPs and primers, respectively. A second PCR is set up to incorporate the Ms-SNuPE primers just before the CpG of interest. The single base extension (SBE) reaction then incorporates a fluorescently labeled dideoxyribonucleoside triphosphate (ddNTP), with different emission wavelengths

for each of the four nucleotides, which then prevents the incorporation of any more nucleotides. This PCR product undergoes yet another purification by SAP to remove unincorporated ddNTPs, and the resulting sample can be analyzed by capillary electrophoresis.[129] Because each ddNTP is labeled with a different fluorescence color, methylation can be determined by comparing the ratio of the peak heights for the strands corresponding to C and T. This technique allows for quantitative analysis of DNA methylation from fragment analysis.

With this analytical approach, there is no additional instrumentation necessary for a forensic DNA laboratory. Data analysis is fairly straight forward and troubleshooting for assays would be performed in a manner similar to that used with STR kits. The disadvantages of this approach include the numerous reactions that are required prior to capillary electrophoresis which would reduce the laboratories throughput capabilities.[130] Other disadvantages include the difficulty in creating multiplex panels which require primers to create amplicons that are sufficiently different in size to allow for easy interpretation and relatively balanced peak heights in the electropherogram across multiple target regions. When targets are either hyper- or hypomethylated, the smaller peak could be lost in the baseline noise of the electropherogram, necessitating the need to have a much higher representation of that target in the assay versus other targets. This solution in turn can cause another problem, overloaded peaks in the electropherogram, which prevents the accurate calculation of peak height ratios.[131]

D. Matrix-assisted Laser Desorption/Ionization-Time of Flight Mass Spectrometry

The results of a MALDI-TOF MS run gives the specific mass determination of the DNA products in a sample, rather than the indirect analysis of DNA by fluorescent signal described in the previous methods. This means that the results give highly accurate recordings of the methylation content of any sample.[132] A MALDI-TOF MS experiment consists of combining DNA products with a suitable matrix that can be deposited on a sample holder. This DNA-matrix surface is ablated with a high-powered laser which pulses on the sample stage, ionizing the DNA-matrix surface. The ions are accelerated by an electric field through the flight tube analyzer, separating by charge and mass, until they reach the detector.[133] The collected mass spectra are collected and analyzed for fragment size and content, and a specialized software recreates the sequence of the PCR product as well as the methylation status of various CpG sites.[134] This technique offers one of the most precise methods for evaluating PCR products for their methylation content. Sample preparation is fairly straightforward and the ability to process samples back to back allow for reasonable throughput, although it would still be only one sample analyzed at a time. There is, however, one key issue with the implementation of this instrument in a forensic DNA laboratory. Mass spectrometers are large deviations from the traditional instrumentation in a forensic DNA laboratory and are incredibly expensive at over $200,000 for the instrument and up to $20,000 per year in maintenance. Their inclusion into a forensic DNA laboratory would require extensive training of personnel for testimony in a court of law, and for the price of just one mass spectrometer, several other instruments and machines could be purchased for the laboratory. For this reason, it is unlikely that the MALDI-TOF MS becomes a staple of forensic laboratories.

E. Pyrosequencing

Pyrosequencing, proposed as an alternative to Sanger sequencing, was first described in 1986, and realized in 1987, as a method to monitor in real time the release of pyrophosphate as a byproduct of nucleotide incorporation during a sequencing reaction of a PCR product.[135,136] The resulting signal represents the order of nucleotides in a strand of DNA as the strand of DNA is being synthesized, making pyrosequencing one of the earliest iterations of sequencing-by-synthesis (SBS). The set up for pyrosequencing is fairly straight forward but has several important requirements. First, the PCR amplicon needs to be immobilized in the reaction well. With recent instrumentation, this is primarily achieved by the use of streptavidin-coated magnetic beads in the reaction well binding to the biotin-labeled primers used in PCR amplification (Figure 3.3A). This PCR amplicon will contain what is referred to as the target region: the region to be sequenced. After immobilization, the amplicon will be denatured, and the non-labeled strand of DNA is washed away. The second requirement is the use of sequencing primer that is specific to the portion of the amplicon just before the target region. This sequencing primer, along with a polymerase, is what allows the target region to be amplified during the pyrosequencing process.[137] Pyrosequencing is so called because it is predicated on the release of pyrophosphate during replication that feeds an enzymatic cascade which results in the emission of light.

Modern instrumentation allows for significant automation of the pyrosequencing process. Throughout the present dissertation the Qiagen PyroMark® Q48 Autoprep pyrosequencer was used to automate the pyrosequencing process with the utilization of three separate reagent cartridges which dispense precise volumes of each reagent for each of the specified pyrosequencing reactions. The first cartridge contains the sequencing

primers and binding buffer to initiate the reaction. The second cartridge contains the denaturing solution, enzymes, substrates, and annealing buffer that are core to the pyrosequencing process. The third cartridge contains the four dNTPs that are dispensed in a specified order for the region of DNA that is being sequenced in the reaction well. (Figure 3.3B).

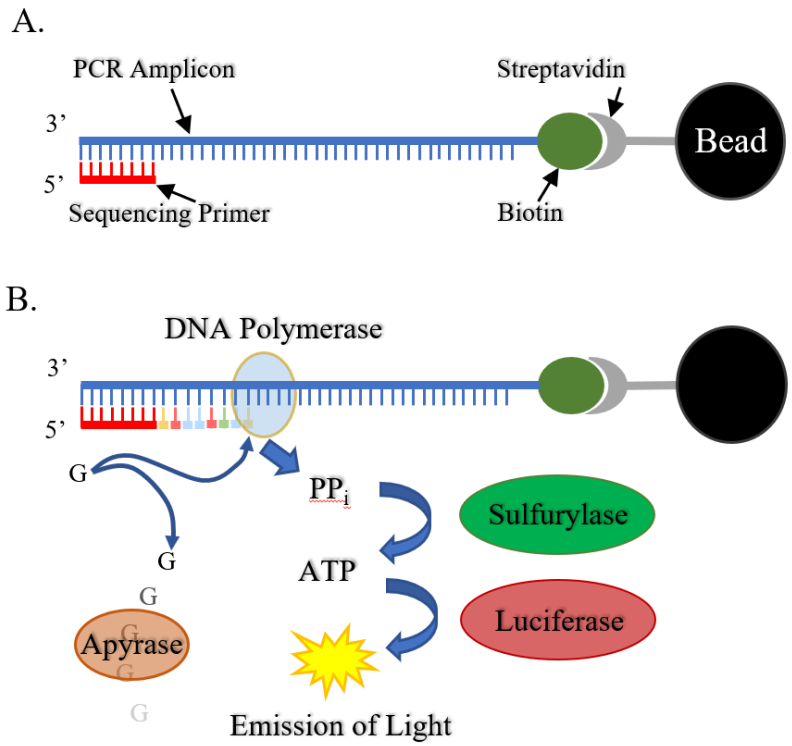

Figure 3.3 – Schematic representation of A) the process for capturing PCR amplicons to be analyzed via pyrosequencing and B) the enzymatic cascade that produces the light signal recorded for a pyrogram. Inspired by Diggle and Clarke, 2004.[138]

For pyrosequencing there are four enzymes (DNA polymerase, ATP sulfurylase, luciferase, apyrase) and two enzymes (adenosine 5' phosphosulfate (APS), luciferin). Each enzyme and substrate.[139] The DNA polymerase functions as normal, incorporating the dNTP on the new strand of DNA that is complimentary to the template strand of DNA. Upon dNTP incorporation, the enzymatic reaction cascade starts. First, the inorganic

pyrophosphate (PP_i) is released during dNTP incorporation, catalyzed by DNA polymerase, as follows:

$$(DNA)^n + dNTP \rightarrow (DNA)^{n+1} + PP_i + H^+$$

The incorporation of nucleotides releases an equimolar quantity of PPi. Pyrophosphate, in the presence of APS, is then converted to ATP by ATP sulfurylase, as follows:

$$PP_i + APS \rightarrow ATP + SO_4^{2-}$$

The newly created ATP then serves as a cofactor for the oxidative reaction of luciferin to oxyluciferin, catalyzed by luciferase, as follows:

$$ATP + Luciferin \rightarrow AMP + PP_i + oxyluciferin + CO_2 + h\nu$$

The monitored result of this enzymatic cascade is the production of light that is proportional to the number of nucleotides that were incorporated after each dispensation. This light is recorded by a charge coupled device (CCD) detector and displayed in the form of a pyrogram. The last enzymatic reaction to take place is the degradation of unincorporated dNTPs and ATP to adenosine monophosphate (AMP) and deoxyribonucleoside monophosphate (dNMP) so that the signal in the pyrogram can return to baseline before the next nucleotide is dispensed. This reaction, catalyzed by apyrase, is as follows:

$$ATP \text{ or } dNTP \rightarrow AMP \text{ or } dNMP + 4P_i$$

It should be noted that during the pyrosequencing process, the incorporation of adenine to the new strand of DNA is through the use of Adenosine-5'-(α-thio)-triphosphate (ATPαS) and not a regular ATP.[140] This modified nucleotide is used because it is still recognized by the DNA polymerase and apyrase, but not the luciferase. This means that the dispensation

of A during pyrosequencing will only result in a light signal if the dNTP is actually incorporated into the strand of DNA.

The results of a pyrosequencing reaction, displayed as a pyrogram, are graphed with the light signal in the form of relative light units (RLU) for each dispensation of a dNTP (Figure 3.4). The height of each peak will be proportional to the number of incorporated dNTPs, meaning that if the targeted sequence contains 3 Thymine followed by 1 Adenine, the peaks would be in a 3:1 ratio. Additionally, the nucleotide dispensation will include dead injections which are nucleotides that are known to be incorrect for that portion of the DNA. This servers as a negative control to verify that the correct fragment of DNA is being sequenced.

For DNA methylation analysis, the process of pyrosequencing remains unchanged, save for the inclusion of two characteristics. First, each CpG will be evaluated as a known variable position. The instrument will dispense a C followed by a T at each CpG site in order to keep both the methylated and unmethylated strands of DNA in sync during sequencing. Second, bisulfite control dispensations will be added to the nucleotide dispensation order. These controls consist of the attempted incorporation of a C at a point in the target region that is not a CpG and, therefore, is extremely unlikely to have been methylated prior to bisulfite conversion of the DNA. This bisulfite control is followed by a T dispensation. The presence of any signal at the C would indicate incomplete bisulfite conversion of the DNA sample and render the results of the methylation analysis inaccurate.[141]

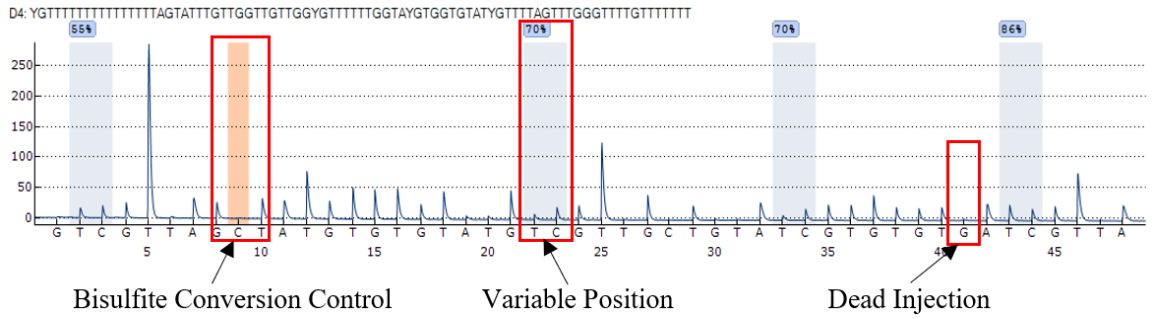

Figure 3.4 – Pyrogram of marker VE_8 for the identification of vaginal epithelial cells. The dispensation order can be seen above the pyrogram dictating the order in which nucleotides will be introduced to the pyrosequencing process. It includes nucleotides for the known sequence of DNA as well as injections for the bisulfite control, variable position, and dead injections. Each of these injections and the subsequent peaks are used by the pyrosequencing data analysis software to determine the quality of the data and the percent methylation observed at each CpG site.

For CpG analysis via pyrosequencing, the percent methylation at each CpG is calculated by comparing the peak heights of the C and T at each variable position to the established peak height that a single nucleotide would be expected to produce based on the rest of the peaks in the pyrogram. The first example of DNA methylation analysis via pyrosequencing for body fluid identification was in 2012 and since then there have been a multitude of other studies expanding on the list of body fluids that can be identified and various other applications, such as monozygotic twin differentiation and biological age determination.[63,142]

One of the promising advantages of pyrosequencing in forensic DNA laboratories is how time and cost effective the process is. Modern pyrosequencers are relatively inexpensive instruments that require very little maintenance and can be efficiently washed in between runs with simple deionized water. The commercial kits for these instruments allow for a fairly low per sample cost and the capacity for up to 48 samples in a single automated run allows for a high level of processivity for most laboratories [141].

Additionally, it will be demonstrated that, although not designed for multiplex PCR, it is possible to have multiple primers in a single PCR reaction followed by separate sequencing reactions. This approach, while still having some constraints, offers the advantage of probing a larger number of target regions without the need to consume too much DNA, which is often a limiting factor in forensic DNA laboratories. Additionally, multiplex PCR reactions allow for a larger amount of information to be conveyed versus multiple monoplex reactions.

F. Targeted Methyl Sequencing

Next Generation Sequencing (NGS), or Massively Parallel Sequencing (MPS), has been around long enough now that many more applications have been developed beyond just whole genome sequencing, and some are quite elegant and precise. For example, the major companies have, for years now, offered NGS library preparation kits that target for a wide selection of clinical diagnostics, like cancer screening panels, and for forensic applications. Research in the realm of DNA methylation has spurred manufacturers to also develop pre-defined Methylation Panels that can examine methylation sites related to the gene expression pathways of various diseases.[143] It is also possible to contract to create a panel that will target whichever methylation sites a customer wants and sequence them on an NGS platform.

Although several companies have developed their own versions of this technique, the main objective for all of them is the same: use bisulfite converted DNA as the template for targeted amplification of regions of interest and then sequence the amplicons.[144] The following methodology covers the specifics of the QIAseq Targeted

Methyl Panel as it is the library preparation kit utilized herein, but it should be noted that other proprietary techniques from the major companies are conceptually indistinguishable.

This library preparation begins with the DNA end repair of bisulfite converted DNA. The reason for the end repair is twofold: bisulfite conversion causes significant fragmentation of the DNA and the repair process allows for more efficient ligation of adapters in the second step.[145] The adapter ligation begins the process of making the template DNA truly prepared for targeted sequencing by making the DNA capable of being captured on the surface of the flow cell prior to sequencing. The adapter consists of three main sections: the unique molecular index (UMI), the sample index, and the homologous sequence for PCR and flow cell capture. The UMI consists of a 12-base design of alternating random and cytosine bases. This results in 4^8 possible UMI sequences per adapter resulting in each molecule of DNA in the sample receiving a different UMI sequence. The sample index consists of eight bases in a specific order. Every single adapter across the entire sample will have the same sample index. Finally the adapter contains a region of DNA that allows for hybridization to the DNA anchors on the surface of the flow cell and a small region that allows for non-specific primers to bind for various portions of library preparation and sequencing.[146] The result of adapter ligation is a sample of DNA where every single fragment contains a sequence that is specific to the whole sample and a separate sequence that is specific to just that molecule of DNA. This allows for the multiplexing of samples after library preparation and for quality control of the data resulting from sequencing. Any unincorporated adaptors are

removed in a purification step using QIAseq Beads, a magnetic bead that binds the DNA sample and allows for the removal of other components.

Next, target enrichment amplifies the desired regions of DNA using a primer that has one half that is specific to the DNA upstream of the region of interest and the second half that will be used as a primer binding region for the subsequent universal PCR. During target enrichment, the ligated DNA molecules undergo nine cycles of PCR using one gene-specific primer and a forward primer that is complimentary to the ligated adapter. A subsequent purification via QIAseq bead cleanup is performed to remove unincorporated primers. Finally, a universal PCR utilizes a primer that is complimentary to the second half of the gene-specific PCR primer, and containing the second index, and a universal primer. The addition of this second index is necessary for the process of bridge amplification and cluster generation prior to sequencing and further allows for the differentiation of specific samples after sequencing. After library preparation a quality control check of each library is conducted to evaluate the size distribution of the libraries as well as the concentration of DNA in each sample. The samples can then be diluted and combined in equal concentration in a single tube prior to being loaded on to the flow cell for sequencing. The sequencing process proceeds as previously described in Chapter 2.

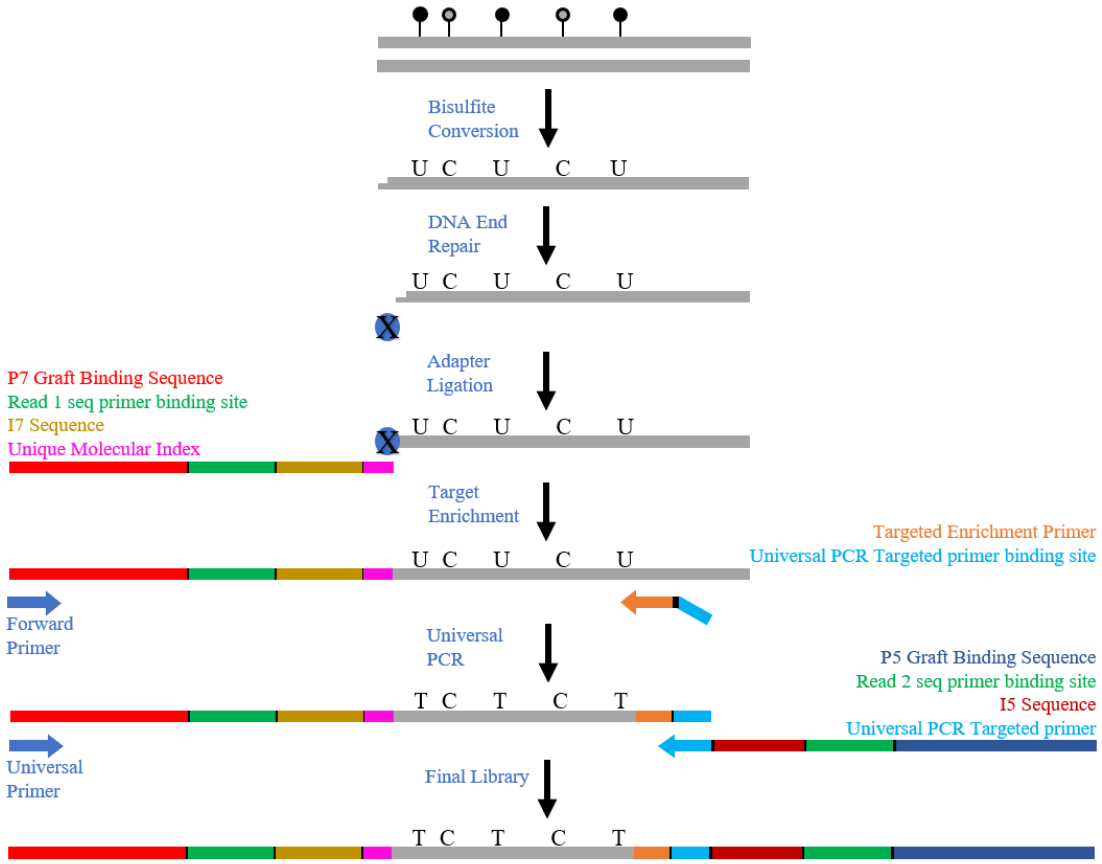

Figure 3.5 – Schematic representation of the library preparation for a QIAseq Targeted Methyl Panel. Inspired by Qiagen, 2019.[147]

E. Statement of the problem

As previously discussed, the presumptive and confirmatory tests for body fluid identification contain a multitude of drawbacks: false positives and false negatives, sample destruction, and tests are for designed for only one body fluid at a time. All of the currently employed serology tests rely on the presence of a protein in order to identify, either directly or indirectly, the body fluid. Given that there is a vast range of scenarios that can affect the samples prior to collection, this approach is not favorable. For example, the pH, humidity, and temperature of a crime scene can cause proteins to denature, which in turn causes them

to lose enzymatic activity.[148] If the proteins are no longer functional, many of the serological tests will not function properly.

With this in mind, much of the research of body fluid identification over the past two decades has been focused on the building blocks of proteins, mRNA. Assays evaluating mRNA as a body fluid identification method have seen great success with the ability to differentiate between body fluid cell types based on the differential expression of mRNA leading to proteins that are body fluid specific.[149] Research into mRNA for body fluid analysis has offered an attractive solution for forensic laboratories given that the samples can be handled in largely the same way that DNA samples are handled. Part of the sample collected at a crime scene undergoes automated extraction, quantification, PCR, and capillary electrophoresis with only minor changes from the protocols used for DNA.[150] Work in mRNA body fluid identification has produced assays capable of identifying saliva, blood, vaginal epithelia, and semen, as well as others like menstrual blood.[151–154] These assays have relied on a combination of real-time PCR, end point PCR, and capillary electrophoresis for the evaluation of the results. This means that forensic DNA analysts would have no problem at all incorporating these methods into their routine workflows for body fluid identification. There are, however, several drawbacks to the use of mRNA for body fluid identification in forensic laboratories. The RNA molecule, given its intermediate nature in the central dogma, is not a particularly stable molecule.[155] Also, the abundance of mRNA within a cell can vary greatly as a result of various physiological conditions, like disease or malnourishment.[156] Additionally a routine step in RNA analysis is the use of DNase I which completely degrades all DNA within a sample while leaving the RNA untouched.[157] It would be an unmitigated disaster for a forensic laboratory to have an

accidental contamination of DNase I in a lab that is primarily focused on human DNA analysis. There have been, however, reported protocols for the simultaneous extraction and separation of DNA and RNA from the same sample.[158] There has not, as of yet, been a huge appetite to incorporate the mRNA approach in to forensic laboratories given the fact that all considerations for body fluid identification would have to take place right at the beginning of sample accessioning. This could increase costs to laboratories that are chronically underfunded.

The use of DNA methylation is an alternate technique for body fluid identification that has received widespread attention. DNA methylation involves a covalent bond to cytosine, and its storage stability has been shown to last decades.[159] Body fluid identification via DNA methylation has produced a number of assays capable of identifying most of the forensically relevant body fluid types including saliva, blood, vaginal epithelia, semen, menstrual blood, and urine by targeting tDMRs.[61,63,64,160,161] One of the drawbacks in this area has been the need to develop cost effective methods for the multiplexing multiple body fluid identification assays together.

But it is not just body fluid identification that DNA methylation offers to the forensic community. Several lifestyle traits, like tobacco and alcohol consumption, drug use, and biological age can also be determined using the information contained within the methylation of various CpGs in the human genome.[60,162–166] And given that previous reports have indicated that age determining assays rely heavily on the knowledge of which body fluid the DNA originated from, it is only logical that these applications be analyzed simultaneously in a single experiment. This would offer forensic laboratories new and

exciting ways to increase the information that can be offered to law enforcement throughout the course of an investigation.

One approach that has attempted for years to break into the forensic community has been the use of Next Generation Sequencing platforms. With the upfront costs for these instruments and the per sample cost so high in the early years, most laboratories have been reluctant to implement massively parallel sequencing in routine DNA typing. Sequencing methods give forensic DNA analysts the opportunity to interrogate over 200 STR and single nucleotide polymorphisms (SNPs). The data can be used to identify unknown suspects when compared to databases as well as give phenotypic and biogeographical information about the unknown individual, such as hair color, eye color, and skin color.[167] Over the past few years the upfront cost of sequencers and the per sample cost of sequencing has come down dramatically, to the point that forensic laboratories are exploring ways to implement them in to their workflow, eliminating in large part the traditional PCR and capillary electrophoresis portions of the DNA analysis workflow. Additionally, Targeted Methyl sequencing library preparation kits are available that allow the creation of custom assays to interrogate a large number of CpG sites simultaneously.[168]

In Chapter 4 of this dissertation, we will discuss the creation and optimization of a body fluid identification multiplex via pyrosequencing that identifies saliva, blood, vaginal epithelia, and semen as body fluid sources. In Chapter 5, with the multiplex created, a developmental validation will examine the reproducibility of data, sensitivity of the assay, an inhibition study, a degradation study, and finally a mixture study. These tests will examine the robustness of the assay and its suitability for use in forensic settings. Chapter 6 will evaluate the use of statistical modeling to identify the body fluid origin of a sample

using methylation data from the multiplex without the need for human interaction, reducing the subjective bias of data interpretation. Finally, in chapter 7, we will discuss the preliminary construction of a targeted methyl sequencing assay run on an MPS platform for the simultaneous identification of body fluid and age determination to supplement the current assays offered for the forensic community on massively parallel sequencers. A forensic focused methyl assay for massively parallel sequencing would be yet another justification for the adoption of sequencing platforms into forensic laboratories.

CHAPTER IV – BODY FLUID MULTIPLEX VIA PYROSEQUENCING FOR SALIVA, BLOOD, VAGINAL EPITHELIA, AND SEMEN

The first task in creating a multiplex for body fluid identification and biological age determination was to start with what was already known in the literature. Given that biological age determination through DNA methylation is much more easily calculated with tissue-specific age assays, the creation of a body fluid multiplex was a natural starting point.[169] A number of DNA methylation markers for body fluid identification had previously been discovered, evaluated, and published within our research group. However, each marker had been studied in isolation; they were amplified as singleplexes and only evaluated for their ability to differentiate one body fluid in comparison to others. These markers were developed to identify saliva, blood, vaginal epithelia, and semen. To achieve the future goal of a single assay identifying both the body fluid origin and biological age, an initial body fluid identification multiplex was created and analyzed via pyrosequencing. Efforts were made to optimize this body fluid identification method. This included variations in PCR and sequencing primers, primer concentrations, input DNA for PCR, the number of magnetic beads for DNA capture prior to pyrosequencing, and the inclusion of formamide, a DNA denaturant, to increase the stringency of the sequencing primers. The following are the assays that were used for the construction of the initial multiplex with each genomic location data and surrounding features (UCSC Genome Browser GRCh37 – hg19).

A. Marker selection

For the identification of saliva, the BCAS4 marker described in Madi et al. was used.[63] This marker is named for the protein downstream, breast carcinoma amplified sequence 4 (BCAS4), which has been characterized to show overexpression resulting in tumor progression in breast cancer cell lines [170]. This assay, targeting Chromosome 20, is 158 bases long and consists of 7 CpGs sites, though only one CpG site, cg01997006, is included in the Illumina HumanMethylation450 beadchip array. This array is for probing the human genome for possible methylation markers and serves as the basis for the discovery of many new CpG markers for various applications. This marker was originally evaluated for its ability to identify semen, but upon further analysis, it was determined that saliva shows a much higher level of methylation across the various CpGs than in blood, vaginal epithelia, and semen [171].

For the identification of blood, the cg06379435 assay was used. The cg06379435 marker was first described by Park et al., but contained only the one CpG site.[161] It has no formal name as it has not been formally associated with a specific gene as it exists approximately 15 kilobases away from the Nuclear Factor I C (NFIC) gene which codes for a DNA-binding transcription activator.[172] In the developmental validation by Silva et al. the assay was expanded to include four more CpG sites in order to have a more complete methylation profile.[173] This assay, targeting a region of Chromosome 19, is 210 bases long and consists of 5 CpGs. This marker is characterized by hypomethylation in saliva, vaginal epithelia and semen and intermediate methylation in blood.

For the identification of vaginal epithelial cells, PFN3 A and VE_8 were examined. The PFN3 A marker, first described by Lee et al. in a methylation array study and then

characterized by Antunes et al. in a pyrosequencing assay, exists within CpG Island 82 on Chromosome 5.[174,175] This CpG island influences the transcription of Profilin-3 (PFN3), a protein that binds to actin, affecting stability of cytoskeletons, and may be associated with spermatogenesis.[176] There are nine CpG sites that are targeted in the 215-base long PFN3 A marker, a subregion of the full PFN3 CpG Island 82. This marker is characterized by the hypermethylation in saliva and blood, hypomethylation in semen, and intermediate levels of methylation in vaginal epithelia. In the initial stages of the multiplex development, it was determined that the PFN3 A amplicon was contributing to a number of issues that will be detailed later in this chapter. For this reason, it was later replaced by the VE_8 marker. The VE_8 marker was determined by Antunes et al. through the bioinformatics analysis of a data set produced by Park et al. in 2014.[161] This marker is 131 bases long and contains 4 CpG sites with cg08751438 being the CpG contained in the beadchip array that was initially identified. The lack of a formal name for this marker was because the CpG sites are nearly 20 kilobases downstream of LINC00197, a long noncoding RNA, and there were no other expression features nearby that could be influenced by the methylation of these CpG sites. This marker is characterized by the hypermethylation seen in saliva, blood, and semen, and intermediate methylation seen in vaginal epithelia.

Lastly, the ZC3H12D marker was used for the discrimination of semen from other body fluids. This marker originates from CpG island 41 in one of the introns of the Zinc Finger CCCH-Type Containing 12D (ZC3H12D) protein, also referred to as Monocyte chemo-tactic protein-induced protein 4 (MCPIP4), on chromosome 6.[177,178] This protein is possibly linked to cell growth regulation by ribonuclease 1 phosphorylation and some endonuclease activity in conjunction with the paralog ZC3H12A.[179] In the initial study by

Madi et al., this marker, at just 91 bases in length, contains 5 CpGs that are hypomethylated in semen, but hypermethylated in saliva, blood, and vaginal epithelia.

B. Standard Method

Buccal swabs, blood, vaginal swabs and semen samples were collected from volunteers under the conditions set forth under the approved protocol of IRB-17-0210 from Florida International University. Swabs were air-dried before being stored at -20 ºC or proceeding directly to extraction.

DNA extraction was performed either by manual or automated extraction protocols. The manual extraction involves the use of Phenol:Chloroform:Isoamyl alcohol and a separation filter as described in Appendix 1.[180] Automated extraction and purification were performed using the EZ1® DNA Investigator kit (Qiagen, CA) and the BioRobot® EZ1 automated purification workstation (Qiagen, CA) according to the manufacturer's specifications, detailed in Appendix 1. Samples were eluted in volumes of 40 µL Tris-Ethylenediaminetetraacetic acid (TE).

Quantification of DNA was performed using the ALU qPCR and Rotorgene thermal cycler method as described in Appendix 1. After concentration was determined, 200 nanograms of DNA were bisulfite modified using the EpiTect® Fast DNA Bisulfite Kit (Qiagen, CA) according to manufacturer's protocol, as detailed in Appendix 1. The elution volume after modification was 20 µL in order to achieve approximately 10 ng/µL concentration of bisulfite modified DNA.

The initial PCR primers and all variants were designed using the PyroMark® Assay Design software version 2.0 (Qiagen, CA). DNA amplification reactions were performed

using the PyroMark® PCR kit (Qiagen, CA) by adding 2 μL of bisulfite-modified DNA to each reaction according to manufacturer's protocol, which also specified a 0.2 μM final concentration for all PCR primers. A slight deviation from this protocol was made to scale up the final PCR volume to 45 μL, rather than 25 μL, in order to have enough volume for the subsequent pyrosequencing reactions. Primer sequences for the initial multiplex are specified in Table 4.1.

Table 4.1 – Panel of markers used in the initial multiplex. The reverse primer of each assay is the biotinylated primer.

Marker		Sequence
BCAS4	Forward	5'-AGT GGG TGA GGT TGT GAA ATG T-3'
	Reverse	5'-CCC ATC CTA CTA AAA CAT CTA ATT-3'
	Sequencing	5'-AGT TTT TTG GTG AAG TTT AT-3'
cg06379435	Forward	5'-AGT AGA GGT GGG GGT TAA TAA TT-3'
	Reverse	5'-CCA CAC AAC AAA ACA ACT ATC TCT-3'
	Sequencing	5'-GTT AGG AAA GAA AAA TGT AAT TTA-3'
PFN3 A	Forward	5'- GTG TAT AGT TTT GTT GAG GAT GTT TT - 3'
	Reverse	5' - ACA AAC ACA CCT TCC TAC AA – 3'
	Sequencing	5' - GTT TTG TTG AGG ATG TTT TT – 3'
ZC3H12D	Forward	5'-GGG TGA GGG TTT AAG GGT-3'
	Reverse	5'-CTC CCC TCA AAA CCT CAT-3'
	Sequencing	5'-GTT TTT GAG AAT TAT TTT TAA-3'

Pyrosequencing reactions were carried out on the PyroMark® Q48 Autoprep pyrosequencer (Qiagen, CA) with 10 μL of PCR product as the template for each of the four pyrosequencing reactions. Interpretation of the pyrogram and calculation of the percent methylation at each CpG was conducted through the PyroMark® Q48 Autoprep software (Qiagen, CA). The software evaluates the expected peak heights at each nucleotide dispensation versus what is observed in the pyrogram , and flags poor quality

data by issuing warnings to the user. These warnings include peak height deviations, peaks called at dead injections, suspected errors in dispensations and specific warnings such as high peak height deviations in the variable position which affect the accuracy of methylation calculations for each CpG. Warnings concerning variable positions in the pyrogram will cause the results to change color from blue, indicating good quality data, to yellow, indicating that additional scrutiny of the position is required by the analyst. If there are too many warnings issued in the variable position and the regions surrounding it, the results will be color coded red, indicating that the data is unreliable and should not be used for further interpretation.

C. Multiplex creation and optimization

The first attempts to develop an epigenetic multiplex were intended to determine if the four body fluid identification markers could be amplified together. To address this goal, the PCR primers for all four body fluid identification markers were used at a 0.2 μM final concentration in a PCR set up that with the final volume scaled up to 45 μL to accommodate the larger volume of PCR product needed to conduct the four subsequent pyrosequencing reactions necessary to sequence each locus. The initial results are shown below in Figure 4.1.

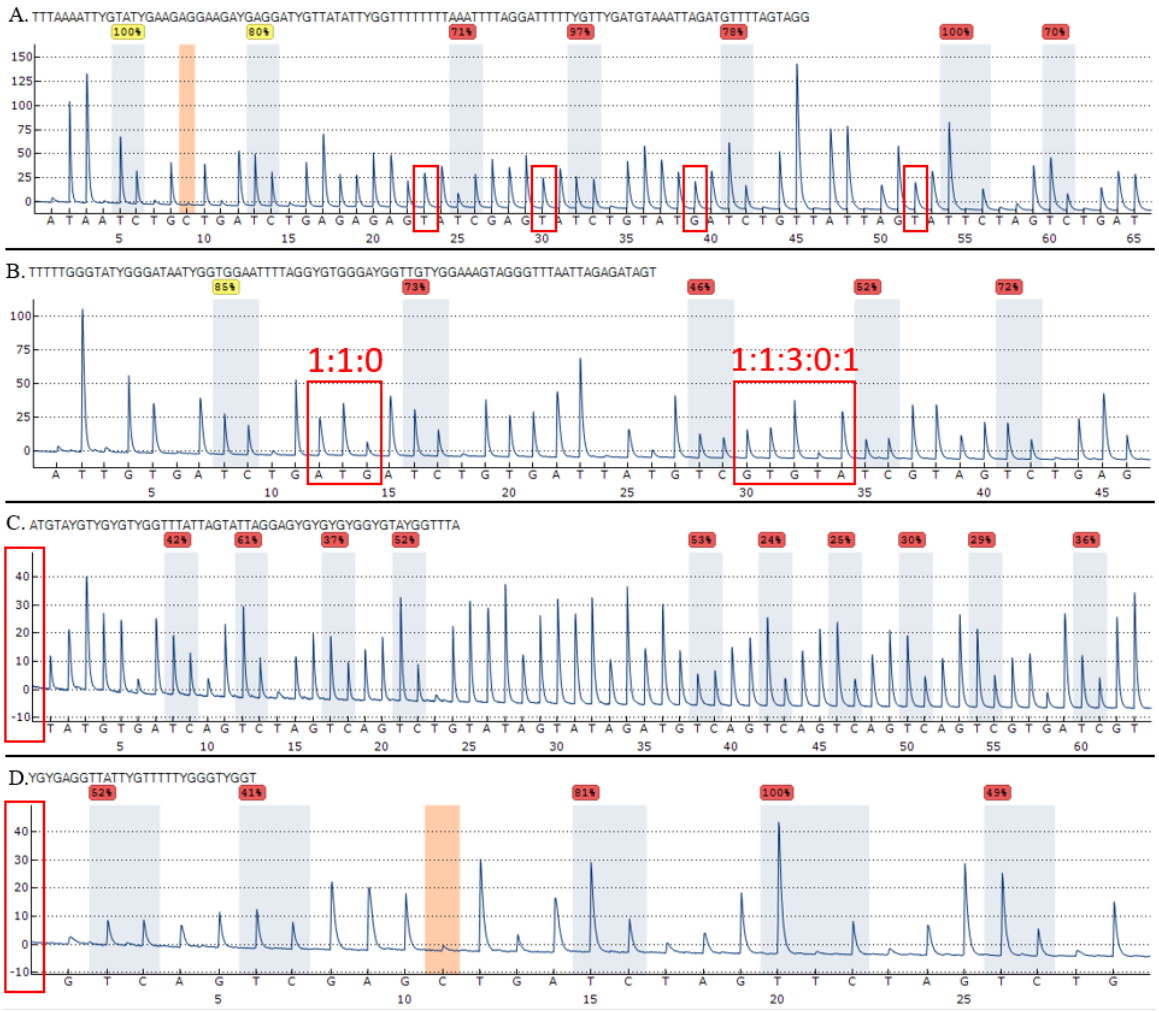

Figure 4.1 – Initial pyrosequencing results of the body fluid multiplex consisting of BCAS4 (A), cg06379435 (B), PFN3 A (C), and ZC3H12D (D). The results include peaks at locations where there should be no signal, incorrect peak height ratios when compared to the known sequence, low peak heights for the whole pyrogram and nearly all variable locations flagged red.

The initial results of the multiplex showed several deficiencies. Of primary concern was the quality of the pyrogram results. As seen in Figure 4.1A, the saliva marker contained additional peaks at dead injection locations. The inclusion of these peaks in the pyrogram caused the analysis software to misinterpret the height other nearby peaks resulting in a loss of accuracy of the methylation levels observed at variable positions. In Figure 4.1B, the peak height ratios were inconsistent with the expected peaks in the sequence for this blood marker. Similar to the previous issue, this prevents the software

from accurately interpreting methylation levels as the peak heights at each position is used to estimate the peak height equivalent to one nucleotide incorporated. In Figure 4.1C and Figure 4.1D for the vaginal epithelia and semen markers, the overall peak heights for the reaction were extremely low – approximately 10 Relative Light Units (RLU), making the data unusable. For reference, a monoplex reaction for any of these markers regularly shows RLUs of 50.

To improve these results experiments were designed targeting either the PCR or the pyrosequencing protocols. The main assumption involved the pyrosequencer, as typically only one PCR amplicon is added to the pyrosequencing reaction at a time, and therefore all of the DNA captured by the magnetic beads is the correct DNA template for the sequencing primer to bind. However, in this multiplex there were four different PCR products that were added to the reaction well. Because the magnetic beads do not distinguish between PCR products, there would be a competition for the PCR products to bind to a limited number of beads, resulting in a low recovery and poor balance in the capture of product. This was confirmed through analysis of the PCR markers run in 2% agarose gels, An experimental design was set up to optimize the PCR reaction and pyrosequencing by altering the concentration of PCR primers (0.2-0.6 μM), the concentration of $MgCl_2$ (from 1.5 mM to 2.5mM), the amount of DNA added to the reaction (up to 50ng), and the PCR primer sequences (using results obtained from PyroMark® Assay Design software version 2.0.) The goal in each case was to increase the representation of one marker, while not adversely affecting the resulting pyrograms for all of the other markers. Ultimately, the inclusion of additional $MgCl_2$ in the PCR set up, optimizing the forward and reverse PCR primer concentrations, and alteration of the

forward primer for the cg06379435 marker produced an increase in the concentration of PCR product, as demonstrated by increased peak heights in the resulting pyrograms.

A second issue to be resolved was indiscriminate binding of the sequencing primer. Because the multiplex contained four PCR products, the stringency of the sequencing primer was not sufficient to ensure that the primer would only bind to the correct template DNA. Several of the sequencing primers were capable of partially binding to random parts of the PCR products. Then, once pyrosequencing started, the signal produced was nonsensical. Figure 4.2 demonstrates that the specificity of binding of the sequencing primer to its target was insufficient and caused problems with the sequencing readout. In this figure, each of the markers was amplified individually and then each was sequenced with the BCAS4 sequencing primer. The resulting pyrograms indicate that this sequencing primer was quite capable of binding to the three incorrect PCR products and in the case of PFN3 A, creating large peaks that interfered in the interpretation of the data.

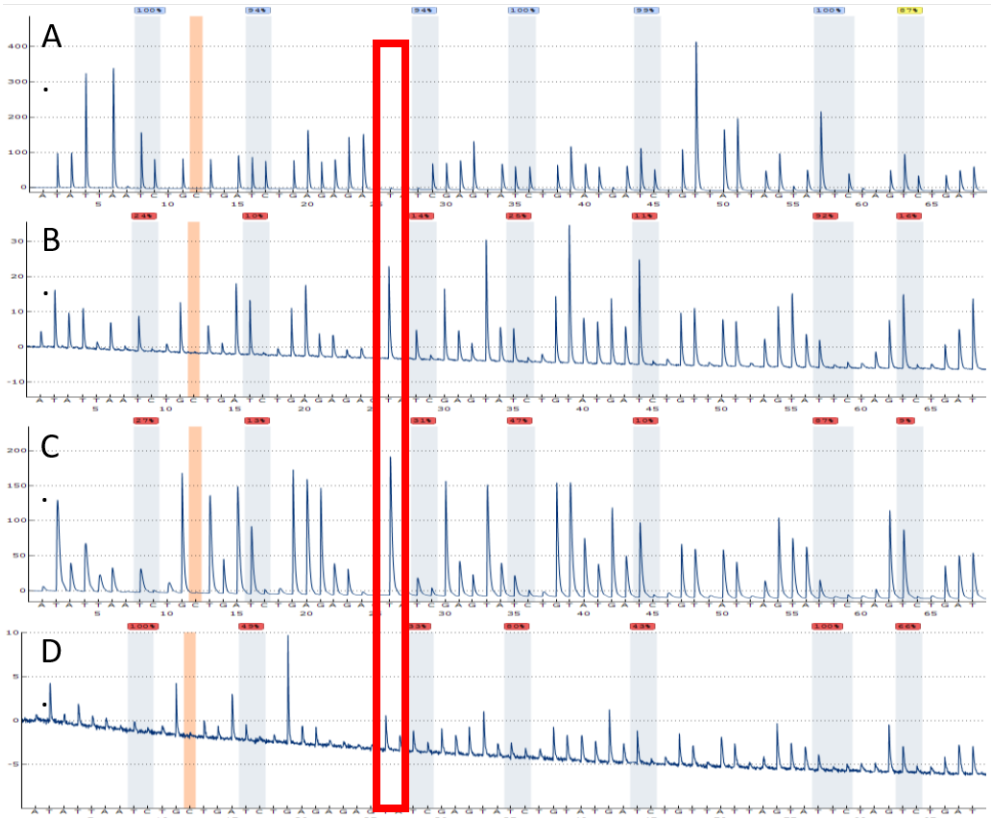

Figure 4.2 – Pyrograms resulting from amplifying BCAS4 (A), cg06379435 (B), PFN3 A (C), and ZC3H12D (D) in monoplex and then sequencing using the BCAS4 Sequencing Primer. Red bar indicates one of the interfering peaks seen in BCAS4 pyrograms that is a result of the sequencing primer improperly binding to other PCR products.

To minimize this problem, additional sequencing primer variants were created and evaluated for their ability to produce strong signals for their respective targets, but not bind other PCR products in the reaction well. Of the sequencing primer variants tested, only BCAS4 Sequencing Variant 1 produced an increase to the results of the pyrograms when evaluated in the presence of multiplex PCR product. Unfortunately, there were still a number of interfering peaks and improper peak height ratios occurring throughout the four pyrograms. To further reduce the probability of a sequencing primer from binding to the incorrect PCR product, the addition of formamide to the sequencing primers was explored for its known effects increasing the stringency of the primer binding. The unbound primers

would then be washed away prior to the actual start of sequencing and not interfere with the sequencing reaction. The results of the formamide experiments demonstrated that using up to 90% formamide in BCAS4, PFN3 A, and ZC3H12D sequencing assays did not significantly reduce the peak heights when compared to the monoplex results, but significantly reduced the peak heights of assays in which the incorrect sequencing primer was present. Figure 4.3 shows the difference between 0 and 90% formamide in the BCAS4 sequencing primer when combined with the BCAS4 PCR product (4.3A and 4.3B) and the PFN3 A PCR product (Figure 4.3C and Figure 4.3D). Similar results were observed in all other combinations of sequencing primer and PCR products, except that the cg06379435 marker suffered a decrease in pyrogram peak heights when over 40% formamide was used. Ultimately, the protocol included 90% formamide in the BCAS4, PFN3 A, and ZC3H12D sequencing primers and 40% formamide in the cg06379435 sequencing primer.

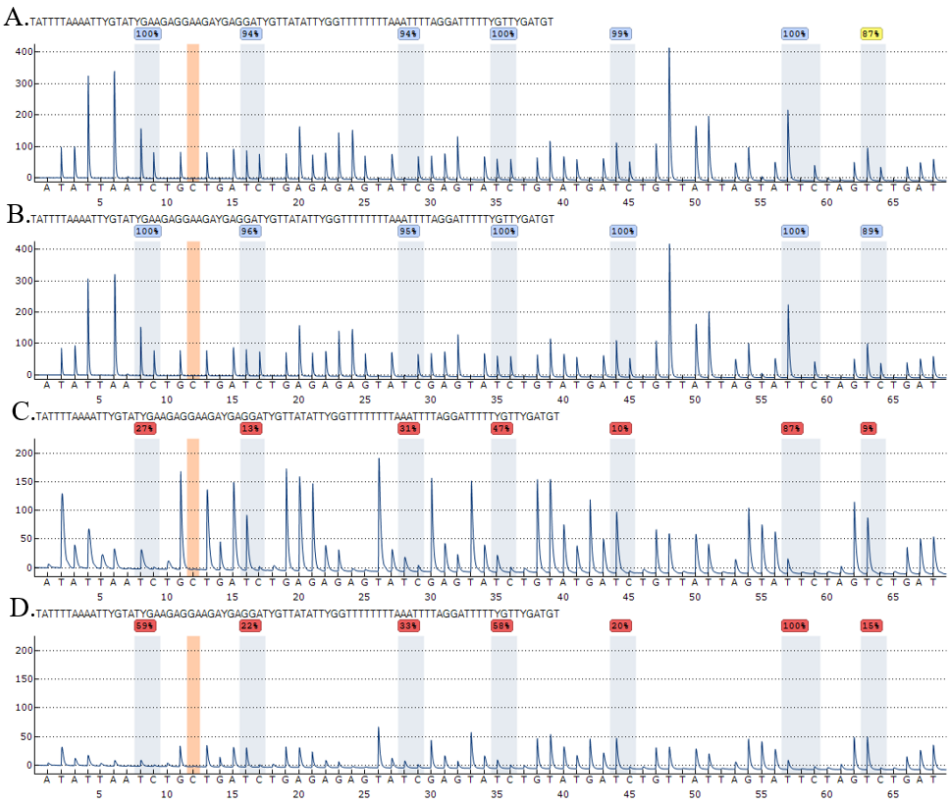

Figure 4.3 – Effects of the inclusion of formamide in the sequencing primer solutions. Peak heights are not significantly affected in the sequencing of BCAS4 PCR product when using 0% formamide (A) and 90% formamide (B) in the BCAS4 Sequencing primer. For the PFN3 A PCR product, the decrease in peak heights from 0% formamide (C) and 90% formamide (D) is quite noticeable.

The end result of these modifications is seen below in Figure 4.4. The changes made to the concentration and designs of the PCR primers and the $MgCl_2$, resulted in a semi-functional body fluid identification multiplex. The BCAS4 and ZC3H12D markers had improved recorded peak heights. Additionally, the use of a new sequencing primer for BCAS4 and the presence of formamide in all of the sequencing primers had a significant effect on reducing the number of CpG sites that the software deemed unusable. Ultimately all of these efforts did not sufficiently increase the peak height of the PFN3 A marker. Additionally, it was obvious from the formamide and monoplex reactions, that the PFN3A PCR product was responsible for the interfering peaks that remained, even if they were

below the threshold for the software to flag them. For this reason, a switch to a new vaginal epithelial marker, VE_8, was made.

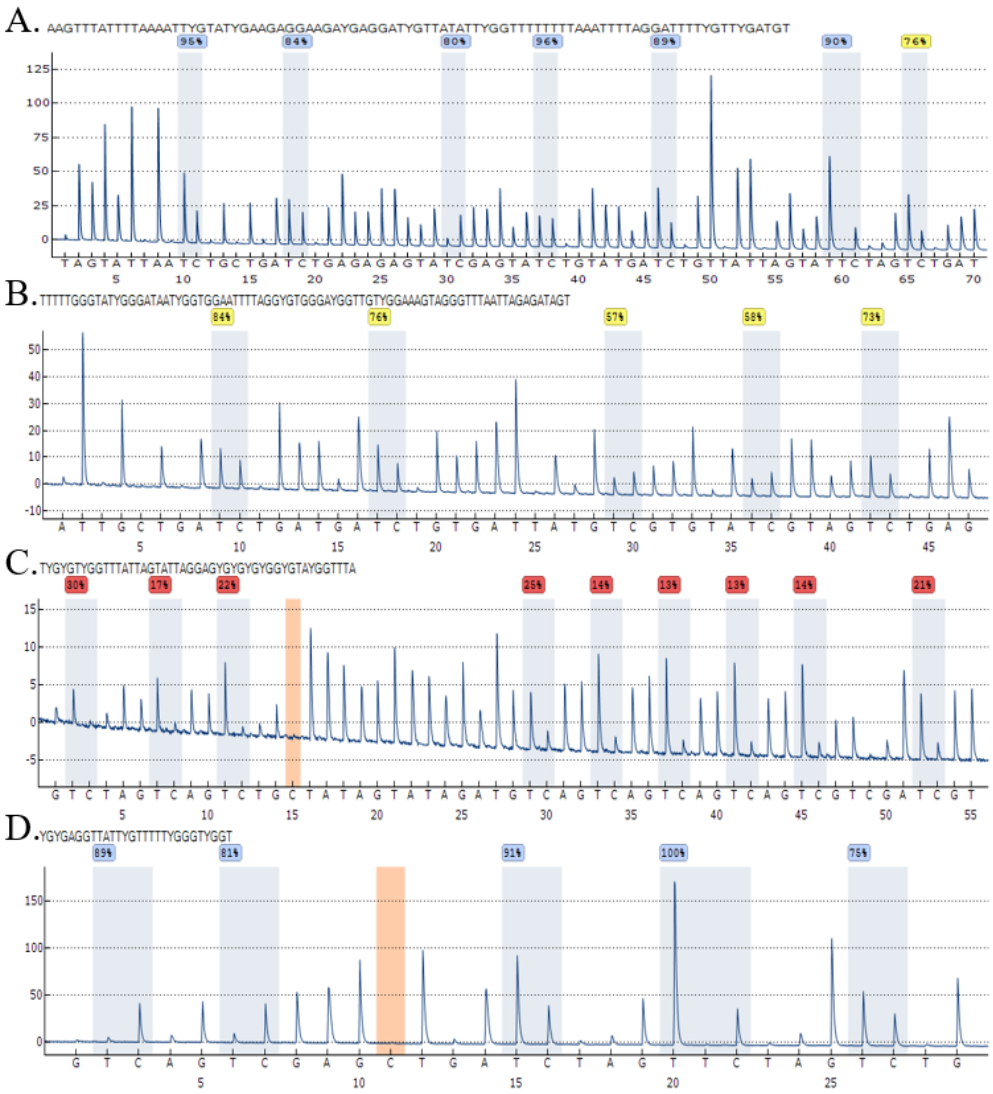

Figure 4.4 – Results of optimizations made to the multiplex containing BCAS4 (A), cg06379435 (B), PFN3 A (C), and ZC3H12D (D). The optimizations have resulted in acceptable peak heights for the CpG sites in BCAS4 and ZC3H12D. Cg06379435, while having low peak heights was also usable. Unfortunately, the results for the PFN3 A locus were subpar.

The VE_8 marker, identified by Antunes et al., was incorporated into the multiplex in place of PFN3 A, and immediately resulted in a significant improvement to the overall process. The removal of the PFN3 A PCR primers eliminated many of the secondary PCR

products and improved the recovery of the other PCR products and sequences. This was observed as a global increase to the observed peak heights in the pyrogram of each marker. Additionally, the use of formamide (90%) in the sequencing primer of VE_8 reduced incorrect primer binding while not affecting the overall peak heights of the assay. The only drawback to this switch was the emergence of interfering peaks for the BCAS4 marker in the second half of its pyrogram, These peaks were later identified as being caused by a secondary PCR product created by the combination of the cg06379435, VE_8, and ZC3H12D PCR primers. To counteract these effects, the primer concentrations were reduced across the board so that the erroneous PCR products influence was less noticeable. Additionally, a decision was made to cut the size of the BCAS4 marker in half and focus the pyrosequencing reaction on the first four CpGs. The finalized multiplex PCR mixture and primer sequences with concentrations (Table 4.2 and Table 4.3) are shown below along with an example of the final multiplex (Figure 4.5).

Table 4.2 – Finalized PCR setup using PyroMark® PCR kit. Volumes listed are for one sample.

PCR Master Mix	Volume(µL)
Pyromark Master Mix, 2x	22.5
Coral Load, 10x	4.5
Primer mix, 10x	4.5
$MgCl_2$	1.08
H_2O	10.42
Sample DNA	2
Total	45µL

Table 4.3 – Sequence of PCR and sequencing primers used in the final multiplex. The reverse primer of each assay is the biotinylated primer.

Marker		Sequence	Final Concentration (µM)
BCAS4	Forward	5'-AGT GGG TGA GGT TGT GAA ATG T-3'	0.2
	Reverse	5'-CCC ATC CTA CTA AAA CAT CTA ATT-3'	0.15
	Sequencing	5'-AGTTTAATAGTTTTTGGTG-3'	4
cg06379435	Forward	5'-AGT AGG GGT TTA GGT TAT GTT ATT GT-3'	0.175
	Reverse	5'-CCA CAC AAC AAA ACA ACT ATC TCT-3'	0.135
	Sequencing	5'-GTT AGG AAA GAA AAA TGT AAT TTA-3'	4
VE_8	Forward	5'-GTT TTA AAT TAG GGT GTG GGT AGA G-3'	0.11
	Reverse	5'-CAT ACC AAA AAA ACA AAA CCC AAA CTA-3'	0.105
	Sequencing	5'-AGA GTT GTG TTT TTT TTG GA-3'	4
ZC3H12D	Forward	5'-GGG TGA GGG TTT AAG GGT-3'	0.165
	Reverse	5'-CTC CCC TCA AAA CCT CAT-3'	0.165
	Sequencing	5'-GTT TTT GAG AAT TAT TTT TAA-3'	4

The decrease in the concentration of the PCR primers in the finalized multiplex caused the overall peak heights to be lower than the initial multiplex experiments with the VE_8 marker. However, the pyrograms still had peak heights well over the threshold for the software. Although there were still two interfering peaks present in the BCAS4

pyrogram, they did not cross the threshold to be considered real peaks, and therefore they were not considered in the calculation of peak heights and percent methylation.

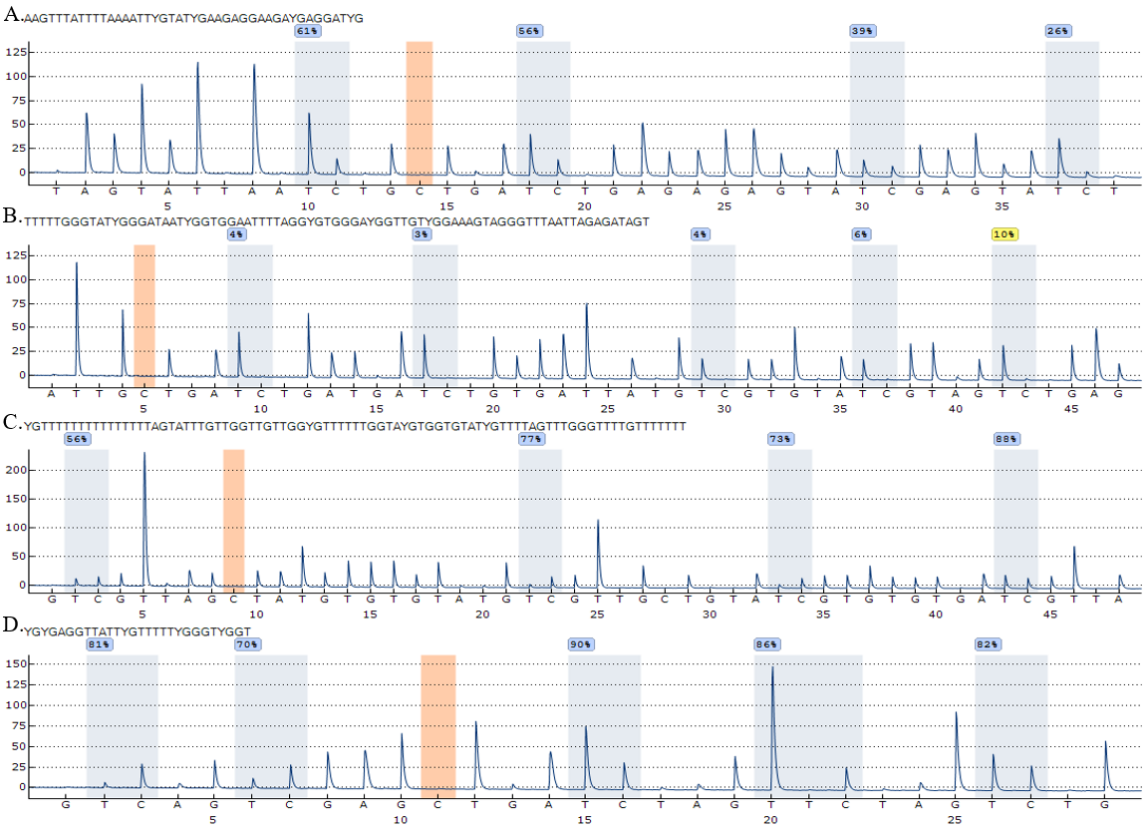

Figure 4.5 – Pyrograms of the finalized multiplex consisting of BCAS4 (A), cg06379435 (B), VE_8 (C), and ZC3H12D (D).

To verify the reproducibility of the multiplex, 10 samples of sample, blood, vaginal epithelia, and semen, were analyzed and the results are shown in Figure 4.6. The observed methylation values in the multiplex were reproducible and produced means and standard deviations consistent with the literature values of the markers in monoplex.

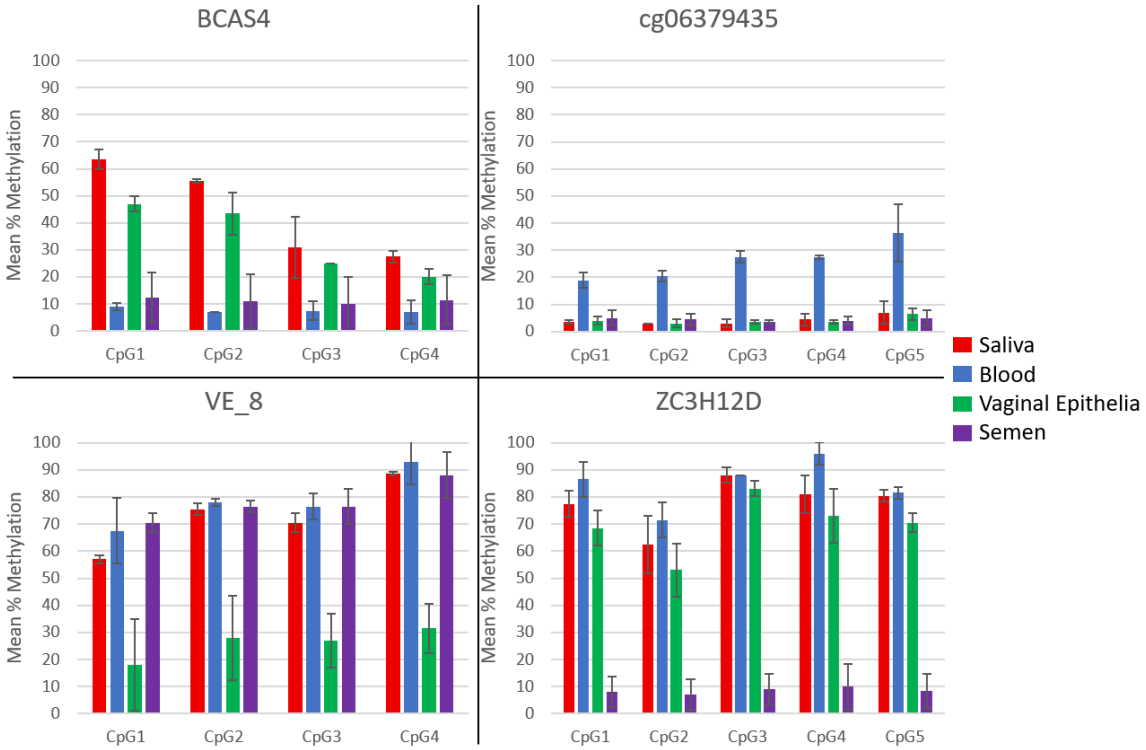

Figure 4.6 – Graph showing the mean % methylation and standard deviation for samples of saliva (n=10), blood (n=10), vaginal epithelia (n=10), and semen (n=10). Observed methylation values in the multiplex were consistent with the values in the literature for monoplex reactions.

D. Concluding Remarks

A body fluid identification multiplex was developed using pyrosequencing that was optimized to give reproducible results across 4 different sample types. This was the first body fluid identification multiplex via pyrosequencing reported in the literature.[181] The results demonstrate that although the pyrosequencing process was intended to accept only a single PCR product at a time, it is possible to run a multiplex on the pyrosequencing platform, through optimization of experimental conditions. The advantage of this multiplex approach is that it reduces the total amount of sample consumed throughout the analytical process by requiring a single amplification instead of four, increasing throughput and

providing the ability to conclusively identify several different body fluids in a single assay. The strengths of this assay will be explored further in Chapter 5 through a developmental validation, and in Chapter 6 the statistical evaluation of this multiplex's capability to positively identify a body fluid will be explored.

CHAPTER V – DEVELOPMENTAL VALIDATION OF THE BODY FLUID IDENTIFCATION MULTIPLEX

The identification of a body fluid during forensic investigations can give important context and clues to help elucidate the series of events that occurred at a crime scene. The creation of a body fluid identification multiplex offers a powerful tool for forensic investigators because it gives a confirmatory test to analysts to be able to identify several of the most commonly found body fluids at crime scenes. However, as with all tests, the extent of the capabilities of the assays need to be explored. To verify the efficacy of the multiplex, a developmental validation consisting of several studies was performed to determine the limitations of this assay. These validations include a population study to verify the reproducibility of the multiplex results, a sensitivity study to determine the minimum concentration of DNA for use in the multiplex, a mixture study to determine the ability of the assay to detect multiple body fluid present in a sample, and inhibition and degradation studies to gauge the robustness of the approach.

Studies of body fluid specificity and species specificity were not conducted because Silva et al. had already detailed the body fluid and species specificity of these markers in previous works using monoplex amplifications. Thus there was no expectation that the use of multiplex amplification would alter the results of those studies.[173] To obtain body fluid specificity, the multiplex was optimized to reduce interfering signals across the different probed CpG sites in order to produce results similar to those obtained when samples were amplified in monoplex. For species specificity, Silva et al. compared a human blood sample to DNA samples of dog, cat, mouse, chicken, bovine, equine, pig, chimpanzee, orangutan, gorilla and a microbial pool consisting of *Escherichia coli*, *Staphylococcus aureus*,

Enterococcus faecalis, and *Pseudomonas aeruginosa*. These species represent a large extent of the genomic material that might be present in evidence samples and could cause confusion in the interpretation of results. The results of the species specificity showed that certain non-human primates produced pyrograms similar to humans, while the other more commonly found animals did not produce any results. This is largely to be expected as primates are evolutionarily much closer to humans than the other species targeted.[182]

Implementation of a body fluid identification multiplex in forensic laboratories offers a confirmatory method that could increase the evidentiary value of any single piece of evidence. However, the assay needs to still be forensically relevant, i.e. reproducible with low input DNA, resistant to degradation and inhibition, and, if possible, offering the ability to detect DNA mixtures. The validation study will explore each of these points to determine the suitably of this assay for its intended purpose.

A. Methods and Materials

Buccal swabs, blood, vaginal swabs and semen samples were collected from volunteers under the conditions set forth under the approved protocol of IRB-17-0210 from Florida International University. Swabs were air-dried before being stored at -20 ºC or proceeding directly to extraction.

DNA extraction was performed either by manual or automated extraction protocols. The manual extraction involves the use of Phenol:Chloroform:Isoamyl alcohol and a separation filter as described in Appendix 1.[180] Automated extraction and purification were performed using the EZ1® DNA Investigator kit (Qiagen, CA) and the BioRobot® EZ1 automated purification workstation (Qiagen, CA) according to the manufacturer's

specifications, detailed in Appendix 1. Samples were eluted in volumes of 40 μL Tris-Ethylenediaminetetraacetic acid (TE) buffer.

DNA Quantification was performed using the ALU qPCR and Rotorgene thermal cycler method as described in Appendix 1. After the concentration was determined, 200 nanograms of DNA were bisulfite modified using the EpiTect® Fast DNA Bisulfite Kit (Qiagen, CA) according to manufacturer's protocol, Appendix 1. The elution volume after modification was 20 μL in order to achieve approximately 10 ng/μL concentration of bisulfite modified DNA.

Multiplex PCR reactions were carried out using the PyroMark® PCR Kit (Qiagen, CA) on a GeneAmp® PCR System 9700 (Applied Biosystems, Foster City, CA). All samples were prepared according to the PCR master mix specified in Table 4.2 and the thermal cycling parameters specified in the manufacturer's protocol (Qiagen, CA). After amplification 10 μL of PCR product was used for each of the four pyrosequencing reactions corresponding to the body fluid assays. Pyrosequencing was carried out on the PyroMark® Q48 Autoprep pyrosequencer (Qiagen, CA) following the manufacturer's protocol, but with the addition of formamide in the sequencing primers as previously described. Following pyrosequencing, the percent methylation at each CpG site was automatically calculated using the PyroMark® Q48 Autoprep software and the results were displayed as a pyrogram. The CpG sites for each of the body fluid markers in the multiplex were analyzed, and the mean and standard deviation for each body fluid was calculated. To compare the means observed in the population study to the literature value, a T-test Assuming Unequal Variance was used with a p-value of 0.05. For the sensitivity studies, a

one-way analysis of variance (ANOVA) test with a p-value of 0.0083 was used to determine if the differences between DNA input levels was significant.

B. Validation Studies

To generate a database of characteristic methylation values for each CpG in the multiplex, a population study was conducted consisting of approximately 30 samples, using 10-20 ng input DNA for each sample, from each body fluid type (saliva, blood, vaginal epithelia, and semen). The pyrograms for each sample were analyzed and compared to ascertain whether any of the body fluids produced statistically similar results that would negatively impact their ability to determine body fluids. Additionally, the results of the multiplex amplifications were compared to the literature values for the markers in monoplex amplification to determine if the differences in approach caused any significant changes in the methylation values observed.

As the Scientific Working Group for DNA Analysis Methods (SWGDAM) guidelines describe, the evaluation of any test's limits regarding DNA input is necessary to evaluate the reliability of results. Several publications have detailed the sensitivity of the markers used in the body fluid multiplex, including an assessment of input DNA levels ranging from 500 ng to 1 ng.[63,161,173] However, these sensitivity studies were carried out on the PyroMark® Q24 pyrosequencer (Qiagen, CA). Therefore it was necessary to examine results with the upgraded system used in these studies, the PyroMark® Q48 Autoprep pyrosequencer (Qiagen, CA) In this sensitivity study the body fluid multiplex was tested with the following DNA inputs: 20ng, 10ng, 5ng, 2ng, 1ng, 500pg, 250pg, 100pg. One

sample of each body fluid was amplified in five replicates to assess the accuracy of the results.

As was seen in previous studies, the mixture of body fluid types in the same sample produced intermediate methylation values at each CpG which varied with the relative concentration of each cell type.[173] To detect the effect of mixtures, samples of DNA from either saliva, blood, vaginal epithelia or semen were combined to produce 6 different mixture types (Saliva/Blood, Saliva/Vaginal Epithelia, Saliva/Semen, Blood/Vaginal Epithelia, Blood/Semen, Vaginal Epithelia/Semen) at three different ratios (75:25, 50:50, 25:75).

In order to evaluate the body fluid multiplex's ability to produce results in samples that may contain inhibitors or have been degraded, several mock samples were created and tested. To assess the effects of inhibition, two well characterized inhibitors were used: hematin and humic acid.[183] Samples of control DNA (10ng/μL) were combined with the inhibitors (hematin 0.08M and humic acid 0.24mg/mL) either before or after bisulfite conversion. For degradation, samples of control DNA (EpiTect PCR Control Methylated Converted, Qiagen, CA) were heated at 95 ºC for 14, 20, and 25 minutes to induce DNA fragmentation and compared to a sample that was just incubated at room temperature [183].

In the work by Wang and McCord, heating DNA samples at 95 ºC caused intact DNA (15 kilobases) to fragment to approximately 200-700bp in length when heated for 10-25 minutes. The extent of fragmentation was further enhanced by reducing the input DNA, so forensically relevant (sub 20ng) input DNA to PCR reactions should be affected.

C. Results and Discussion

Population Study

For the population study, over 120 samples were tested in the multiplex and the resulting methylation patterns for each marker in the multiplex can be seen in Figure 5.1. A comparison of the multiplex and monoplex data are shown in Table 5.1. Initially, the data sets appear to show the same trends across body fluids and markers. The data was evaluated first by comparing the means with a student's *t*-test to evaluate if there were statistically significant differences in observed means.

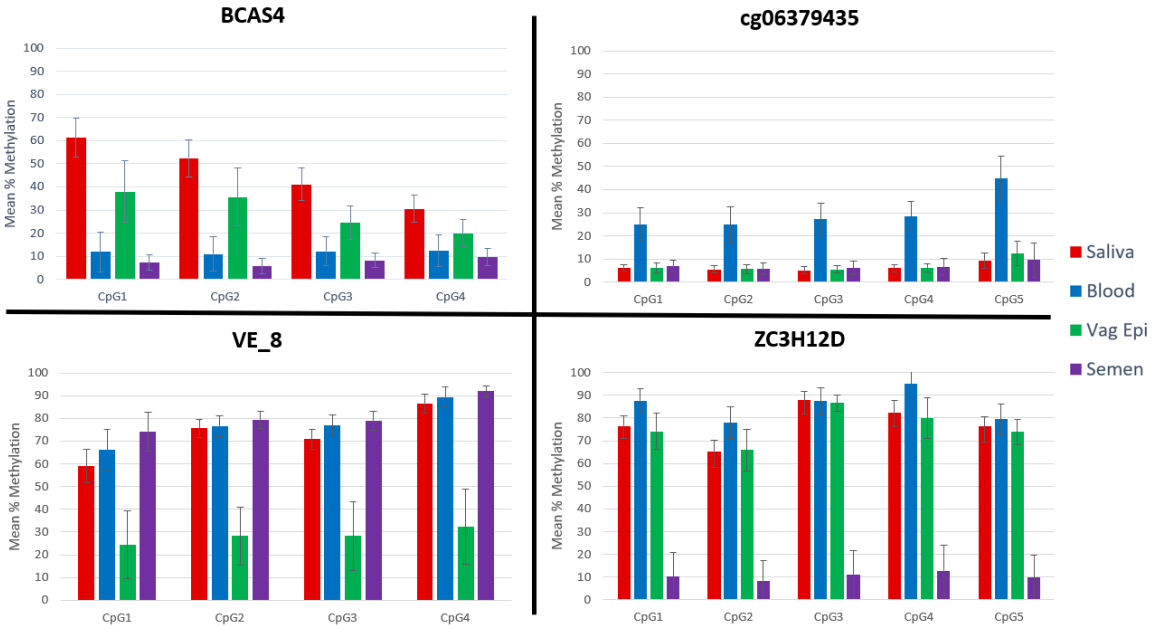

Figure 5.1 – Mean percent methylation values observed for saliva (n=38), blood (n=32), vaginal epithelia (n=26), and semen (n=28) when amplified in the body fluid identification multiplex. Error bars are one standard deviation.

Table 5.1 – Methylation profiles of saliva (n=38), blood (n=32), vaginal epithelia (n=26), and semen (n=28) when tested in the multiplex and compared to the methylation profiles of these markers when tested in monoplex, according to literature values. The values for BCAS4 CpG2 and CpG3 are not reported in the literature.

Marker	Body Fluid	CpG (Mean % Methylation ± SD)									
		CpG1		CpG2		CpG3		CpG4			
		Multiplex	Monoplex	Multiplex	Monoplex	Multiplex	Monoplex	Multiplex	Monoplex		
BCAS4	Saliva	61.4±8.5	64±7.1	52.4±8	N/A	41.1±7.1	N/A	30.5±5.9	27±5.6		
	Blood	11.9±8.7	6.1±1.4	10.9±7.4	N/A	12.2±6.3	N/A	12.4±6.9	3.2±2.8		
	Vaginal Epithelia	38±13.2	36±16.7	35.7±12.6	N/A	24.4±7.2	N/A	19.9±5.8	13±7.5		
	Semen	7.2±3.5	3.9±1.6	5.7±3.1	N/A	8.2±3.2	N/A	9.6±3.8	2.3±0.9		
		CpG1		CpG2		CpG3		CpG4		CpG5	
		Multiplex	Monoplex	Multiplex	Monoplex	Multiplex	Monoplex	Multiplex	Monoplex	Multiplex	Monoplex
cg06379435	Saliva	6.1±1.4	8.7±7	5.5±1.5	2.6±1.4	5.1±1.5	6±3.9	6±1.4	3.5±2.6	9.4±3.3	7.7±4.7
	Blood	25±7.2	24±7.8	24.9±7.5	22±6.7	27.2±7	33±7.4	28.4±6.5	30±8.2	44.7±9.8	49±12
	Vaginal Epithelia	6.3±2.2	2.3±0.6	5.7±1.9	1.3±0.6	5.4±1.5	3±1	6.2±1.8	2±1	12.3±5.2	8.3±7.8
	Semen	7.1±2.6	3.4±1.7	5.8±2.7	2.4±1.8	6±3	2.7±1.1	6.5±3.9	1.8±1.3	9.5±7.3	3.13±1.2
		CpG1		CpG2		CpG3		CpG4			
		Multiplex	Monoplex	Multiplex	Monoplex	Multiplex	Monoplex	Multiplex	Monoplex		
VE_8	Saliva	59±7.5	62±10.1	75.7±4	65.3±9	70.9±4.3	57.7±3.8	86.5±4.3	74.7±5.9		
	Blood	66.1±8.9	59.3±15.3	76.6±4.4	66.7±4.5	76.8±4.6	69.7±3.8	89.4±4.4	90.3±4.2		
	Vaginal Epithelia	24.4±15.1	13±10.8	28.2±12.8	14.3±8.4	28.2±15.2	12±9.5	32.3±16.6	16±12.2		
	Semen	74.2±8.4	85.3±1.2	79.3±3.7	67.7±0.6	78.9±4.1	68.3±2.9	91.9±2.4	79±1.7		
		CpG1		CpG2		CpG3		CpG4		CpG5	
		Multiplex	Monoplex	Multiplex	Monoplex	Multiplex	Monoplex	Multiplex	Monoplex	Multiplex	Monoplex
ZC3H12D	Saliva	76.5±4.4	81±4.1	65.3±4.8	78±4.5	87.9±3.8	99±2.1	82.5±5.1	79±4.9	76.2±4.4	82±3.4
	Blood	87.6±5.5	94±1.6	77.9±7.1	94±2.7	87.5±6	100±0	95±6.1	97±7.9	79.4±6.9	86±3.5
	Vaginal Epithelia	74.1±8	77±9.7	65.9±9.3	77±10	86.6±3.6	97±2.5	80±9	73±10	73.9±5.7	80±4
	Semen	10.2±10.6	5.4±4	8.4±8.9	5.3±3.9	11.2±10.3	6.7±4.3	12.9±11.2	6.3±4	9.8±9.8	5.1±3.7

The *t*-test revealed that the CpGs from each marker in the multiplex returned a statistically significant difference in observed means when compared to the monoplex studies for the body fluid that that marker was intended to identify. This was initially a concern for the reliability of the proposed method for body fluid identification. However, given the previously discussed changes in experimental conditions, there is a plausible explanation for the difference in the difference in the means. In the validation study conducted by Madi et al., the identification of a body fluid was made using cut-off thresholds for each CpG in an assay. The specific threshold is irrelevant as long as the methylation data returned for samples are reproducible. If the observed variance between the two treatments is found to be statistically insignificant, then the multiplex assay could still be used to differentiate between body fluids, given that the observed means in the multiplex, while statistically different from the monoplex, still do not overlap amongst body fluid types.

To compare the variances, an F-test was used with an alpha of 0.05, and the results are seen in Tables 5.2-5.5. The results indicate that there are some statistically different variances in the data resulting from the analysis of the markers in multiplex versus monoplex, as indicated by calculated F test values above the critical F value. However, it should be observed in Tables 5.2-5.4 that the F-tests for BCAS4, cg05379435, and VE_8 show insignificant differences in variance for the body fluid that each marker is intended to identify. The F-test revealed that there is a statistically significant difference in the methylation variance observed in the ZC3H12D marker for semen samples. However, given the nearly 70% difference in mean methylation between semen and the other body

fluids in this marker, the larger variation doesn't negatively impact that ability to differentiate body fluids. This means that the results across the population study showed a level of variation that is more likely attributed to the natural variation in methylation across individuals. With large differences in mean methylation values for the four body fluids tested in the multiplex and variation across body fluids that is consistent with monoplex reactions, the body fluid multiplex's results are still able to be used to differentiate between body fluids.

Overall, the multiplex offers a greater amount of information in a single test than any of the monoplex reactions by themselves. Operationally, the multiplex permits the evaluation of all methylation values simultaneously. This permits the simultaneous application of methylation values across multiple loci when determining the presence of different body fluids, increasing specificity. Various prediction models and expert systems can then be applied to assist the analyst in determining the origin of the sample.

Table 5.2 – Results of F-test for the BCAS4 marker comparing the variance observed in the results of the multiplex and monoplex reactions. Although blood and semen show statistically significant differences in variance across the CpGs, the variance observed in the saliva samples is not statistically significant, indicating that the assay is still able to produce reliable results for the body fluid it is intended to identify. *= the body fluid the assay is specific for.

F-Test for BCAS4

	Saliva*		Blood		Vaginal Epithelia		Semen	
CpG1	Multiplex	Monoplex	Multiplex	Monoplex	Multiplex	Monoplex	Multiplex	Monoplex
Variance	72.25	50.41	75.69	1.96	174.24	278.89	12.25	2.56
F test	1.43		38.62		1.60		4.79	
α	0.05		0.05		0.05		0.05	
F critical	3.05		3.07		3.1		3.09	
CpG2								
Variance	64	N/A	54.76	N/A	158.76	N/A	9.61	N/A
F test	-		-		-		-	
α	0.05		0.05		0.05		0.05	
F critical	3.05		3.07		3.1		3.09	
CpG3								
Variance	50.41	N/A	39.69	N/A	51.84	N/A	10.24	N/A
F test	-		-		-		-	
α	0.05		0.05		0.05		0.05	
F critical	3.05		3.07		3.1		3.09	
CpG4								
Variance	34.81	31.36	47.61	7.84	33.64	56.25	14.44	0.81
F test	1.11		6.07		1.67		17.83	
α	0.05		0.05		0.05		0.05	
F critical	3.05		3.07		3.1		3.09	

Table 5.3 – Results of F-test for the cg06379435 marker comparing the variance observed in the results of the multiplex and monoplex reactions. Statistically significant differences in variance were observed in several CpGs across the saliva, vaginal epithelia, and semen samples, but the variance observed in the blood samples is not statistically significant, indicating that the results for the assay are reproducible for the body fluid it is intended to identify. *= the body fluid the assay is specific for.

F-test for cg06379435

	Saliva		Blood*		Vaginal Epithelia		Semen	
CpG1	Multiplex	Monoplex	Multiplex	Monoplex	Multiplex	Monoplex	Multiplex	Monoplex
Variance	1.96	49.00	51.84	60.84	4.84	0.00	6.76	2.89
F test	25.00		0.85		0.00		2.34	
α	0.05		0.05		0.05		0.05	
F critical	3.05		3.07		3.10		3.09	
CpG2								
Variance	2.25	1.96	56.25	44.89	3.61	0.36	7.29	3.24
F test	1.15		1.25		10.03		2.25	
α	0.05		0.05		0.05		0.05	
F critical	3.05		3.07		3.10		3.09	
CpG3								
Variance	2.25	15.21	49.00	54.76	2.25	1.00	9.00	1.21
F test	6.76		1.12		2.25		7.44	
α	0.05		0.05		0.05		0.05	
F critical	3.05		3.07		3.10		3.09	
CpG4								
Variance	1.96	6.76	42.25	67.24	3.24	1.00	15.21	1.69
F test	0.29		0.63		0.31		9.00	
α	0.05		0.05		0.05		0.05	
F critical	3.05		3.07		3.10		3.09	
CpG5								
Variance	10.89	22.09	96.04	144.00	27.04	60.84	53.29	1.44
F test	2.03		1.50		2.25		37.01	
α	0.05		0.05		0.05		0.05	
F critical	3.05		3.07		3.10		3.09	

Table 5.4 – Results of F-test for the VE_8 marker comparing the variance observed in the results of the multiplex and monoplex reactions. Statistically significant differences in variance were observed in 2 CpGs in semen and 1 CpG in saliva, indicating that the variance observed in multiplex and monoplex reactions are quite similar. Of particular note, even though the standard deviation for vaginal epithelial samples analyzed with the VE_8 marker in the multiplex is quite large, it is still in line with the monoplex analysis. *= the body fluid the assay is specific for.

F-Test for VE_8

	Saliva		Blood		Vaginal Epithelia*		Semen	
CpG1	Multiplex	Monoplex	Multiplex	Monoplex	Multiplex	Monoplex	Multiplex	Monoplex
Variance	56.25	102.01	79.21	234.09	228.01	116.64	70.56	1.44
F test	1.81		2.96		1.95		49.00	
α	0.05		0.05		0.05		0.05	
F critical	3.05		3.07		3.10		3.09	
CpG2								
Variance	16.00	81.00	19.36	20.25	163.84	70.56	13.69	0.36
F test	5.06		1.05		2.32		38.03	
α	0.05		0.05		0.05		0.05	
F critical	3.05		3.07		3.10		3.09	
CpG3								
Variance	18.49	14.44	21.16	14.44	231.04	90.25	16.81	8.41
F test	1.28		1.47		2.56		2.00	
α	0.05		0.05		0.05		0.05	
F critical	3.05		3.07		3.10		3.09	
CpG4								
Variance	18.49	34.81	19.36	17.64	275.56	148.84	5.76	2.89
F test	1.88		1.10		1.85		1.99	
α	0.05		0.05		0.05		0.05	
F critical	3.05		3.07		3.10		3.09	

Table 5.5 – Results of F-test for the ZC3H12D marker comparing the variance observed in the results of the multiplex and monoplex reactions. For this marker, there were significant differences in the methylation variance for semen when comparing the multiplex and monoplex results. This would suggest that the multiplex is giving results that are inconsistent with the monoplex assay. However, the multiplex results retain the large difference in mean methylation between semen and the other tested body fluids. *= the body fluid the assay is specific for.

F-test for ZC3H12D

	Saliva		Blood		Vaginal Epithelia		Semen*	
CpG1	Multiplex	Monoplex	Multiplex	Monoplex	Multiplex	Monoplex	Multiplex	Monoplex
Variance	19.36	16.81	30.25	2.56	64.00	94.09	112.36	16.00
F test	1.15		11.82		1.47		7.02	
α	0.05		0.05		0.05		0.05	
F critical	3.05		3.07		3.10		3.09	
CpG2								
Variance	23.04	20.25	50.41	7.29	86.49	100.00	79.21	15.21
F test	1.14		6.91		1.16		5.21	
α	0.05		0.05		0.05		0.05	
F critical	3.05		3.07		3.10		3.09	
CpG3								
Variance	14.44	4.41	6.00	0.00	12.96	6.25	106.09	18.49
F test	3.27		0.00		2.07		5.74	
α	0.05		0.05		0.05		0.05	
F critical	3.05		3.07		3.10		3.09	
CpG4								
Variance	26.01	24.01	37.21	62.41	81.00	100.00	125.44	16.00
F test	1.08		1.68		1.23		7.84	
α	0.05		0.05		0.05		0.05	
F critical	3.05		3.07		3.10		3.09	
CpG5								
Variance	19.36	11.56	47.61	12.25	32.49	16.00	96.04	13.69
F test	1.67		3.89		2.03		7.02	
α	0.05		0.05		0.05		0.05	
F critical	3.05		3.07		3.10		3.09	

Sensitivity Study

The purpose of the sensitivity study was to determine the lower limit of input DNA that can be used while still producing reliable results. In previous studies, while low level (0.1-1 ng) PCR products were detected on a gel, the methylation data was found to be inconsistent across replicates, causing some low level samples to produce methylation beyond the thresholds that were used as cutoffs.[173] There is no doubt that the PCR amplification successfully amplified the DNA in the reaction. However, the resulting DNA

methylation content can be influenced by stochastic effects due to the relatively low number of cells present, as little as 15 cells in a 100pg reaction, and DNA degradation effects produced through bisulfite modification. Further studies were needed to define this effect which is sometimes referred to as PCR bias.[184]

In this study, a sample of each body fluid was analyzed in five replicates. The mean percent methylation for each of the 18 CpG sites across the multiplex did not show large deviations from the expected values, however the standard deviations for the lower input levels increased by anywhere from 3- to 10-fold, depending on the marker. To evaluate the results, the standard deviation of the 20ng input levels was compared to each of the subsequent input levels. The F-tests revealed that below 1ng, the body fluid and marker combinations that present intermediate methylation values (25-75%) throughout the multiplex had a statistically significant increase in variance. The primary implication is that if the methylation value observed at a CpG for a particular sample deviates significantly from that body fluid's known profile due to low level of input DNA, then the assay's ability to correctly identify body fluid samples is lost. In Figure 5.2, the results of the saliva replicates at different input levels are shown. For the saliva replicates, input levels of 500pg produced methylation profiles in one of the five replicates that was greater than two standard deviations of the 20ng input level across each CpG in the BCAS4, VE_8, and ZC3H12D markers. At 250pg and 100pg, two of the five replicates produced methylation profiles with variances greater than two standard deviations of the 20ng input level across each CpG in each of those same markers. The methylation observed at the cg06379435 marker was stable from 20ng to 100pg, likely due to the hypomethylated nature of saliva samples at this marker. Similar results were seen across the replicates of

the other body fluids tested, i.e. any body fluid and marker combination that should produce a hyper or hypomethylated methylation profile was stable down to 100pg, while intermediate methylation levels caused samples with less than 1ng input to be inconsistent with high input DNA samples. This trend is consistent with the sensitivity study conducted by Madi et al. as the effects of input DNA is irrespective to the body fluid origin of the DNA.

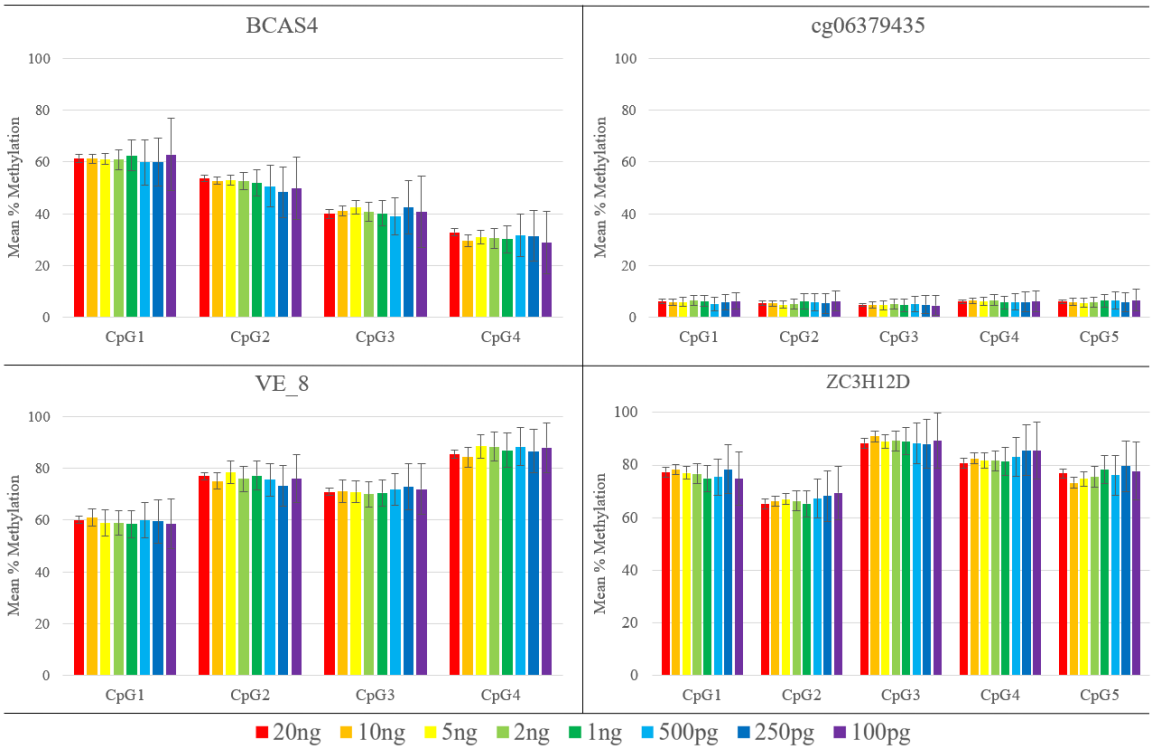

Figure 5.2 – Methylation profile of five replicates of a saliva sample analyzed at various input levels in the body fluid identification multiplex. Similar trends were observed for blood, vaginal epithelia, and semen.

There are several proposed reasons for this to occur. First, there is the question of sampling. A cell contains approximately 6.6 pg of gDNA and if a PCR reaction includes only 100pg of DNA, then the reaction can be presumed to have roughly 15 copies of the target region for DNA methylation analysis. It is also important to remember that a CpG site is not partially methylated; the CpG is either methylated or unmethylated. The percent

methylation at a CpG site that is reported after pyrosequencing is the ratio of methylated and unmethylated DNA in the sample. So, when using very low input levels of DNA for pyrosequencing, there is a larger probability that differences in percent methylation will be a function of sampling, rather than an actual difference in methylation.[185] The second driving factor for differences in expected DNA methylation and low input samples would be PCR bias that occurs during amplification.[186] It is possible that with low levels of input DNA, there can be a more efficient amplification of either the unmethylated or the methylated bisulfite converted DNA, depending on the construction of the PCR primers.[184] To combat these effects, there are two likely remedies. Future body fluid identification marker development should focus on markers that are either hypermethylated or hypomethylated for the target body fluid compared to other body fluids, like in the cg06379435 and ZC3H12D markers. Secondly, combining replicates of low input samples as a routine analytical procedure and using the mean methylation of the replicates to identify body fluids would increase accuracy, and is similar to methods currently utilized in forensic laboratories for low-copy number samples.[187]

Mixture Study

In short tandem repeat analysis, the use of probabilistic genotyping software has made significant strides in determining the contributors of a DNA sample containing multiple contributors. However, there still remains the question of the origin of the DNA sample and which body fluid provided the STR result. In this study, mixtures of two different types of body fluids were prepared at various ratios. The resultant pyrogram provides an intermediate ratio of the methylation state of the two different body fluids,

while generally not influencing methylation data for body fluids not present. (Figure 5.3). The resultant data demonstrates a proportionate response to each mixture for the majority of queried CpGs. As can be seen in the mixtures of saliva and blood, the observed methylation at each CpG site across the entire multiplex is an intermediate value that is proportional to the ratios of the two pure body fluid profiles. CpG2 of VE_8 and CpG3 of ZC3H12D do not change almost at all across the mixture ratios as the pure samples had nearly identical methylation. These trends remain true for the 5 other mixtures that were evaluated (Saliva/Vaginal Epithelia in Figure 5.4, Saliva/Semen in Figure 5.5, Blood/Vaginal Epithelia in Figure 5.6, Blood/Semen in Figure 5.7, and Vaginal Epithelia/Semen in Figure 5.8).

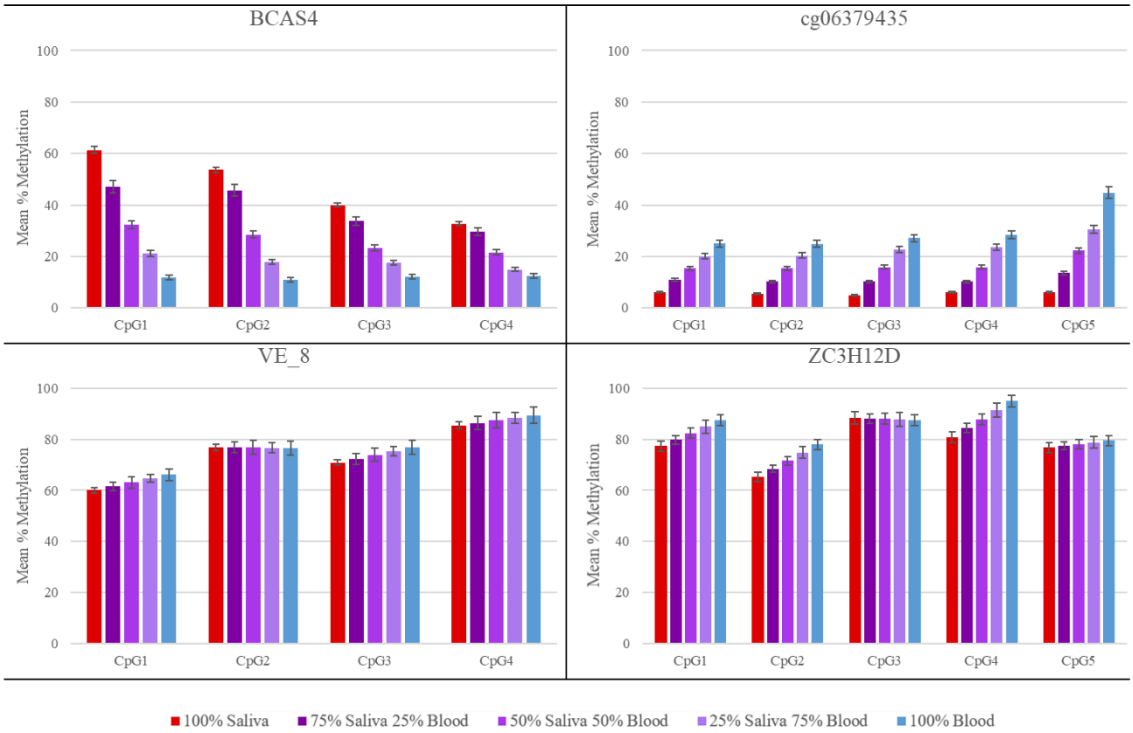

Figure 5.3 – Methylation results for mixture of saliva and blood at different ratios run on the multiplex. The major impact for the mixture occurs for the BCAS4 (saliva marker) and cg0637935 (blood marker). Minor variations are also seen for the other two markers, VE_8 (vaginal epithelial) marker and ZC3H12D (semen marker).

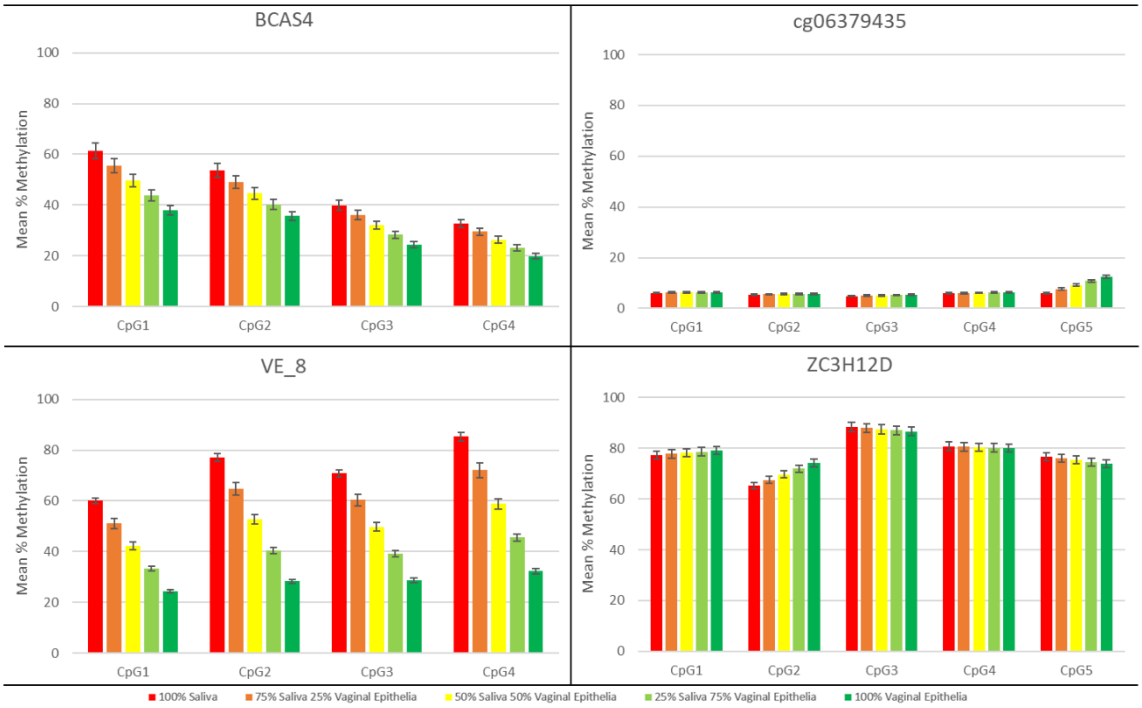

Figure 5.4 – Methylation results for mixture of saliva and vaginal epithelia at different ratios run on the multiplex. The major impact for the mixture occurs for the BCAS4 (saliva marker) and VE_8 (vaginal epithelia marker). Minor variations are also seen for the other two markers, cg06379435 (blood marker) and ZC3H12D (semen marker).

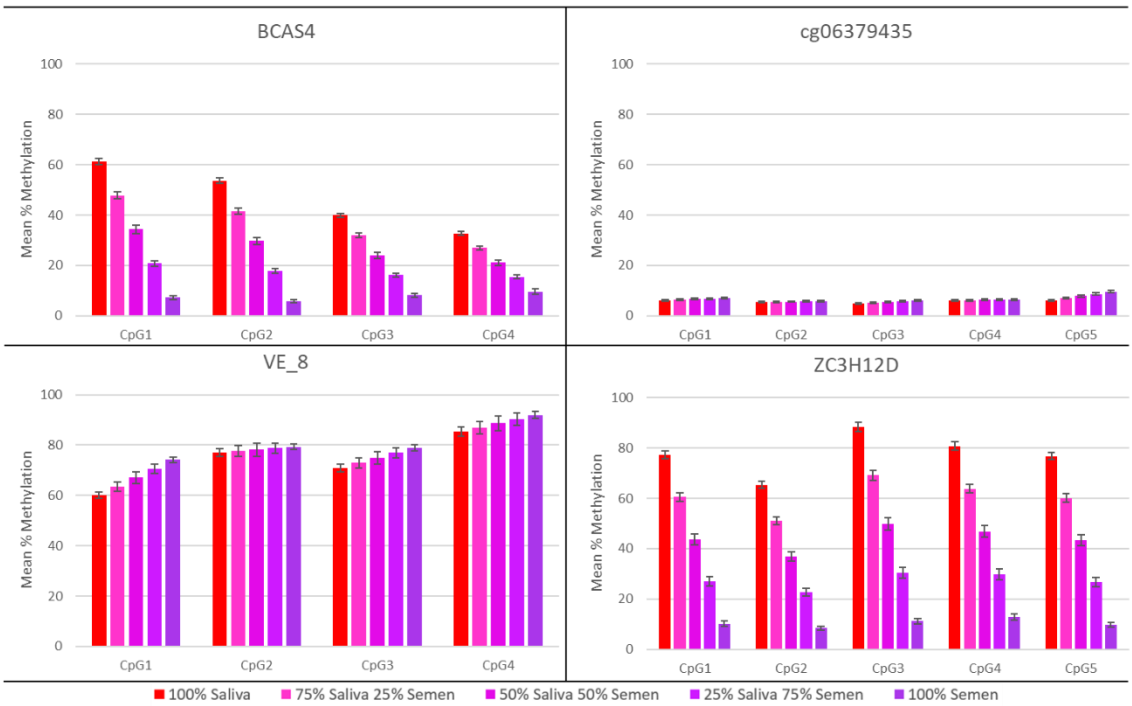

Figure 5.5 – Methylation results for mixture of saliva and semen at different ratios run on the multiplex. The major impact for the mixture occurs for the BCAS4 (saliva marker) and ZC3H12D (semen marker). Minor variations are also seen for the other two markers, cg06379435 (blood marker) and VE_8 (vaginal epithelia marker).

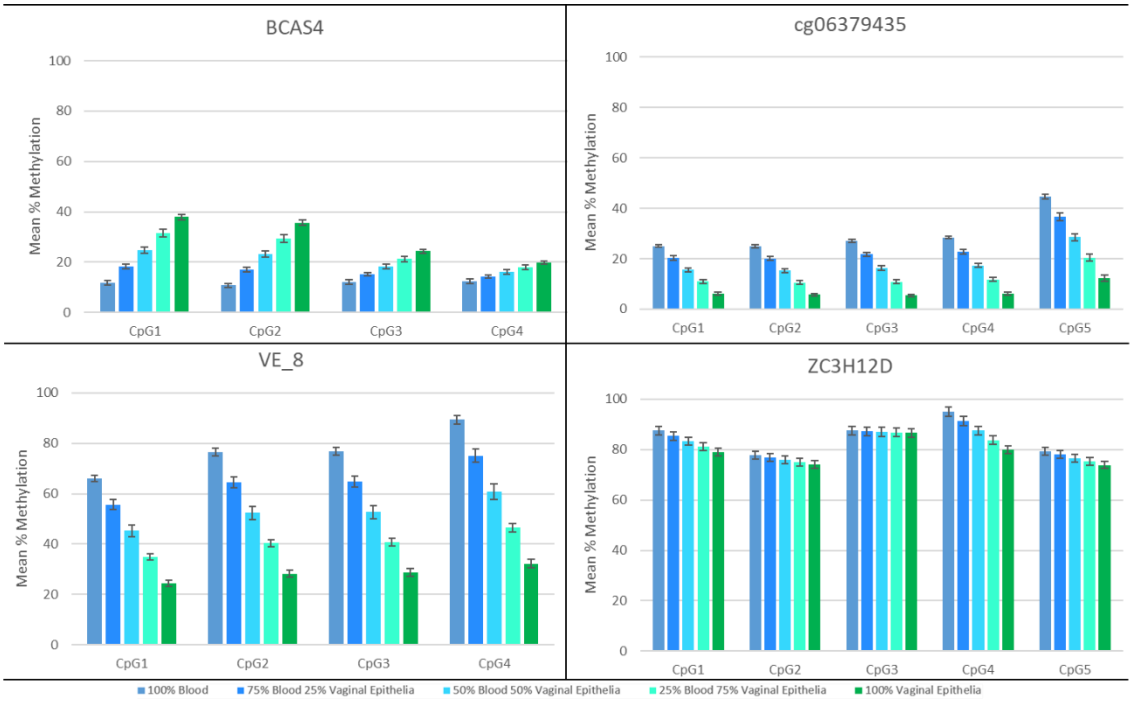

Figure 5.6 – Methylation results for mixture of blood and vaginal epithelia at different ratios run on the multiplex. The major impact for the mixture occurs for the cg06379435 (blood marker) and VE_8 (vaginal epithelia marker). CpG1 and CpG2 of the BCAS4 (saliva marker) also are impacted, due to the difference in methylation between vaginal epithelia and blood on the saliva marker. Minor variations are also seen for the ZC3H12D (semen marker).

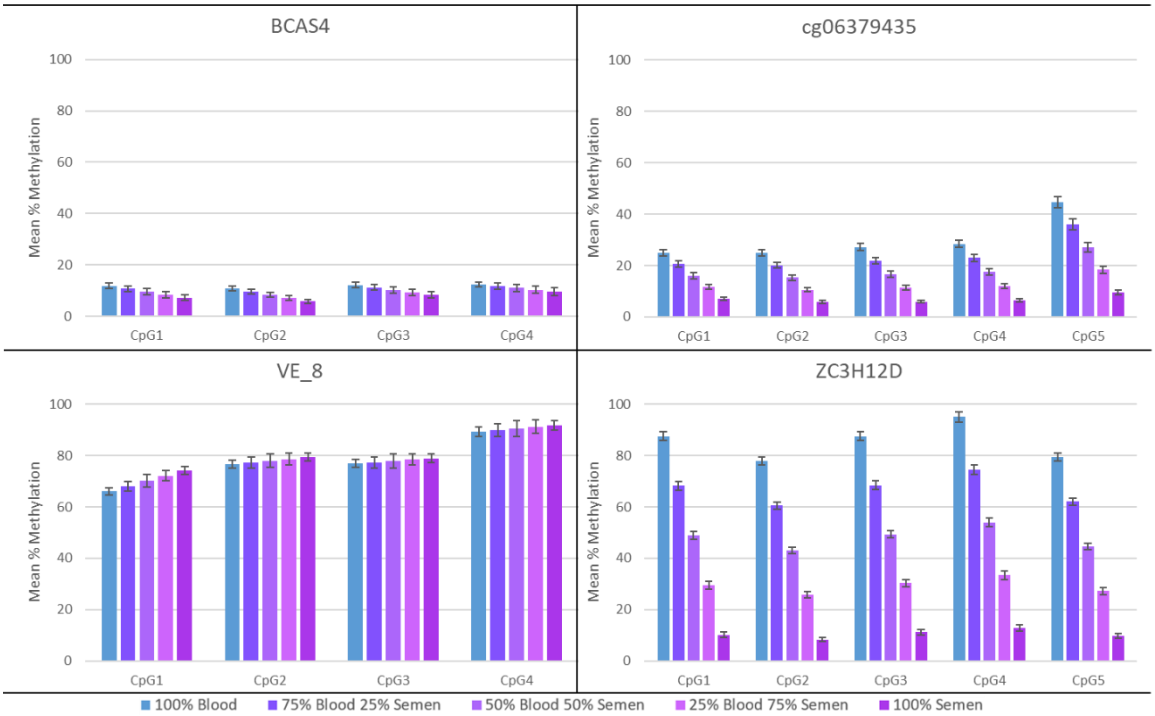

Figure 5.7 – Methylation results for mixture of blood and semen at different ratios run on the multiplex. The major impact for the mixture occurs for the cg06379435 (blood marker) and ZC3H12D (semen marker). Minor variations are also seen for the other two markers, BCAS4 (saliva marker) and VE_8 (vaginal epithelia marker).

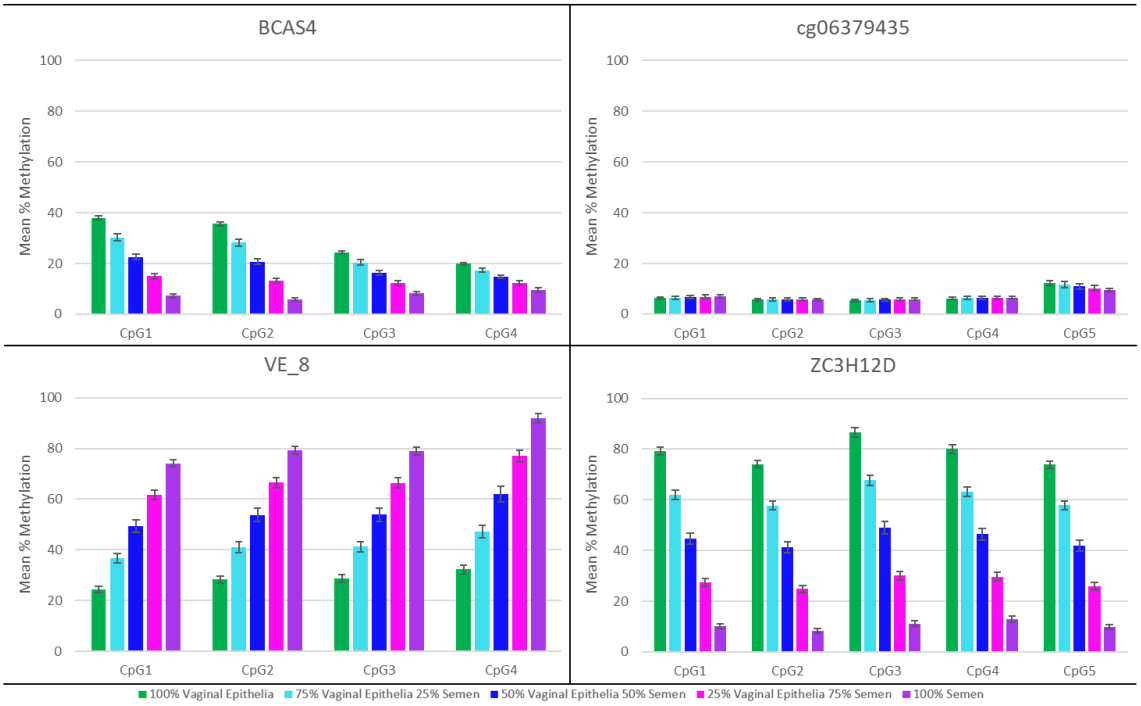

Figure 5.8 – Methylation results for mixture of vaginal epithelia and semen at different ratios run on the multiplex. The major impact for the mixture occurs for the VE_8 (vaginal epithelia marker) and ZC3H12D (semen marker). CpG1 and CpG2 of the BCAS4 (saliva marker) also are impacted, due to the difference in methylation between vaginal epithelia and semen on the saliva marker. Minor variations are also seen for the cg06379435 (blood marker).

As seen in the figures above, the mixture of two body fluids analyzed as a single sample clearly alters the resulting methylation profile of that sample. In each scenario, the resultant values show a linear response to the presence of each mixture. This should allow for the interpretation of mixture data by a trained analyst possible. Each one of the mixtures above is quite obviously not a single source sample, as the methylation profile of the mixture does not fit with profiles, and means, of single source samples developed in the population study.

As observed in the sensitivity study and population study, if sufficient quantities of DNA are present for the amplification of the multiplex, the resulting methylation profile of each body fluid has variance across the multiplex of approximately 2% (cg06379435 for

saliva, blood, and semen) to 15% (vaginal epithelia in VE_8), Given that the difference in methylation across each body fluid in each marker ranges from 20-75%, it should be possible to confirm the presence of a mixture, and to approximate the ratio. Single source data for the presumed two body fluids present in a mixture combined with the observed methylation value at each CpG in the multiplex could be used to give an approximate ratio of the two body fluids. This information could prove useful to analysts in scenarios when the presence of semen in a sexual assault sample is in dispute. In this study, the use of a multiplex for body fluid identification is able to presumptively identify mixtures of two body fluids present and help in excluding the body fluids that are not present in the sample.

Inhibition and Degradation Studies

It is not uncommon for forensic casework samples arriving in the laboratory to contain inhibitors or be significantly degraded due to time or exposure to the elements. To recreate inhibition, DNA samples were spiked with hematin at final concentration of 0.08M, and humic acid at final concentration of 0.24 mg/mL. The samples were spiked either before or after bisulfite conversion. The results of the inhibition study show that if the inhibitor is added before bisulfite conversion, there is a less than 10% decrease in peaks heights observed at each peak in the resulting pyrograms, while if the inhibitor was added following bisulfite conversion the resultant amplification was very poor with over 90% loss in peak intensity observed in the pyrogram, (Figure 5.9). It should be noted that the process of bisulfite conversion includes a sample purification utilizing a silica membrane filter in a spin column that adsorbs DNA while other molecules are not retained. While the

DNA is adsorbed to the membrane, ethanol wash steps likely remove inhibitors, thus purifying the DNA.

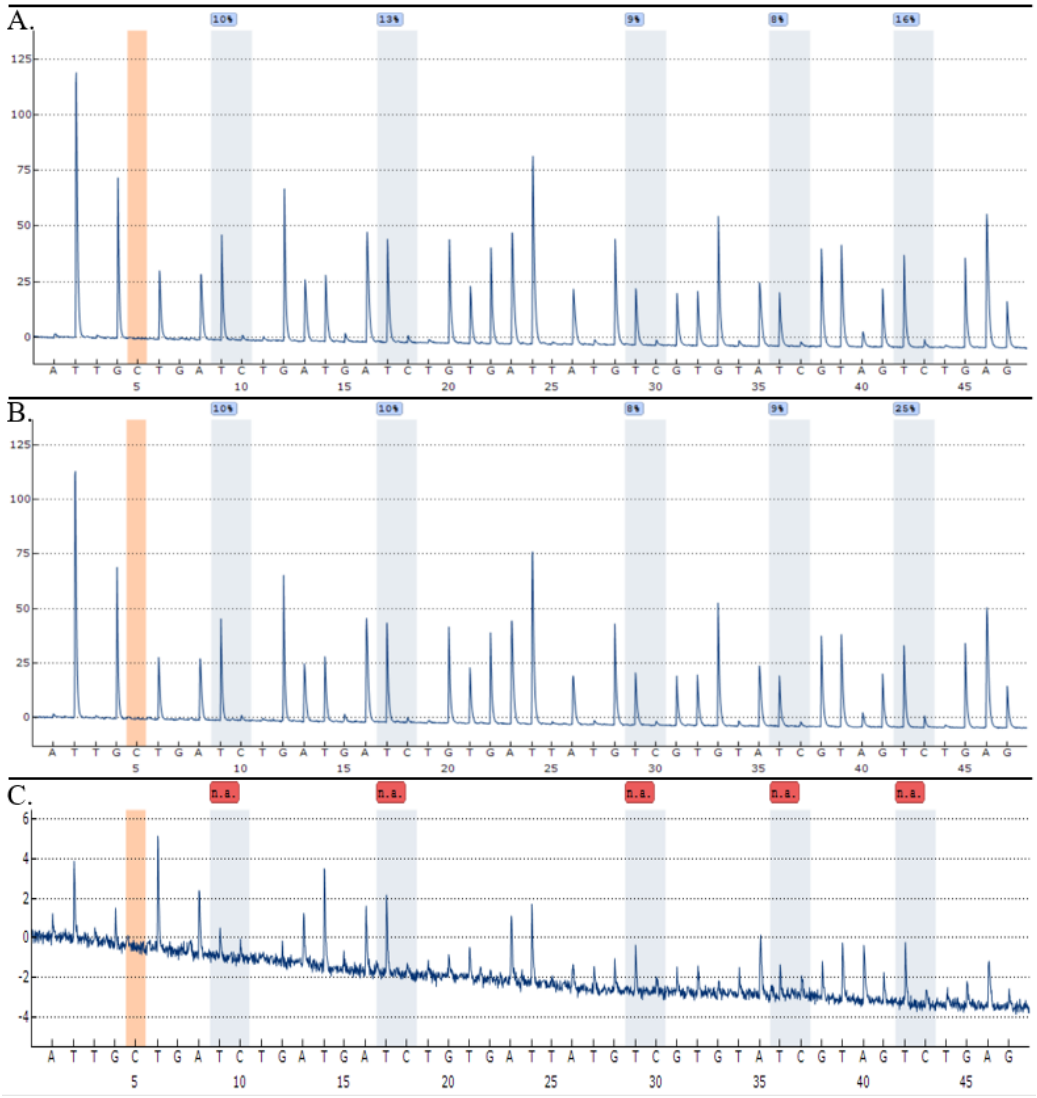

Figure 5.9 – Inhibition study showing the cg06379435 pyrogram of an unmethylated control DNA sample with A) no humic acid added B) humic acid added before bisulfite conversion and C) humic acid added after bisulfite conversion.

Four 10µL aliquots of 10ng/µL EpiTect Methylated Converted Control DNA were incubated in 0.2mL PCR tubes on a thermal block for 0, 14, 20 or 25 minutes. The samples were then amplified in triplicate with the body fluid identification multiplex and sequenced on the pyrosequencer. In the previous study by Madi et al., the monoplex reactions of these

body fluid markers produced functioning pyrograms with no discernible difference between the samples that were not heated, and the samples that were heated for 25 minutes, suggesting the degradation had little to no effect on the amplification of the targets. In this degradation study, effects on the peak height of the resulting pyrograms was observed at each time point in all samples, Figure 5.10. The level of degradation at the 14- and 20-minute marks caused an approximately 25-35% decrease in observed peak heights across the pyrogram. At the 25-minute mark, the degradation caused a 70% reduction in observed peak heights. The loss in peak height caused the PyroMark Q48 Autoprep software to flag most of the variable positions for review, decreasing confidence in the results of the assay.

A reason for the reduced peak heights seen in this degradation study can be attributed to the different experimental conditions of the multiplex assay versus the monoplex assays. When sequencing the multiplex PCR products, all four PCR products are present and competing for magnetic bead binding. It is possible that degradation of the DNA was enough to affect amplification, but when sequencing the monoplex PCR products, the magnetic beads captured the full PCR product, resulting in pyrograms with relatively unchanged peak heights. In the multiplex assay, the decreased level of amplification is exacerbated by the limiting reagent of the sequencing reaction, magnetic beads, and the reduced peak heights of the pyrogram is observed.

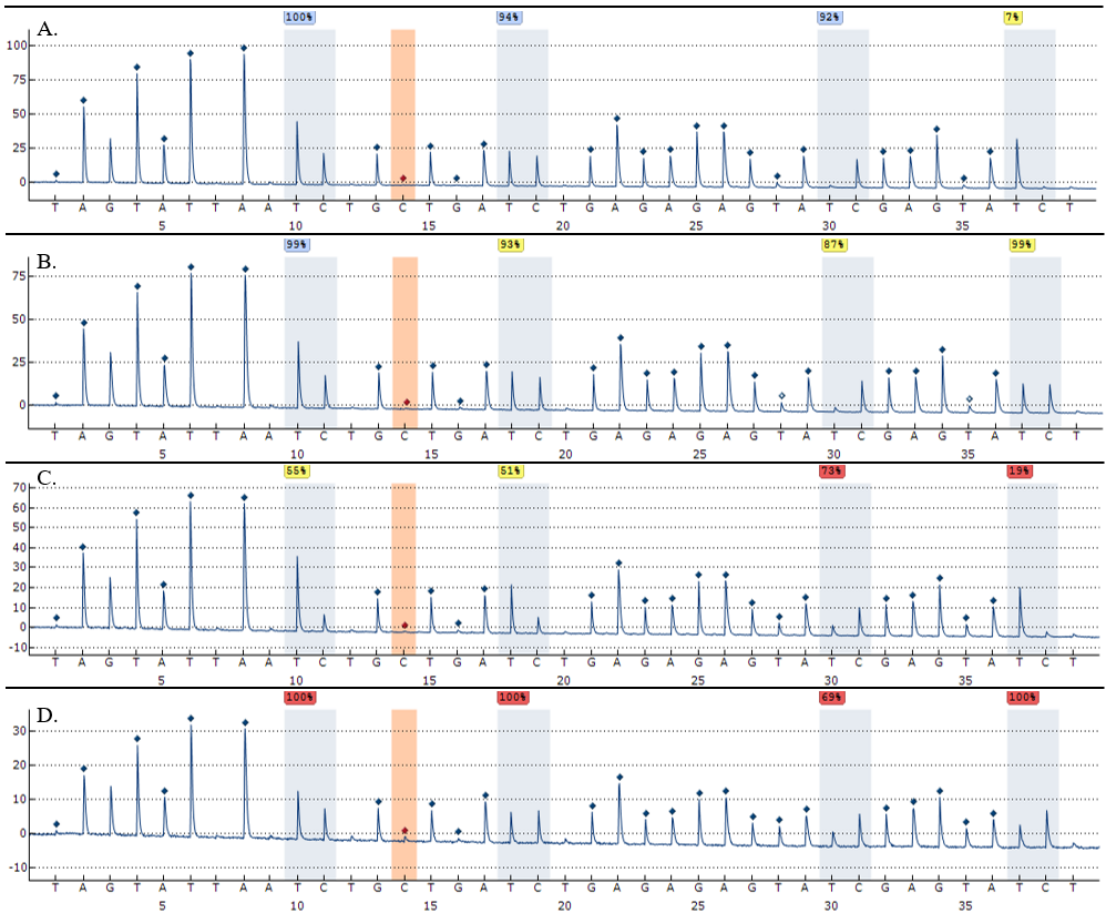

Figure 5.10 – Degradation study showing the BCAS4 pyrogram of a methylated control DNA sample after incubation at 95 ºC for A) 0 minutes B) 14 minutes C) 20 minutes and D) 25 minutes. The reduced peak heights at longer incubation times is likely due to the fragmented DNA not amplifying to the same extent as unfragmented DNA and therefore not enough PCR product is available for sequencing.

D. Concluding Remarks

When trace levels of DNA are present, it becomes difficult to use standard serological tests to determine body fluid type. The ability to determine trace levels of body fluids in a concise and reproducible manner at a crime scene can provide important evidence to the trier of facts. The results of these validations studies demonstrate that the newly created body fluid identification multiplex is both reliable and robust, and fit to purpose. The results of the validation study show that, as expected, sub nanogram levels of DNA produced increased stochastic variation, mixtures can be identified and

approximated, that bisulfite modification removes inhibitors and that the analysis can be used on moderately degraded DNA.

CHAPTER VI – AUTOMATED BODY FLUID IDENTIFICATION USING THE BODY FLUID MULTIPLEX

It is possible to differentiate between saliva, blood, vaginal epithelia, and semen using 4 different tissue-specific DNA methylation loci, analyzing the data one locus at a time. However as shown in the previous chapter, precious sample is conserved by combining a single evidentiary extract into a four-locus multiplex. An additional question can be asked: Instead of using each locus as a simple yes/no, might additional specificity be obtained by simultaneously using all the methylation data obtained by the four assays, to make a conclusion? The goal of this portion of the project was to examine the use of data analysis tools combine all methylation results and with a goal to improve the specificity of determining the body fluid origin of a sample. The use of an expert system should result in a faster interpretation of the data, and also makes the result less subject to unconscious bias.

In this chapter, two methods, cluster analysis and latent profile analysis, were applied to a data set consisting of various body fluid type samples. Their relative ability to differentiate body fluids was compared and the result examined for potential use in forensic analysis.

A. Cluster Analysis Primer

Cluster analysis is a multivariate method that classifies a given sample set based on a set of experimental responses. The classification process sorts similar samples into the same group.[188] There are a number of ways to achieve this result, but the general process is to examine experimental results and develop classifying variables from this data. These

variables are used to place individual samples into sets of clusters with similar behavior. Placing samples from the experimental data set into clusters based on similarity between samples will increase the size of the clusters while decreasing the total number of clusters. In this project body fluid identification from the multiplex was performed using an agglomerative hierarchical cluster analysis through squared Euclidean distances and Ward's linkage method via SPSS software (Statistical Product and Service Solutions) version 20 (IBM, NY).

In agglomerative hierarchical cluster analysis, the model begins with each sample in a set of experimental results being its own cluster. The two most similar clusters are identified, collapsed into one, and then the process repeats until there is a single cluster containing all of the samples. An agglomeration schedule coefficient is determined which gives a numerical analysis at each stage of the cluster solution of successive collapsing of clusters. As the cluster analysis continues, this coefficient becomes larger as the combined clusters become more different from each other. By the end of the agglomerative hierarchical cluster analysis, the increase in the agglomeration coefficient becomes extremely large, suggesting that clusters that do not have much in common have been combined. The relative size of this coefficient can be used to determine how many clusters naturally exist in the sample set.[189]

To determine the distance between clusters, Euclidean distance is commonly used for interval data sets, such as methylation at CpG sites that can be anywhere from 0-100%.[190] For this measurement, if p variables X_1, X_2,..., X_p for *n* samples exist, then the data for sample *i* can be written as x_{i1}, x_{i2}, ..., x_{ip} and sample *j* as x_{j1}, x_{j2}, ..., x_{jp}. The Euclidean distance is calculated using the following formula:

$$d_{ij} = \sqrt{(x_{i1} - x_{j1})^2 + (x_{i2} - x_{j2})^2 + \ldots + (x_{ip} - x_{jp})^2}$$

In this study, the squared Euclidean distance was used. By squaring the value obtained with the formula above, the differences between clusters can be increased, emphasizing the importance of larger distances, while deemphasizing the importance of smaller distances. The agglomeration schedule coefficient is the within-cluster sum of squares, which is calculated using the following formula:

$$\sum_{k=1}^{K} \sum_{i \in S_k} \sum_{j=1}^{p} (x_{ij} - \bar{x}_{kj})^2$$

Where k is the cluster number, p is the number of variables, S_k is the set of observations in the k^{th} cluster and $\bar{\mathbf{x}}_{kj}$ is the j^{th} variable of the cluster center for the kth cluster.[191]

Finally, the method to decide which clusters should be combined must be used. While several methods exist, this study utilized Ward's linkage method. In Ward's method all possible pairs of clusters are combined, and the within-cluster sum of the square is calculated for each combination. The sum of squared distance within each possible cluster pair is then compared to the squared distance over all of the clusters. Whichever combination of possible cluster pairs results in the smallest sum of squares is then confirmed as the two clusters to be collapsed at that stage, and the process repeats for each subsequent stage.[192]

Prior to using cluster analysis on the population data set from the developmental validation, the data had to be checked to ensure that it met the assumptions of the cluster method.[193] The first assumption is that the samples are representative of the population and the second is that the input variables are not dependent on each other. For the first

assumption, the samples from the population study were collected from random individuals without regard for any underlying information. For the second assumption, a bivariate correlation analysis of the CpGs revealed only a weak statistical correlation with each other, likely due to the persistent methylation profiles that are observed in various body fluids. There is nothing in the literature to suggest that the methylation status of a CpG in one genomic location can affect the methylation status of a CpG elsewhere in the genome, and therefore the input variables for the cluster analysis are presumed to not be dependent on each other.

Over the course of the analysis, the population data set was split in to two groups, the training set and the test set. The training set consists of 74 samples (20 saliva, 20 blood, 17 vaginal epithelia, and 18 semen) and was used to identify the number of clusters that naturally occur within the data. This process was repeated ten times with the samples in a random order each time, as it is possible for the order of samples to affect the agglomeration schedule.[194] The hope was to create a model that has a cluster number that corresponds to the number of body fluid groups in the training set. The test set consists of 46 samples (17 saliva, 10 blood, 9 vaginal epithelia, and 10 semen) and was used to verify the ability of the cluster model to categorize the unknown samples into the correct body fluid cluster. Using the results of the cluster analysis, an ANOVA with Tukey *post-hoc* analysis was used to determine which of the CpG sites within the multiplex were most discriminatory and if the model accuracy could be improved by focusing on those more discriminatory CpGs. Then a new model was developed consisting of only the most discriminatory CpGs. This model was analyzed, again repeating the analysis 10 times to validate the model. The

refined model was used to again categorize the test set into the correct body fluid clusters, and the accuracy of the model was evaluated.

B. Cluster Analysis Results

The results of the first cluster analysis using the full set CpGs in the multiplex are shown in Table 6.1. Within the column for the agglomeration schedule coefficient, it can be seen that the agglomeration of the last three stages caused the largest increases to the coefficient. This would suggest that stage 71 produced an optimum size with four naturally occurring clusters within the data set This same result was achieved following ten replicate analyses. The replicate analyses continued to utilize the same training set, but the order of samples was randomized each time.

Table 6.1 – Agglomeration schedule for 74 samples using data from the 18 CpGs in the multiplex

Agglomeration Schedule						
Stage	Cluster Combined		Coefficients	Stage Cluster First Appears		Next Stage
	Cluster 1	Cluster 2		Cluster 1	Cluster 2	
1	57	61	16.00	0	0	6
2	59	60	36.50	0	0	9
3	58	72	58.00	0	0	26
4	62	64	80.00	0	0	18
⋮	⋮	⋮	⋮	⋮	⋮	⋮
68	1	8	35586.80	65	60	71
69	57	66	40288.75	63	61	73
70	41	43	51169.79	62	64	71
71	1	41	150477.48	68	70	72
72	1	21	295503.25	71	67	73
73	1	57	711840.88	72	69	0

Next, the test set of 46 samples was run through the same cluster analysis with the same parameters, but with the stipulation that there must be four clusters at the end, and the cluster membership for each sample should be specified. The resulting dendrogram is shown in Figure 6.1. Within this model, there was one miscategorized sample; a single vaginal epithelial sample was categorized as a blood sample. This sample showed higher than normal methylation values at all four CpG sites in the VE_8 marker (64-74%) when compared to the expected values seen in the population data (24-32%). The percent methylation observed in the other three markers was unremarkable, which would suggest that high methylation in the VE_8 marker is not a result of cross-contamination. Given that vaginal epithelial samples are collected without control for proximity to menstrual cycles, it is possible that some menstrual blood existed in the sample, which could cause the sample to exhibit methylation different from a normal vaginal epithelial sample. A Rescaled Distance Cluster Combine was next examined in order to determine the level of difference between each cluster. It shows that the saliva and vaginal epithelial clusters are the most similar at a cluster combine distance of just six, likely due to the similarity in methylation values for saliva and vaginal epithelia in the BCAS4 marker. The blood cluster is not far from those two clusters with an additional cluster combine distance of 3. The semen cluster is the most distinct cluster in the model, with a cluster combine distance of 16 from any other cluster. These results demonstrate discrimination between the body fluids however, the relatively small distance between the clusters may cause the model to incorrectly place certain samples.

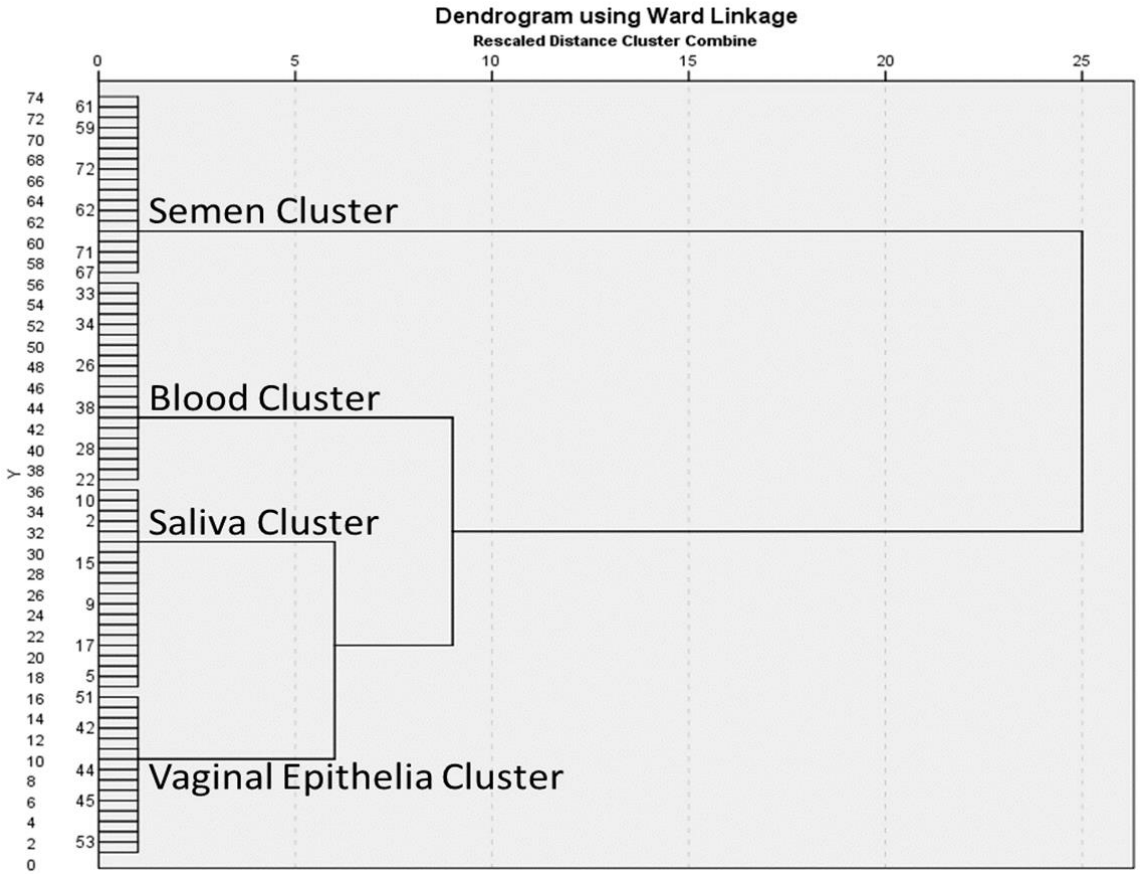

Figure 6.1 – Dendrogram resulting from the categorization of saliva (n=20), blood (n=20), vaginal epithelia (n=17) and semen (n=17) using 18 CpGs in the multiplex.

One crucial detail about cluster analysis is that it does not possess the ability to differentiate between relevant and irrelevant variables. Every data point fed into the algorithm is treated as equally valid for the determination of clusters. In an attempt to increase the accuracy of the identification model, an ANOVA and Tukey *post-hoc* analysis of the cluster analysis data was used to determine which of the CpG sites in the multiplex were most discriminatory for their respective body fluid. For the Tukey's *post-hoc*, given the multiple comparisons being made, a Bonferroni corrected p-value of 0.0083 was used. CpG sites that displayed a significance value below this p-value were determined to be

highly discriminatory. The results, show in abbreviated form in Table 6.2, indicate that five of the 18 CpG sites are the most discriminatory for the identification of body fluids.

Table 6.2 – Results of Tukey's post-hoc analysis for the five most discriminatory CpGs in the body fluid identification multiplex.

Tukey's Test						99.17% Confidence Interval	
CpG			Mean Difference (I-J)	Std. Error	Significance	Lower Bound	Upper Bound
BCAS4_CpG1	Saliva	Blood	41.471*	3.326	<0.001	30.24	52.71
		Vaginal Epithelia	26.846*	3.685	<0.001	14.40	39.29
		Semen	51.571*	3.426	<0.001	40.00	63.14
cg06379435_CpG1	Blood	Saliva	18.433*	1.022	<0.001	14.98	21.89
		Vaginal Epithelia	19.352*	1.228	<0.001	15.21	23.50
		Semen	16.527*	1.154	<0.001	12.63	20.43
cg06379435_CpG2	Blood	Saliva	20.086*	1.360	<0.001	15.49	24.68
		Vaginal Epithelia	19.659*	1.633	<0.001	14.14	25.18
		Semen	18.909*	1.536	<0.001	13.72	24.10
VE_8_CpG3	Vaginal Epithelia	Saliva	-31.103*	3.211	<0.001	-41.95	-20.26
		Blood	-33.477*	3.480	<0.001	-45.23	-21.72
		Semen	-36.750*	3.552	<0.001	-48.75	-24.75
ZC3H12D_CpG2	Semen	Saliva	-53.659*	3.212	<0.001	-64.51	-42.81
		Blood	-61.782*	3.522	<0.001	-73.68	-49.88
		Vaginal Epithelia	-53.600*	3.824	<0.001	-66.52	-40.68

*. The mean difference is significant at the 0.0083 level.

With the understanding that only 5 CpGs are necessary for the cluster analysis to categorize body fluid samples, the cluster analysis method was repeated from the beginning, but using only the five most discriminatory CpGs. The resultant analysis of the 76 samples in the training set defined four clusters. Furthermore, analysis of the 46 test samples correctly categorized each of the samples into the correct body fluid cluster. The results of this new model are seen in the dendrogram in Figure 6.2. In this new dendrogram there are two sub-motifs for the blood and vaginal epithelia cell clusters. It is important to note that the SPSS software does not distinguish these sub-motifs as independent clusters; the individual samples are assigned the same cluster membership regardless of which sub-motif they are placed in. The methylation values at each CpG for these samples do not indicate why they would be considered as belonging to different subgroups and given the

limited information that is collected from donors via IRB consent forms, the reason for these sub-motifs is not apparent.

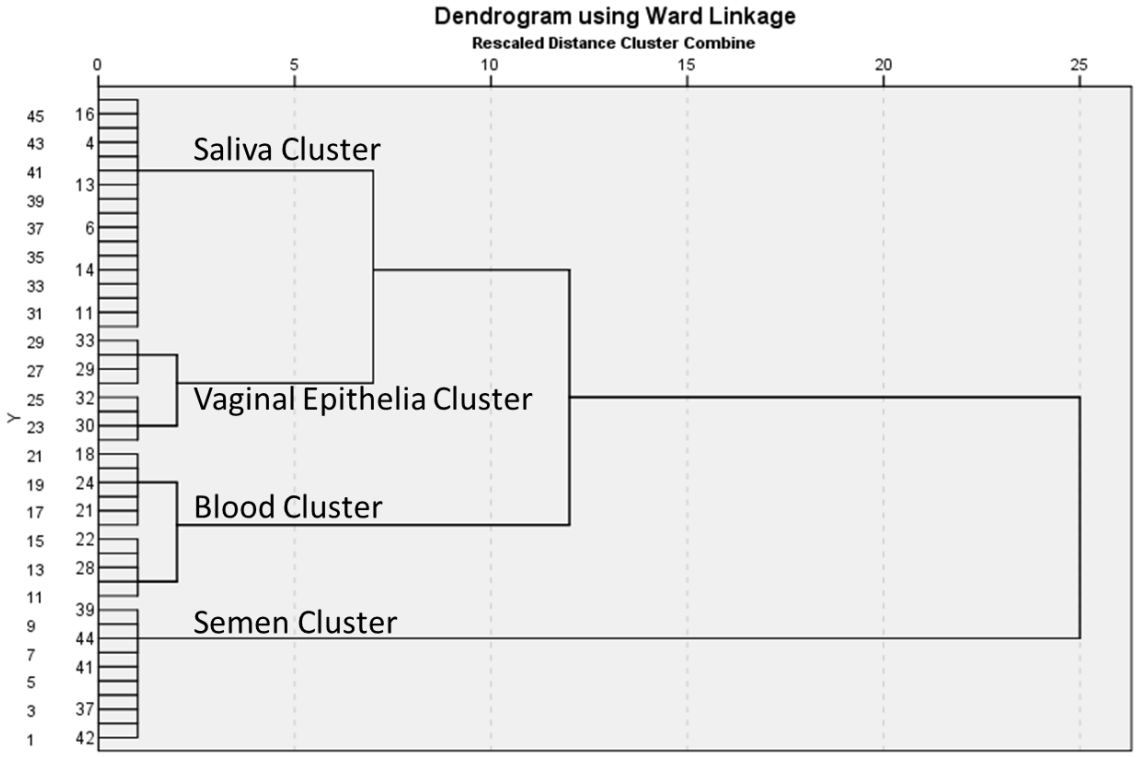

Figure 6.2 – Dendrogram resulting from the categorization of saliva (n=17), blood (n=11), vaginal epithelia (n=8) and semen (n=10) using 5 CpGs from the multiplex. Reproduced with permission from Gauthier, Cho, Carmel, and McCord, 2019.[181]

Within the dendrogram resulting from the use of 5 CpGs, the rescaled distance cluster combine demonstrated that the distance between the saliva and vaginal epithelia clusters was increased, as was the distance between those two clusters and the blood cluster. This fact, combined with the correct categorization of every sample, provided strong evidence that the more parsimonious model performs better than the model with all 18 CpG sites in the multiplex.

The successful identification of unknown samples using cluster analysis demonstrates that this method could easily be employed by a forensic laboratory. To implement this method, a laboratory would only need to develop a database of known body fluid origin samples and then use it to compare to unknowns. The number and type of samples incorporated into the model could also be explored and evaluated by individual laboratories.

C. Latent Profile Analysis Primer

Latent Profile Analysis (LPA) is a finite mixture model that proposes the use of underlying categorical variables to differentiate a population set into mutually exclusive latent profiles.[195] For the above data set, LPA presents an alternative method to calculate the probability that an unknown sample belongs to a latent profile that correlates to one of the body fluids. The LPA process makes assumptions about an unknown variable, *X*, that ties all observations within the model together in the context of the observable variables. In the LPA framework, *X* denotes the latent variable and the categories of *X* are the latent profiles. In latent profile analysis, all observed variables of a particular sample are called its manifest variables, and the set of manifest variables that are intended to directly measure *X* are called indicators.[196] For example, in the body fluid multiplex, the sequence data gives manifest variables in the form of nucleotides, but only a select few nucleotides, the CpGs of interest, would be considered indicators for *X*. Five indicators, *A, B, C, D,* and *E,* are used, each of which corresponds to the five most discriminatory CpGs that were observed in the Cluster Analysis. These are used to measure some unknown true value, *X*, which is theorized to correspond to the body fluid origin. In this analysis, *A, B, C, D,* and *E*, are all

observed variables, a requirement of latent profile analysis, with classes relating to percent methylation.[197] With the assumption of local independence, as was previously explored during the cluster analysis, the likelihood kernel, calculated probabilities of inclusion in each profile, for the ABCDE cross-classification table in terms of X is represented by the following formula:

$$\pi_{abcde}^{ABCDE} = \sum_{x} \pi_{x}^{X} \pi_{a|x}^{A|X} \pi_{b|x}^{B|X} \pi_{c|x}^{C|X} \pi_{d|x}^{D|X} \pi_{e|x}^{E|X}$$

where *X* is the latent profile variable, π_x^X the size of profile *x* and, , $\pi_{a|x}^{A|X}$ is the probability that variable *A* takes on the value *a* in the latent profile *x.*[198] The above equation describes the probability of seeing any combination of values for *a, b, c, d,* and *e* as depending solely on the differences in latent profile sizes combined with how different the profiles are in the context of the observed variables, *A, B, C, D,* and *E*.

The above formula is used to first approximate the makeup of a profile using the means and standard deviations of all samples used in the analysis. From there it is used to calculate the probability that each sample belongs to each of the defined profiles. Using the calculated posterior probabilities that an observation belongs to a specific profile, the parameters of the profiles are updated, and the posterior probabilities for all observations are recalculated. This process is continued until the parameters that define a profile stop changing, indicating that the profile is fully defined for a particular data set.[198]

Several software programs aid in Latent Profile Analysis. These programs range from early forms of the software such as MLSSA from Clifford Clogg, and LCAG from Jacques Hagenaars, to the more contemporary packages such as MPlus from Muthén & Muthén, and mclust, an open source statistical analysis package for Gaussian Mixture

Modeling. The mclust software package, which is used in the R coding language, has been favored by many data scientists for its ease of use, wide availability of sample data sets, and reliability. However, it does not directly perform for Latent Profile Analysis; as each individual step needs to be manually executed. To fix this, a second R-based package named tidyLPA was created.[199] This second package effectively acts as wrappers for the mclust functions, stringing all of the steps together and outputting the data in an accessible format.

All analyses for LPA took place using RStudio Desktop 1.3.1073 with mclust version 5.4.6 and tidyLPA version 1.0.8 and using the reference manual for the package provided by the authors.[199] To first identify the number of profiles that naturally exist in the data set, the training set of 74 known samples were used. The first instruction to the software was to first estimate the number of profiles that exist, up to 6, and to compare the resulting solutions using several parameters to determine the optimal number of profiles. To compare the solutions, the Akaike information criterion (AIC), Bayesian information criterion (BIC), Bootstrap Likelihood Ratio Difference Test (BLRT), and Entropy were used.[195] The AIC and BIC indicate how well the solutions fit the data set with lower values indicating a better solution fit. [200,201] The BLRT compares each solution to the neighboring solution with one less profile to determine whether the solution fits better with more profiles or not.[202] A significant value for the BLRT suggests that the increase in solution fitness compared to the previous solution is not due to random chance.[203] The entropy measures the accuracy of the classification with values approaching 1 indicating better classification.[204]

Upon determining the appropriate number of profiles that exist in the data set, the profiles were plotted (Figure 6.3), placing each individual sample into one of four profiles corresponding to body fluid according to the observed data of that sample at each of the 5 CpG sites. Based on the results obtained from this plot, each profile can be correlated to a specific body fluid. With the profiles set, a test set of 40 samples, 10 from each body fluid, was then tested against the model to determine its ability to identify unknowns. The results for the unknown samples (Table 6.4) are shown as the raw data at the five CpG sites, followed by the posterior probabilities that the sample belongs to each of the profiles. Based on the probabilities, the most likely identity of the unknown sample could be determined.

D. Latent Profile Analysis Results

LPA was used to identify the most likely number of profiles that exist within the dataset of 74 known samples consisting of 20 saliva, 20 blood, 17 vaginal epithelia, and 17 semen. The samples were formatted into a data.frame for the R package and uploaded into the local desktop application of RStudio. The tidyLPA program is used to load the data.frame, select the columns corresponding to the 5 CpG sites, impute the data, and then estimate the number of profiles that exist. The result suggested that there are four profiles that exist in the data set. To verify this, the package was instructed to calculate the AIC, BIC, BLRT and Entropy of the models containing one to six profiles. The AIC was minimized at the four-profile model, and the BIC minimized at the five-profile model. The BLRT suggested the four-profile model because the P-value for the five-profile model was not significant. The entropy of the four-profile model also indicated that correct

classification was maximized when compared to the other models. The high entropy value calculated for the five-profile model, and the fact that the BIC minimized at the five-profile model is likely related to the sub-motif of vaginal epithelial samples that was previously observed in the cluster analysis. The results, shown in Table 6.3, indicated that the model containing four profiles had the best fit for the data.

Table 6.3 – Fit statistics for 6 models and selection criteria for latent profile analysis. N = 74. AIC, Akaike information criterion; BIC, Bayesian information criterion; BLRT, Bootstrap Likelihood Ratio Difference Test.

Number of Profiles	AIC	BIC	BLRT	Entropy
1	4637.61	4572.38	-	-
2	4316.65	4349.59	P < 0.01	0.79
3	4278.17	4337.36	P < 0.01	0.87
4	**4248.16**	**4326.21**	**P = 0.01**	**1.00**
5	4256.73	4321.27	P = 0.06	0.96
6	4251.82	4328.16	P < 0.01	0.93

Once the model was confirmed using fit statistics, the package was next used to plot the data points on a graph, Figure 6.3. Each data point was color-coded based on its profile and was represented at each of the CpG sites utilized. Samples are initially sorted into random profiles and a latent profile variable is calculated that combines the information from each indicator. Each individual sample is then evaluated for the probability that it belongs to the profile it was placed in, or one of the other three profiles, based on the latent profile variable, *X*, for each profile. If a sample is moved to a new profile, the latent profile variable is updated. Samples continue to be moved between profiles in an iterative process until each sample is in the profile for which it has the highest probability of belonging.[205] The profiles were next categorized based on the pattern of methylation for the five CpGs to determine body fluid type by accessing the unknown variable that ties all of the samples to their respective profiles.

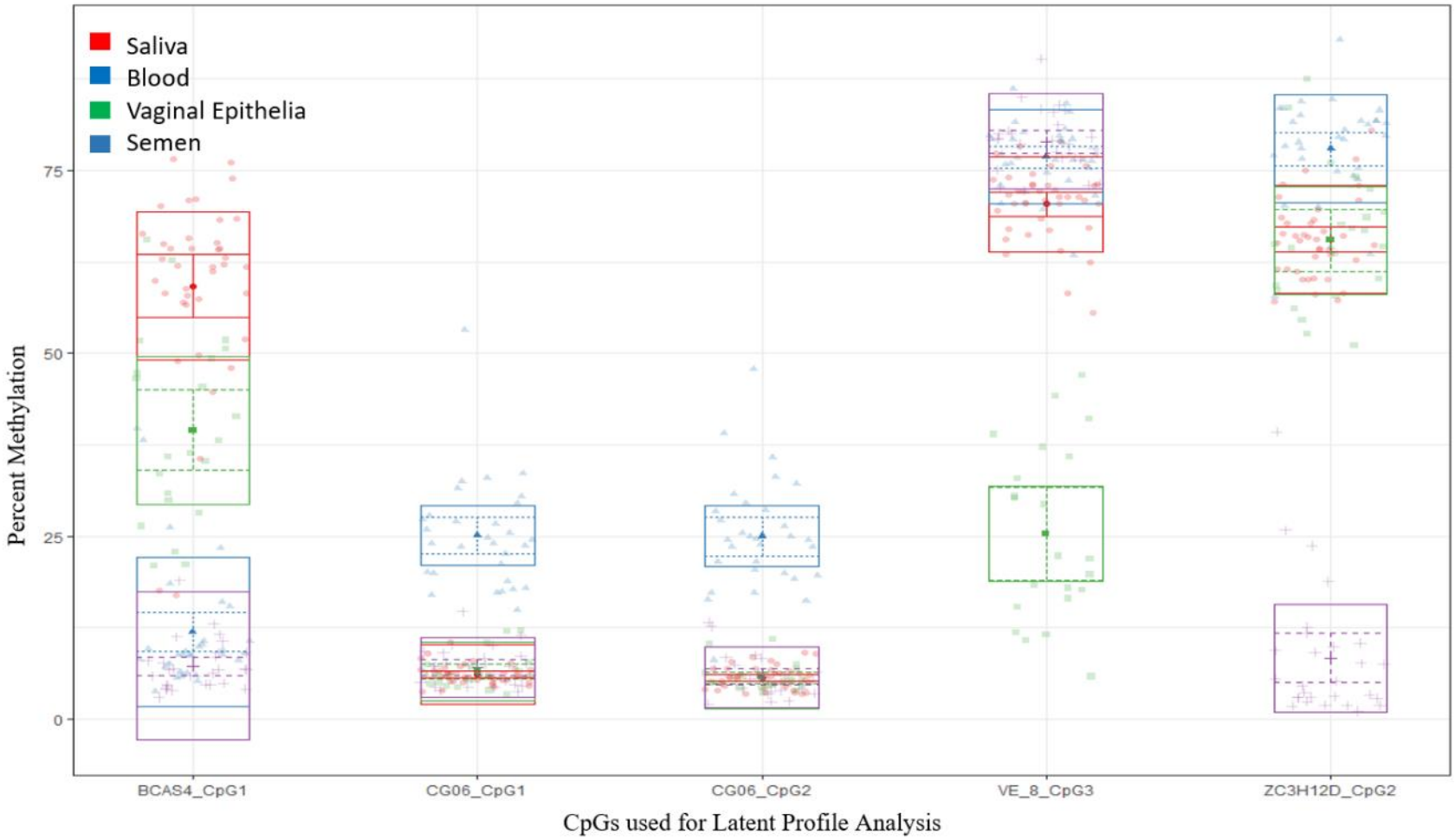

Figure 6.3 – Plotted profiles resulting from Latent Profile Analysis of 74 known samples. Bars reflect the 95% confidence interval of the profile centroid. Boxes reflect the standard deviation (+/- 64%) within each profile.

The next step was to apply the model to unknown samples. Forty samples that were deidentified from the user were processed by the model and a posterior probability for each of the four profiles was estimated. The results are seen in Table 6.4. In this approach, the posterior probability for each profile was considered for each sample. Whichever profile had the highest probability was used to determine the which body fluid the sample originated from. The identity of the assumed body fluids was then confirmed by comparing to the sample data prior to deidentification. In this test, all 40 samples were correctly identified, and the calculated probability for the corresponding profile could be used as a means to express the confidence in the answer.

Table 6.4 – Calculated posterior probabilities for 10 saliva, 13 blood, 9 vaginal epithelia, and 8 semen deidentified samples via Latent Profile Analysis. Each sample was calculated to have over 99% probability of belonging to the correct body fluid profile.

Sample #	Observed Methylation					Posterior Probability of				Body Fluid Identified
	BCAS4 CpG1	cg06379435 CpG1	cg06379435 CpG2	VE_8 CpG3	ZC3H12D CpG2	Profile 1	Profile 2	Profile 3	Profile 4	
30	8.79	29.5	28.5	84.1	84.4	3.07×10^{-21}	>0.999	1.00×10^{-34}	3.72×10^{-36}	Blood
17	49	6.32	6.27	72.9	67.8	>0.999	5.48×10^{-13}	1.49×10^{-12}	1.44×10^{-18}	Saliva
21	8.76	24.5	23.5	80.2	84.7	4.09×10^{-16}	>0.999	7.90×10^{-28}	1.15×10^{-31}	Blood
13	56.7	6.57	5.16	70.9	65.1	>0.999	2.43×10^{-15}	2.81×10^{-12}	3.48×10^{-19}	Saliva
33	57.4	5.25	4.83	78.2	76.5	>0.999	1.04×10^{-14}	9.08×10^{-16}	5.62×10^{-24}	Saliva
7	13	9.26	8.73	77.4	12.6	1.67×10^{-16}	5.30×10^{-24}	2.67×10^{-27}	>0.999	Semen
9	48	6.35	6.61	58.3	66.2	>0.999	9.45×10^{-14}	1.13×10^{-5}	6.92×10^{-19}	Saliva
8	4.78	7.19	5.54	77.3	4.58	7.02×10^{-22}	4.27×10^{-31}	5.92×10^{-32}	>0.999	Semen
22	6.74	6.02	6.5	79.5	3.36	3.55×10^{-22}	6.93×10^{-32}	1.81×10^{-33}	>0.999	Semen
34	6.22	19.9	20.4	86.1	92.8	5.43×10^{-14}	>0.999	2.83×10^{-28}	2.91×10^{-32}	Blood
11	5.91	4.41	2.12	80	3.67	3.28×10^{-22}	1.33×10^{-34}	1.14×10^{-33}	>0.999	Semen
3	7.02	22.6	24.6	80.6	79.7	1.48×10^{-15}	>0.999	2.69×10^{-27}	1.99×10^{-28}	Blood
31	29.9	6.17	7.49	44.1	74	7.38×10^{-4}	4.24×10^{-13}	0.999	7.00×10^{-23}	Vaginal Epithelia
15	3.92	3.92	2.44	80.4	1.08	7.50×10^{-24}	3.61×10^{-36}	2.60×10^{-35}	>0.999	Semen
6	65.6	5.12	5.91	63.5	66.6	>0.999	7.31×10^{-18}	1.34×10^{-9}	1.47×10^{-22}	Saliva
32	18.4	30.5	39.1	73	57.7	1.61×10^{-21}	>0.999	1.35×10^{-30}	2.39×10^{-27}	Blood
20	33.5	3.38	5.8	30.2	67.5	5.06×10^{-10}	9.00×10^{-24}	>0.999	3.65×10^{-28}	Vaginal Epithelia
19	50.6	6.88	5.69	11.9	87.4	4.36×10^{-17}	8.03×10^{-32}	>0.999	1.34×10^{-49}	Vaginal Epithelia
35	47.2	7.42	5.91	35.8	67.1	2.84×10^{-6}	2.40×10^{-20}	>0.999	1.18×10^{-26}	Vaginal Epithelia
38	10.2	32.9	28.5	79.2	78.8	7.84×10^{-22}	>0.999	4.30×10^{-33}	6.90×10^{-35}	Blood
25	35.8	6.06	4.89	30.4	58.6	9.60×10^{-10}	6.84×10^{-24}	>0.999	3.20×10^{-24}	Vaginal Epithelia
24	5.93	8.59	6.22	83.3	9.35	5.07×10^{-20}	1.45×10^{-27}	5.37×10^{-33}	>0.999	Semen
4	51.8	4.83	3.83	15.3	51.1	2.02×10^{-15}	1.20×10^{-35}	>0.999	2.23×10^{-31}	Vaginal Epithelia
18	36.3	6.22	5.22	37.2	69.2	1.52×10^{-6}	4.55×10^{-19}	>0.999	2.21×10^{-25}	Vaginal Epithelia
26	4.73	4.4	3.4	80	2.98	8.69×10^{-23}	2.10×10^{-34}	3.85×10^{-34}	>0.999	Semen
40	9.57	24.7	25	79	83.5	1.44×10^{-16}	>0.999	8.75×10^{-28}	8.01×10^{-32}	Blood
23	68.4	3.87	3.4	77.3	70.9	>0.999	6.67×10^{-19}	3.01×10^{-16}	6.02×10^{-24}	Saliva
39	26.4	6.54	6.28	47	68.6	7.81×10^{-3}	4.28×10^{-12}	0.992	2.65×10^{-18}	Vaginal Epithelia
27	4.53	4.58	2.44	83.9	2.94	3.35×10^{-23}	6.91×10^{-35}	2.26×10^{-36}	>0.999	Semen
28	16	24	26.4	79.7	79.5	2.99×10^{-15}	>0.999	2.55×10^{-27}	4.33×10^{-30}	Blood
12	65.5	12.1	10.3	5.8	66.7	9.58×10^{-19}	4.70×10^{-34}	>0.999	1.68×10^{-45}	Vaginal Epithelia
2	3.78	17.8	8.09	76.4	73.7	5.49×10^{-5}	>0.999	1.40×10^{-14}	6.32×10^{-15}	Blood
14	7.53	27.7	24.5	76.4	75.3	2.90×10^{-17}	>0.999	4.77×10^{-27}	1.69×10^{-28}	Blood
5	10.7	33.5	35.8	76.3	82.5	1.03×10^{-25}	>0.999	1.09×10^{-35}	8.84×10^{-41}	Blood
36	66.4	7.66	5.76	72.2	65.9	>0.999	2.73×10^{-16}	1.12×10^{-13}	1.47×10^{-21}	Saliva
16	8.09	31.5	32.2	78.9	78	3.30×10^{-23}	>0.999	3.66×10^{-34}	1.78×10^{-35}	Blood
1	59.9	5.3	3.88	71.3	67.8	>0.999	5.98×10^{-17}	9.26×10^{-13}	3.80×10^{-21}	Saliva
29	71	6.93	3.45	66.8	65.4	>0.999	4.23×10^{-19}	1.51×10^{-11}	6.91×10^{-23}	Saliva
37	15.4	25.4	25.4	70.5	81.8	3.55×10^{-15}	>0.999	6.45×10^{-23}	1.11×10^{-31}	Blood
10	35.7	6.24	4.19	78.8	66.1	>0.999	3.94×10^{-11}	3.29×10^{-14}	2.53×10^{-14}	Saliva

E. Concluding Remarks

The two models described above provide a path for the body fluid identification multiplex to become a powerful tool for determining the origin of an unknown sample. The methylation data that is developed in the lab can be interpreted in an impartial manner that quickly and accurately places the unknown into one of four body fluid groups. These methods could be easily implemented in a forensic lab through the use of either a universal database or by creating an internal database based on known samples prepared in the course of an internal validation study.

It should be noted however that cluster analysis does not presently offer a standard measure of statistical confidence that a sample has been placed into the correct cluster. Instead, to gauge the confidence of identifying an unknown, the model's history of identification would have to be used. The process of incorporating a single unknown sample at a time into the model for identification and then reporting how many times in the past the model correctly identified the samples would offer a measure of reliability. Similarly, Latent Profile Analysis does not specifically identify a sample as being from a body fluid. Instead, it presents the probability that an unknown is consistent with a profile that we have assumed to correlate with body fluid origin.

As they currently exist, both the cluster analysis and LPA results dictate that each sample must end up in the four specified groups. They do not allow for a sample to be placed into any other group that may exist outside of the model. This means that if an unknown sample that is not saliva, blood, vaginal epithelia, or semen were analyzed and the methylation data plugged in to the models, it would be misidentified. Similarly, if a mixture of two or more body fluids were encountered, the cluster analysis would insist on

placing that sample into a single group. The LPA would attempt to calculate the probability of a sample belonging to a profile that is representative of single source samples even though this unknown is made up of multiple sources.

For these reasons, future work should focus on the analysis of mixture samples to create new models with a larger number of clusters or profiles that a sample can be placed in. Ideally, these mixture profiles would allow for more accurate identification, but care would be necessary in order to deal with samples such as menstrual blood. Additionally, as the four body fluids tested do not comprise every possible source of DNA from the human body, it would be helpful to expand the data to more body fluids, such as sweat, menstrual blood, urine and bile, or possibly create a cluster and profile that is defined as being not saliva, blood, vaginal epithelia, or semen so that nonsense samples wouldn't be erroneously classified as one of the four body fluids.

CHAPTER VII – BODY FLUID IDENTIFICATION AND AGE DETERMINATION USING A TARGETED METHYL NEXT GENERATION SEQUENCING APPROACH

A. Introduction

The next goal in implementing DNA methylation analysis into the forensic workflow is the inclusion of a method that gives body fluid identification, age determination, and other lifestyle traits in a single tube. One of the pressing demands of any forensic laboratory is to get the most amount of information possible while using the least amount of DNA in order to preserve a precious and limited crime scene sample. Body fluid identification and age determination assays have seen significant development in the past few years and as previously mentioned a sample's body fluid origin can influence the accuracy of age prediction models.[206] Multiplex approaches to body fluid identification have produced the ability to identify a variety of body fluids such as saliva, blood, vaginal epithelia, semen, and menstrual blood from a single tube.[64,181] This data could be combined with other epigenetic information such as age and phenotype.

With the ability to quickly and efficiently identify body fluids from methylation status, the use of DNA methylation for age determination becomes a much more tangible goal in forensics. The first study to examine the methylation status of the genome with age dates back to 1967 when Berdyshev et al. examined the life stages of spawn humpback salmon.[207] They found that the methylation of the humpback salmon decreased with age in a reproducible manner. And although the phenomenon was reproducible in other species, the ability to reliably examine methylation status and correlate it with age has only become possible in more recent years with the various advances in accuracy with modern instrumentation [208]. Recent advances in microarray technologies capable of probing vast

numbers of methylation sites in the human genome have enabled researchers to perform genome wide association studies in order to identify accurate predictors of age.[209] These microarray studies have produced numerous models that utilize a variety of CpG sites for the prediction of biological age in tissue samples.[206,210] Unfortunately, this approach requires large quantities of DNA, and is time consuming and expensive, making it less viable for routine analysis. For the purpose of discovering new CpG sites, however, the microarray studies offer the invaluable ability to identify regions that can be interrogated further with cheaper and quicker methods. This approach has led to the discovery of a number of methylation sites that have been found to be highly correlated with age, such as ELOVL2, ASPA, KLF14, FHL2, and many more.[60,61,165,211]

In a similar fashion to the body fluid assays, a plethora of assays examining the correlation of age and DNA methylation have produced models capable of predicting age from blood samples with a Mean Absolute Deviation (MAD) approaching just 3.4 years.[163] Similarly, Jung et al. have developed an age prediction assay for saliva samples using CpGs found in the same five genes that the Zbieć-Piekarska et al. study utilized with an MAD of 3.5 years.[212] Many more assays exist with overalpping genes and CpGs of interest, however nearly all of them suffer from same deficiency: they utilize singleplex reactions that end up using significant amounts of DNA.

One significant outlier to that trend is the work done by Hwan Young Lee and her associates at Yonsei University College of Medecine. Lee's group has directed significant effort to the development of multiplex DNA methylation assays capable of differentiating body fluids and predicting age in saliva, blood, and semen.[64,212,213] This work, utilizing the SNaPshot kit, has allowed for the determination of body fluid and prediction of age to

occur in just three separate tubes, dramatically reducing the amount of DNA and reagent costs. Additionally, the results are in the form of an electropherogram which is already familiar to forensic laboratories and straightforward for interperetation. There is a drawback, however, to the SNaPshot approach; it targets only the specified CpG and ignores any other possible CpGs surrounding it. As mentioned earlier, there are a number of studies that target the same regions of genes for different CpG sites in their respective prediction models. An assay that analyzed all CpGs in a given region, and from multiple regions simultaneously, would empower the user to mix and match CpGs for different predictve models providing a lot more flexibility.

One way to achieve this goal is to use Next Generation Sequencing in the assay. NGS assays excel when there are multiple targets for sequencing and allow for all of those targets to be probed from a single tube. Advances in the technology have allowed for the creation of Targeted Methyl sequencing panels that can accurately target any region of the human genome for CpG analysis and allows for hundreds of targets if needed. Additionally, NGS provides easy to achieve multiplexing of samples using barcodes to dramatically increase the number of samples in a single sequencing reaction which ultimately decreases the cost to the user.

The focus of this chapter is the development of a Targeted Methyl Sequencing panel that will provide methylation data for a number of published age prediction assays as well as the body fluid identification assay described in Chapter 4. This large panel will allow for the body fluid identification and age prediction of a sample from a single tube while using similar quantities of DNA as the previously described methods and in a format that

would allow for more targets to be added in the future without significant change to the methodology.

B. Selection of Assays

The targeted methyl sequencing panel includes loci for body fluid identification, age prediction, and lifestyle traits and consists of a total of 9 different published assays – 1 body fluid assay, 7 age predicition assays, and one assay for the use of tobacco. A detailed list of the CpG sites being probed in each assay can be seen in Table 7.1. The body fluid identification assay consists of the four markers published by Gauthier et al. in the McCord research group including BCAS4 for saliva, cg06379435 for blood, VE_8 for vaginal epithelia and ZC3H12D for semen.[181] Also from the McCord research group are the age prediction assay for both saliva and blood samples from Alghanim et al. consisting of the genes KLF14 (Kruppel-Like Factor 14) and SCGN (Secretagogin) and the smoking prediction assay using CpG sites from the AHRR (aryl hydrocarbon receptor repressor) gene.[60,162] The age prediction model using KLF14 and SCGN has been validated to predict ages in both saliva and blood with a MAD of 7.1 years and 10.3 years, respectively. Although not the most precise model included in the larger panel, it has the benefit of requiring the methylation status of just three CpG sites in two amplicons, resulting in a straightforward and parsimonius model to interpret. Additionally, as the purpose of the model is to increase the number of assays available for analysis, it is possible that the results of this prediction model could be combined with other models to increase accuracy. The inclusion of the assay for AHRR, fairly distinct from the intended purpose of body fluid and age prediction, is to demonstrate this technique's ability to adapt for a variety of

purposes. As target enrichment during library preparation can be extremely specific, there is the possibility to add a large number of predictive assays to this current panel which can help investigators. The AHRR model developed by Alghanim et al. is capable of predicting whether an individual is currently a smoker, a former smoker, or has never smoked with accuracies in saliva and blood at over 82% and 71%, respectively.

Amongst models chosen from the literature, there were several that have been optimized to predict age in a single body fluid at a time. Although there was some overlap in the models that are used, each one uses a different formula to combine the methylation data that enables age prediction. The first assay chosen was the age prediction in saliva model from Jung et al. This assay utilizes the methylation status of one CpG each from ELOVL2 (Elongation Of Very Long Chain Fatty Acids protein 2), C1orf132/MIR29B2CHG (Chromosome 1 open reading frame 132/ MicroRNA 29b-2 and 29c Host Gene), TRIM59 (Tripartite Containing Motif 59), KLF14, and FHL2 (Four And A Half LIM Domains 2).[212] This assay was reported in the literature to have a MAD of 3.6 years. Additionally for saliva, the assay from Eipel et al. uses one CpG each from ASPA (Aspartoacylase), ITGA2B (Integrin Subunit Alpha 2B), and PDE4C (Phosphodiesterase 4C). This assay predicts age from saliva samples with a MAD of 4.3 years.[214]

For the prediction of age in blood samples, Xu et al.'s model using CpGs from ADAR (Adenosine Deaminase RNA Specific), ITGA2B, and PDE4C has a MAD of just 2.8 years.[215] In addition, a model from Zbieć-Piekarska et al. using CpGs from ELOVL2, C1orf132, TRIM59, KLF14 and FHL2 has a MAD of 3.4 years.[163]

For the prediction of age in semen samples Lee et al.'s assay utilizing one CpG each from TTC7B (Tetratricopeptide Repeat Domain 7B), cg12837463, and NOX4 (NADPH oxidase 4) gives a model with a MAD of 4.2 years.[213]

And finally, for the determination of age in individuals that have undergone severe decay and for which body fluids are no longer an option, the age prediction model using teeth as a DNA source from Bekaert et al. was included. This model, utilizing CpGs from PDE4C, ELOVL2, and EDARADD (EDAR Associated Death Domain), gives an age prediction with a MAD of 4.8 years.[216]

Table 7.1 – Assay information for the custom Targeted Methyl Sequencing panel for body fluid identification, age prediction, and smoking status.

Function	Gene Name	Chromosome	# CpG sites	hg19/GRCh37 CpG sites	MAD (years)	Source
Body Fluid Identification	BCAS4	chr20	1	49,410,865	N/A	Gauthier et al.
	cg06379435	chr19	2	3,344,242; 3,344,251		
	VE_8	chr16	1	86,398,467		
	ZC3H12D	chr6	1	149,778,105		
Smoking	AHRR	chr5	4	373,476; 373,490; 373,494; 373,529	N/A	Alghanim et al.
Saliva/Blood Age	KLF14	chr7	2	130,418,281; 130,418,311	±7.1	Alghanim et al.
	SCGN	chr6	1	25,652,606		
Saliva Age	ELOVL2	chr6	1	11,044,861	±3.6	Jung et al.
	C1orf132	chr1	1	207,997,026		
	TRIM59	chr3	1	160,167,977		
	KLF14	chr7	1	130,419,116		
	FHL2	chr2	1	106,015,739		
Saliva Age	ASPA	chr17	1	3,379,567	±4.3	Eipel et al.
	ITGA2B	chr17	1	42,467,728		
	PDE4C	chr19	1	18,343,915		
Blood Age	ADAR	chr1	2	154,582,187; 154,582,288	±2.8	Xu et al.
	ITGA2B	chr17	1	42,467,780		
	PDE4C	chr19	5	18,343,915; 18,343,937; 18,343,941; 18,343,943; 18,344,003		
Blood Age	ELOVL2	chr6	1	11,044,867	±3.4	Zbieć-Piekarska et al.
	C1orf132	chr1	1	207,997,026		
	TRIM59	chr3	1	160,167,987		
	KLF14	chr7	1	130,419,116		
	FHL2	chr2	1	106,015,745		
Semen Age	TTC7B	chr14	1	91,283,606	±4.2	Lee et al.
	cg12837463	chr7	1	35,300,228		
	NOX4	chr11	1	89,322,851		
Teeth Age	PDE4C	chr19	1	18,343,915	±4.8	Bekaert et al.
	ELOVL2	chr6	5	11,044,858; 11,044,861; 11,044,867; 11,044,873; 11,044,888		
	EDARADD	chr1	1	236,557,695		

C. Methods

Buccal swabs, blood, vaginal swabs and semen samples were collected from volunteers under the conditions set forth under the approved protocol of IRB-17-0210 from Florida International University. Swabs were air-dried before being stored at -20 ºC or proceeding directly to extraction.

DNA extraction was performed by automated extraction protocols. Automated extraction and purification were performed using the EZ1® DNA Investigator kit (Qiagen, CA) and the BioRobot® EZ1 automated purification workstation (Qiagen, CA) according to the manufacturer's specifications, detailed in Appendix 1. Samples were eluted in volumes of 40 μL TE buffer.

DNA Quantification was performed using the ALU qPCR and Rotorgene thermal cycler method as described in Appendix 1. After concentration was determined, 200 nanograms of DNA were bisulfite modified using the EpiTect® Fast DNA Bisulfite Kit (Qiagen, CA) according to manufacturer's protocol, as detailed in Appendix 1. The elution volume after modification was 20 μL. Concentration of samples after bisulfite conversion were verified using the Qubit™ ssDNA Assay Kit (Invitrogen™, Carlsbad, CA).

Library preparation of samples was carried out according to Qiagen's protocol for Targeted Methyl Sequencing Library Preparation for genomic DNA, as detailed in Appendix 1 with a targeted input DNA of 40ng as sample allowed. This process, as described in Chapter 3, includes the end-tail repair of DNA fragments, the incorporation of barcoded adapters on either end of the regions of interest, target enrichment and universal PCR with magnetic bead purification steps throughout to remove unincorporated

primers and leftover components of the previous enzymatic reaction. Details for the Target Enrichment PCR primers can be seen in Table 7.2.

Table 7.2 – Primer sequences for each targeted region as designed by Qiagen for the custom Targeted Methyl Sequencing kit. S = Primer targeting sense strand. A = Primer targeting antisense strand.

hg19 Chr Total	hg19 Start	hg19 End	# of Primers 38	Primer Sequences
chr11	89322800	89322899	2	S – CATAACTAACACCAACTACAACCAACCATTTAAATAAAATA
				A – ATCAATCACAATACCCTACCCAACAACTTTT
chr14	91283550	91283649	2	S – AAAAACACAATCACTAATAAAACCTCCTATCTTAACCA
				A – TTTACTTATTTTTCCCCCAAACTAAAATATAAACTCTCTCA
chr16	86398378	86398508	3	S – AAATATAAATCTCCTATAACTACTATAACCACCAAAAACCA
				S – CCTTTTCCCTCTTCCAACATCTATTAACTACTAA
				A – AATAAAATCATCCCAAATTATCCAAACAACCACTAAAC
chr17	3379500	3379599	2	S – ACCCTTTAAATAAAATCTCATTTACATTTCTAAACCTTTCT
				A – TCAAATCACAATCAATATATCTAATACACTTCTTCACTACT
chr17	42467700	42467799	2	S – AACCCTCAATCCTTTTTAAATAATAAAACTCTTTAACCATT
				A – CTCTAAAACTATAACAAAAAACCTTACTCCCAAAAAACTC
chr19	18343875	18344024	2	S – CTCAACCTACTACAAACCTCTACCCCTT
				A – CTACTCCCTACTATCCCAAACCCCTTT
chr19	3344113	3344322	4	S – CAAAATCACACAACACAATAAAATAAAACCACCTTCTATA
				S – AAAACCCAAACCATACCACTATTACAAAATCTAAAAAC
				A – ACAACAAAACAACTATCTCTAATTAAACCCTACTTTCC
				A – CAAAAAAACCCACCTCAAACCCTTTCATAA
chr1	154582175	154582299	2	S – ATTACTAAACRCCTACCCCTAATAAAACACTTACACACTAC
				A – AATTACTAAAACATCCATCTTCCCTAACAAACTAACTAA
chr1	207997000	207997099	2	S – ACACAAAAACAACRCCCCTAATCCCAACAAATACATA
				A – AAACCAAAAACCTCTAAATAACCTAAACTAAAAAATAACAC
chr1	236557650	236557749	2	S – ACCTACAAATTCCCCAAAAAACTTTCATCTAAAAAATTTA
				A – ACAAATACCTACATACCCTCTTAATAACCAAAAACTTTAAT
chr20	49410801	49410958	2	S – CTCTTCAACCCCAAAACTTATAAAAAATCTATCTAAACC
				A – CCCCACCCRTTCCCATCCTACTAAAACATC
chr2	106015700	106015799	2	S – RATCCCRACCCRTACCCTTTATTTACCAAAACTCCTTTCT
				A – CAAACACATACCTCCTAAAAAATAACCCCCTC
chr3	160167925	160168024	1	A –AAAAAACACTACRCTCCACAACATAACAAAAAACCCC
chr5	373450	373549	2	S – AAAAATAAACCCTAAAAATAATCCTAACAAAACCCCTC
				A –AAAATAAACCATCACCRTAACCCTTACAAAACACAACTAAA
chr6	11044828	11044927	1	A – CATTTCCCCCTAATATATACTTCAAACCCACC
chr6	149778061	149778150	1	A – CCTACCTACTTAACCTAAAACTTCAAACAAATTCAA
chr6	25652575	25652699	2	S – CTCCCCAACAACAATTACTCAAAACTAATCAAATAAC
				A – CTACACCTAAATATACAAATAACTTATTCTACTCACCATCC
chr7	130418250	130418349	1	A – ACCAACAACCTCTAATAAATTCTCTAAAAAAACCCT
chr7	130419075	130419174	1	S – RACCCCCCRACTAAATCATATTTAACAACCTCAAAA
chr7	35300175	35300274	2	S – ACTAAAAACCCACAACAAACCTCAAAACTAAATTTT
				A – ACAAATAAATTCCTTAACCTTCTAAACTTCATTTTCTACAA

After library preparation, all samples were analyzed on the Agilent 2100 Bioanalyzer using the High Sensitivity DNA kit as a quality control step in order to verify the fragment sizes and concentration of the library. After quantification, the libraries were normalized to 1.8nM using Tris-HCl 10mM/pH 8.5 and 0.1% Tween (EBT buffer) and

pooled together in equal amounts for a final library pool volume of 140 μL. After denaturation and dilution to 9pM, 570 μL of library pool were combined with 30 μL of 20pM PhiX internal control (Illumina, Inc). The PhiX control library from Illumina allows for a higher level of nucleotide diversity throughout the sequencing run. This higher level of nucleotide diversity is critical during the first seven rounds of cycling on the MiSeq as this is the time that the instrument is identifying and segmenting the surface of the flow cell according to the observed clusters. If too many neighboring clusters have similar nucleotide content in the first seven rounds of cycling, then the instrument will not be able to differentiate the clusters later on in the cycling, and the Q-score of the sequencing run will suffer.

Libraries were analyzed in a MiSeq FGx in research mode v1.3.1 using a MiSeq Reagent Kit v3. The Qiagen Custom Sequencing Read 1 Primers (3.4 μL) were spiked into the Illumina sequencing primer well on the cartridge for a final concentration of 0.5μM. The library pool was loaded on to the cartridge and the sequencing run was performed using a sample sheet generated using the Illumina Experiment Manager v1.19. The instrument was set to perform a paired-end sequencing of 151bp in each direction and the data generation was set to FastQ files only. BaseSpace ®, Illumina's online platform, monitored the run and was used to retrieve the files and transfer them to Qiagen's GeneGlobe for data analysis.

The GeneGlobe Targeted Methyl Sequencing analysis pipeline was used for data interpretation. The pipeline automatically ingested the FastQ files, conducted trimming of the sequences based on Q-score, deduplication of reads according to Unique Molecular Identifiers, alignment to a human bisulfite converted reference genome, annotation of each

identified CpG site, and calculation of the percent methylation observed for each CpG and provided a report of the process as well as an excel table containing the final results. From this excel sheet, the percent methylation of CpGs specified in Table 7.1 could be used to predict body fluid, and from there the age of the individual and their smoking status.

D. Results and Discussion

Fragment Analysis

As a quality control check prior to sequencing, all samples were analyzed on the Agilent 2100 Bioanalyzer. Figure 7.1 is an example of the resulting electropherogram that was seen for most samples. It shows a significant portion of the final library being comprised of DNA molecules at 246 bp in length, with longer fragments being observed all the way until 800 bp in length. These results indicate an overamplification of the smaller targets in the panel. However, given that most of the targets in the panel have a target region of just 100 bp, the library size of 246 bp, which includes the adapters and primer binding regions on either side of the target region, is consistent with the desired PCR product size.

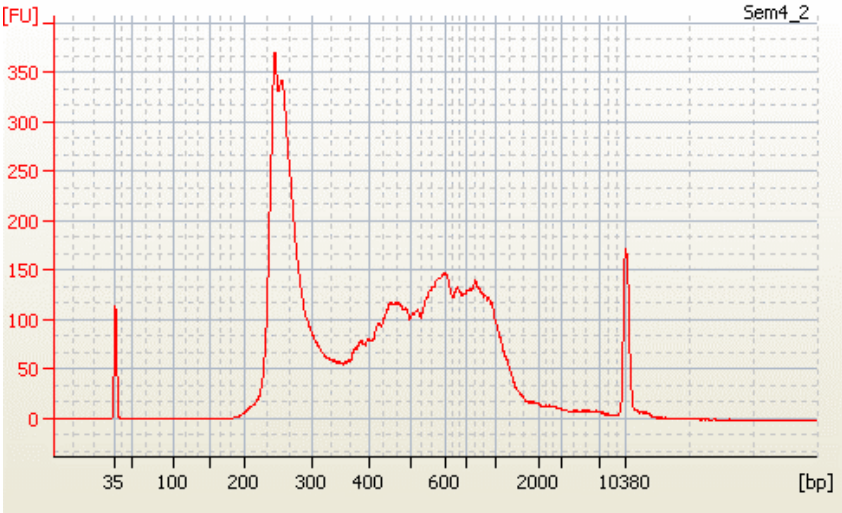

Figure 7.1 – Electropherogram showing fragment analysis of sample Semen4_2 after library preparation. Fragments are primarily 246bp in length.

As determined by the quality control of the library preparations via Agilent 2100 Bioanalyzer, nearly all – 28 of 32 – samples were determined to have sufficient quantities of DNA for subsequent sequencing reactions. The distribution of fragments was extremely consistent across all samples – fragments were primarily centered around 246 base pairs which is consistent with most of the target regions being approximately 100 base pairs long and with the adapters, barcodes, and primer binding regions added. The lower concentration of the larger fragments, which includes the cg06379435 target region spanning 209 base pairs as the largest, proved to not be of particular concern after sequencing – the cg06379435 targeted region ended up having higher coverage in most samples than some of the smaller fragments. The criteria for inclusion in sequencing after library preparation quality control was the ability to have a total concentration of 1.8nM in the pooled libraries. With this criterion in mind two saliva samples and two semen replicates were eliminated from the pool. These four samples showed no signal at all on the electropherogram suggesting that the samples failed to amplify or were lost during

library preparation. Loss of sample is not uncommon during the library preparation process which consists of many tube changes and sample purification steps that can introduce error.

After library quality control checking, the libraries were pooled together at equimolar volumes based on the concentration calculated from the area under the curve of each sample's electropherogram. After pooling, the libraries were loaded in to the MiSeq reagent cartridge with a 5% spike-in of 20 pM PhiX and sequencing began.

Sequencing Quality Control

The first metric for the quality of the data was the observed Q-scores for the base calls during the sequencing run. In Figure 7.2, the Q-scores for over 87% of the called bases are over 30. This means that 87% of the nearly 1.2 billion bases sequenced in the run has less than a 0.1% chance of including a miscalled base. This instills a high confidence that the sequence data for each of the samples will be highly accurate for the called bases. With accurate base calling, the percent methylation at each of the CpGs of interest can be calculated by comparing the proportion of reads containing a cytosine at the CpG site with the total number of reads for that CpG site.

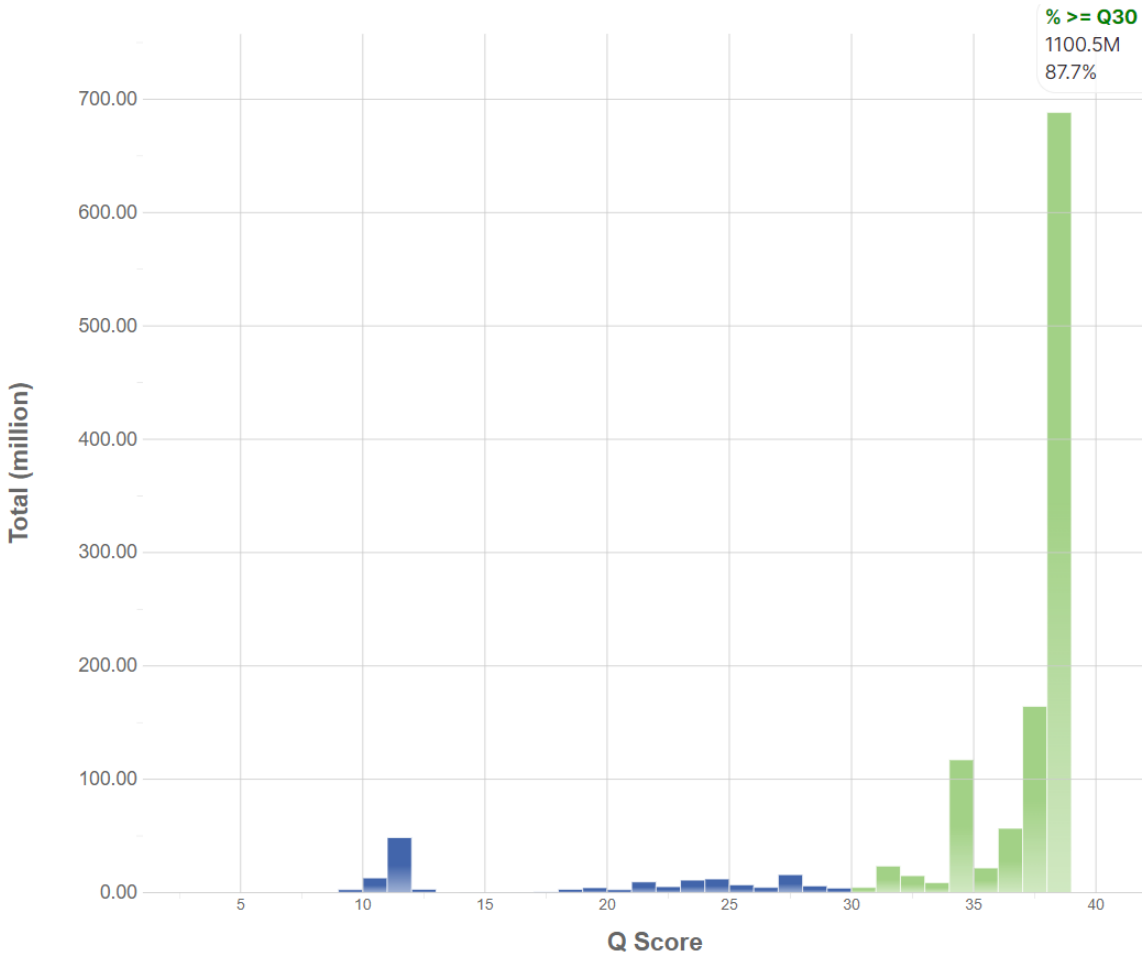

Figure 7.2 – Q-score distrubution for basecalls during the Targeted Methyl Sequencing from Illumina's BaseSpace analysis. Q-score of 30 (99.9% probability of accurate basecall) is considered the threshhold for quality sequencing data.

The primary metric for evaluating the quality of sequencing data, the Q-score, for each cycle and read of the sequencing reaction indicated that the samples that were run on the MiSeq were of excellent quality for sequencing. The overwhelming majority of the reads contained Q-scores over 30. This means that the sequence data contained in the FastQ files generated by the MiSeq could be reliably analyzed for methylation status at each of the targeted CpG sites.

Secondly, Illumina's BaseSpace analysis of the metadata for the sequencing run showed that just over 84% of the reads were identified and separated based on the

barcoding indices that were incorporated to the different samples during library preparation. Given the Q-scores and the inclusion of the PhiX control, which is not indexed and represents approximately 10% of the DNA loaded on to the flow cell, this percent of identified reads means that only about 5% of the sequence data was of such poor quality that it could not be assigned to a sample.

Following adapter trimming, sample grouping, alignment to reference sequence, and deduplication of UMI sequences, the methylation state of each observed CpG site was calculated by dividing the number of cytosines observed at each genomic location by the total read depth at that location. The resulting percent methylation can then be used to compare the results of the targeted methyl sequencing reaction to the literature values.

What is readily apparent in the results across all samples, Appendix II, is the wide variation in coverage across all of the target regions. While some CpG sites have read depths between 100x and 200x, other targets produced read depths under 25x, which makes the determination of percent methylation inconsistent due to stochastic effects as was observed in the sensitivity study of the pyrosequencing multiplex. This coverage falls well short of the 1000x coverage recommended in the literature for accurate methylation calling.[217] Additionally, the handbook provided with the QIAseq Targeted Methyl Sequencing Custom panel indicates that targets should have a mean coverage of 500x across the entirety of the panel. If a threshold of at least 250x coverage were to be applied for each CpG across all assays, none of the assays would have sufficient read depth. However, although the results fell well below this threshold for analysis, there were some encouraging results.

Results for body fluid identification assay

Table 7.3 – Results of Body Fluid Identification assay in the Targeted Methyl Sequencing panel for 26 body fluid samples. Saliva, blood, and vaginal epithelial samples produced methylation profiles consistent with the profiles produced by the body fluid identification multiplex via pyrosequencing. Semen samples produced methylation values at the BCAS4 and ZC3H12D CpGs that are inconsistent with semen as the source body fluid.

Assay	Body Fluid Identification				
CpGs	BCAS4 CpG1	cg06379435 CpG1	cg06379435 CpG2	VE_8 CpG3	ZC3H12D CpG2
Chromosome Position	20; 49,410,865	19; 3,344,242	19; 3,344,251	16; 86,398,467	6; 149,778,105
Saliva2	78.8	0.0	0.0	73.5	70.6
Saliva4	60	0.0	0.0	94.4	100
Saliva5	81.6	0.0	0.0	62.9	61.5
Saliva7	36.8	1.0	0.0	70.4	82.1
Saliva8	X	0.0	0.0	77.3	X
Bl1	1.6	28.0	27.3	79.5	100
Bl2	0.0	21.1	16.8	74.3	84.0
Bl3	9.3	22.4	21.7	77.4	100
Bl4	13.6	1.9	8.6	84.0	100
Bl5	7.2	8.5	8.5	83.7	91.7
VE10	50.9	0.6	0.0	3.2	78.9
VE30	27.0	0.0	0.0	19.4	95.2
Sem1_1	60.3	0.0	0.0	83.9	59.5
Sem1_2	64.2	0.0	0.0	67.9	79.5
Sem2_1	42.4	2.3	2.2	72.7	100
Sem2_2	57.2	0.9	0.3	71.6	76.9
Sem3_1	22.4	0.0	0.0	74.7	93.2
Sem3_2	22.1	0.6	0.9	71.5	91.2
Sem4_1	32.8	1.1	1.0	70.5	78.9
Sem4_2	39.5	1.0	1.6	65.0	80.4
Sem5_2	66.7	1.9	1.8	71.7	64.4
Sem6_1	54.3	0.4	0.0	63.0	80.5
Sem6_2	52.7	0.0	0.5	77.3	85.5
Sem7_2	58.0	1.4	1.0	70.4	70.8
Sem8_1	61.7	0.0	0.5	52.1	93.5
Sem8_2	50.5	0.0	1.1	74.8	95.7

Table 7.3 shows the compiled results of the body fluid identification multiplex assay within the Targeted Methyl Sequencing run. Saliva, blood and vaginal epithelial samples run on the MPS panel resulted in the methylation profiles consistent with the methylation profiles that were developed in the population study of the body fluid identification multiplex via pyrosequencing. This is despite the fact that many of these samples had less than 50x coverage at nearly all CpG sites. The semen samples produced methylation levels consistently over 90%, which is not in line with any of the body fluid profiles. In fact, the methylation values observed in the ZC3H12D marker were consistently higher across all sample types for the MPS panel when compared to pyrosequencing. Similarly, the CpG sites of the cg06379435 marker were much lower in saliva, vaginal epithelia, and semen samples for the MPS panel when compared to pyrosequencing. These results may be caused by PCR bias of either the methylated or unmethylated strands for the primers targeting cg06379435 and ZC3H12D or generally poor amplification efficiency with the MPS panel as designed.

Results for a blood age prediction model

To combat the generally low coverage of many markers across the blood age prediction CpG assays, a blood age prediction model was pulled from the literature. One of the biggest benefits of massively parallel sequencing is that although the panel may have been designed to target a select number of CpG sites, the surrounding sequences are also captured, and that data can be utilized for analyses beyond the initial intended assays in the panel. From the literature, and additional model from Zbieć-Piekarska et al. was identified that utilizes two CpG sites from the ELOVL2 marker to predict age in blood with a MAD

of 7.2 year.[164] That this model uses two CpG sites at are just eight bases away from each other means that the coverages for the two CpG sites are fairly equal, and so the combination of methylation data results in accurate methylation calls.

Table 7.4 – Results of blood age prediction for the 5 blood samples using the two CpG model from Zbieć-Piekarska et al. Four of the five sample's predicted age is within the expected MAD of 7.2 years.

Assay	Zbieć-Piekarska et al. Blood Age assay			
CpGs	ELOVL2 CpG7	ELOVL2 CpG5		
Chromosome Position	6; 11,044,867	6; 11,044,875	Predicted Age	Actual Age
Bl1	80.3	18.2	39.15	34
Bl2	61.9	15.6	21.32	26
Bl3	84.8	17.4	42.56	47
Bl4	83.3	36.0	53.06	63
Bl5	54.8	15.5	15.10	19

The predicted ages of the blood samples using the simple model by Zbieć-Piekarska et al. indicate that when similar coverage is observed (~60x for each CpG in all 5 samples), then the combination of the methylation data to predict age can be used with reasonable accuracy for the predicted age. The results that were obtained in the massively parallel sequencing run are consistent with the results reported in the literature and this assay would be a powerful asset to forensic laboratories when trying to determine the age of an unknown subject.

D. Concluding Remarks

When examining the results, the lack of coverage stands out as the single greatest deficiency of the assay. In this approach over 10 different methylation assays reported in

the literature were amplified and sequenced in a single-tube format and methylation data was recovered for each of the assays. The body fluid identification assay showed results for saliva, blood, and vaginal epithelial samples that are generally consistent with the published literature, The age prediction assay for blood samples utilizing two CpG sites in the ELOVL2 marker provided early indications that this methodology has the potential to work properly if the methylation data for each CpG in an assay is properly represented. However, due to the large inconsistences with read coverage of CpGs within each assay, as well as the generally low coverage, it would not be possible to implement the assay as it currently exists for use in a forensic setting. The low coverage across the panel is likely due to inefficient amplification of the target regions, which then causes insufficient library to be loaded on to the flow cell for sequencing.

The other age predictions produced results outside of the published range for the models. Primer design should be reevaluated to increase the amplification efficiency of the various targeted regions. This can be in the form of improved primers for the assay, as well as varying the concentrations of the primers to allow for a more equal representation of each target. Additionally, the number of PCR cycles should also be explored, particularly during Target Enrichment. The protocol provided by Qiagen dictates that only 8 cycles of Target Enrichment PCR are necessary for this library preparation process, and then universal PCR can include anywhere from 19 to 26 cycles depending on the amount of input DNA. A larger number of cycles during Target Enrichment PCR could help to ensure that each target region is amplified sufficiently and that each unique molecule of DNA in the sample is amplified.

Ultimately, if sensitivity and read depth can be improved, the use of a single reaction for assessing all the assays contained in a large panel would allow for more reproducible data to be obtained. With additional resources and time, this methodology has the potential to dramatically improve the ability of forensic DNA laboratories to determine body fluid origin when the presence of DNA is not in dispute and to provide age as an additional descriptor of unknown individuals being sought in connection with a crime.

CHAPTER VIII – CONCLUDING REMARKS AND FUTURE DIRECTIONS

In the past decade, the goal of forensic DNA researchers has switched from the passive goal of providing a profile for reference to a database to a more proactive goal of providing more information to investigators when no matches exist in the database. One way to do this is to probe epigenetic modifications for differential gene expression. In addition to body fluid type, phenotypic characteristics, such as biological age, lifestyle traits, can be examined.

For DNA methylation analysis to work in a forensic laboratory, the ability to accurately differentiate between a methylated cytosine and unmethylated cytosine needs to be accomplished with methodologies and equipment that are readily available and familiar to forensic analysts. For this reason, DNA methylation assays should be preferable to mRNA or protein analysis methods due to the fact that the same DNA extract used for genotyping can be used for methylation analysis.

With DNA methylation analysis well established in the literature, there has been a push to make this methodology more accessible and implementable to forensic laboratories. Ultimately it will be important to push forward the legal process of getting these technologies accepted in a court of law. In this work we provide evidence that DNA methylation markers for body fluid identification and age determination are best analyzed in a single tube reaction format in order to dramatically decrease the volume of sample needed for analysis while maximizing the information that can be determined. In this thesis the creation, validation, and objective determination of body fluid origin by multiplex amplification has been performed. In addition, the preliminary construction of a next

generation sequencing assay to determine body fluid origin and age determination simultaneously.

A multiplex amplification and pyrosequencing assay was developed using four different body fluid identification markers to determine the body fluid origin of a DNA sample. Specifically, BCAS4 for saliva, cg06379435 for blood, VE_8 for vaginal epithelia, and ZC3H12D for semen were developed to be analyzed as a group to increase the accuracy of the assay. The construction of this multiplex, starting with four monoplex reactions, required the careful balancing of primer concentration ratios, the exploration of new PCR and sequencing primers for greater peak heights, and the use of formamide to increase stringency. The result was a multiplex assay that reduced the number of PCR reactions, and therefore sample volume, required to determine body fluid origin for the four most commonly found body fluids at crime scenes. Each of the four markers was previously described in the literature as being specific for one body fluid, and the resulting multiplex demonstrated methylation values for each body fluid that was consistent with the literature.

To increase the viability of a body fluid multiplex for a forensic laboratory, a developmental validation study of the multiplex was conducted. This validation study included population, sensitivity, inhibition, degradation, and mixture studies. The results of the population study demonstrated the stability of the assay across a wide range of individuals proving that the results were reproducible for body fluid identification. The sensitivity, inhibition, and degradation studies provided results that were consistent with prior works showing that accurate methylation analysis can be achieved with nanogram to subnanogram DNA concentrations and that the assay can be used with degraded and inhibited samples. The mixture study provided a significant update from the mixture

studies of the monoplex reactions. Because the results of all four body fluid markers are from the same amplicon, it is possible to combine the results when determining the presence of a mixture. Although body fluid mixtures still present as methylation profiles that are the intermediate of two separate body fluids, it is possible to exclude the presence of a body fluid, and the presence of a mixture can be confirmed due to the intermediate methylation value observed in the body fluid marker of those two body fluids.

Results were analyzed using, both cluster analysis and latent profile analysis to objectively identify the body fluid origin of an unknown sample. These two methodologies combine the information from all four body fluid identification markers in the multiplex to provide a single result.

Finally, the preliminary results of a targeted methyl sequencing assay for body fluid identification and age determination was presented. While the preliminary data shows that there was not sufficiently high enough coverage, the results demonstrate the potential path forward.

Future work should involve the inclusion of more body fluid loci and continued development of the targeted methyl sequencing assay. For body fluid identification, additional markers that present the opposite methylation profiles would help to increase the accuracy of the assay and could help in determining mixture ratios. For example, the inclusion of a semen identifying marker that is hypermethylated in semen but hypomethylated in other body fluids would complement ZC3H12D by giving more evidence of body fluid mixtures and the additional data could aid in calculating mixture ratios. The inclusion of markers for more body fluids, such as menstrual blood, sweat, and nasal mucosa, would also benefit for the forensic community greatly. However, as

previously discussed, the multiplex assay for pyrosequencing faces a limit for the number of markers that can be reasonably detected during sequencing. Therefore, an additional age multiplex may need to be developed or the jump to massively parallel sequencing will need to be implemented. Additionally, the development of an age prediction model for vaginal epithelia is currently lacking in the literature and could be quite beneficial in certain circumstances. The findings of this study will certainly benefit from further optimizations to increase accuracy of age determination and the number of body fluids being targeted.

The study of DNA methylation and its effects on gene expression in mammals has been known for several decades at this point, and yet its implementation to the forensic field is only ten years old. While this type of research still in its infancy, there are an unknown number of advancements that are yet to be discovered. It is my hope that the results presented in this work act as an important body of knowledge for the current state of DNA methylation analysis, and its implementation in the forensic laboratory.

LIST OF REFERENCES

1. Ulhenhuth P. Das biologische Verfahren zur Erkennung und Unterscheidung von Menschen- und Tierblut sowie anderer Eiweisssubstanzen und seine Anwendung in der forensicher Praxis: ausgewählte Sammlung von Arbeiten und Gutachten. *Dtsch Medizinische Wochenschrift*. Published online 1905:152. Accessed August 14, 2020.

2. Butler J. *Advanced Topics in Forensic DNA Typing: Methodology*. Elsevier Inc.; 2011. Accessed August 14, 2020.

3. Waddington CH. The epigenotype. *Endeavour*. 1942;1:18-20. doi:10.1093/ije/dyr184

4. Han Y, Garcia BA. Combining genomic and proteomic approaches for epigenetics research. *Epigenomics*. 2013;5(4):439-452. doi:10.2217/epi.13.37

5. Goldberg AD, Allis CD, Bernstein E. Epigenetics: A Landscape Takes Shape. *Cell*. 2007;128(4):635-638. doi:10.1016/j.cell.2007.02.006

6. Vidaki A, Daniel B, Court DS. Forensic DNA methylation profiling - Potential opportunities and challenges. *Forensic Sci Int Genet*. 2013;7(5):499-507. doi:10.1016/j.fsigen.2013.05.004

7. Franklin RE, Gosling RG. Molecular configuration in sodium thymonucleate. *Nature*. 1953;171(4356):740-741. doi:10.1038/171740a0

8. Michelson AM, Todd AR. Nucleotides part XXXII. Synthesis of a dithymidine dinucleotide containing a 3′: 5′-internucleotidic linkage. *J Chem Soc*. 1955;(0):2632-2638. doi:10.1039/JR9550002632

9. Crick F, Watson J. The complementary structure of deoxyribonucleic acid. *Proc R Soc London Ser A Math Phys Sci*. 1954;223(1152):80-96. doi:10.1098/rspa.1954.0101

10. Molnar C, Gair J. 9.1 The Structure of DNA. In: *Concepts of Biology - 1st Canadian Edition*. 1st ed. BCcampus; 2012.

11. Alberts B, Johnson A, Lewis J, Raff M, Roberts K, Walter P. *Molecular Biology of the Cell - NCBI Bookshelf*. 4th ed. Garland Science; 2002. Accessed August 14, 2020. https://www.ncbi.nlm.nih.gov/books/NBK21054/

12. Mandel M, Marmur J. [109] Use of ultraviolet absorbance-temperature profile for determining the guanine plus cytosine content of DNA. *Methods Enzymol*. 1968;12(PART B):195-206. doi:10.1016/0076-6879(67)12133-2

13. Fowler S, Roush R, Wise J. Concepts of Biology: OpenStax. Published online 2014. Accessed August 14, 2020. https://openstax.org/details/books/concepts-biology

14. Crick F. The biological replication of macromolecules. *Symp Soc Exp Biol*. Published online 1958.

15. Voet D, Voet JG. *Biochemistry, 4th Edition | Wiley*. John Wiley & Sons, Inc.; 2010.

16. Watson J, Baker T, Bell S, Gann A, Levine M, Losick R. *Molecular Biology of the Gene*. 7th ed. USA: Pearson Education, Inc.; 2017. Accessed August 15, 2020.

17. Van Wazer JR, Griffith EJ, McCullough JF. Structure and Properties of the Condensed Phosphates. VII. Hydrolytic Degradation of Pyro- and Tripolyphosphate. *J Am Chem Soc*. 1955;77(2):287-291. doi:10.1021/ja01607a011

18. Richardson L V., Richardson JP. Rho-dependent termination of transcription is governed primarily by the upstream Rho utilization (rut) sequences of a terminator. *J Biol Chem*. 1996;271(35):21597-21603. doi:10.1074/jbc.271.35.21597

19. Watson J, Baker T, Bell S, Gann A, Levine M, Losick R. *Molecular Biology of the Gene*. 7th ed. USA: Pearson Education, Inc.; 2017. Accessed August 15, 2020. https://www.slugbooks.com/9780805395921-molecular-biology-of-the-gene-6th.html

20. Berg J, Tymoczko J, Stryer L. *Biochemistry*. 5th ed. W H Freeman; 2002. Accessed August 15, 2020. https://www.ncbi.nlm.nih.gov/books/NBK21154/

21. Ramasubbu N, Paloth V, Luo Y, Brayer GD, Levine MJ. Structure of human salivary α-amylase at 1.6 Å resolution: Implications for its role in the oral cavity. *Acta Crystallogr Sect D Biol Crystallogr*. 1996;52(3):435-446. doi:10.1107/S0907444995014119

22. Allis CD, Jenuwein T. The molecular hallmarks of epigenetic control. *Nat Rev Genet*. 2016;17(8):487-500. doi:10.1038/nrg.2016.59

23. Pollard T, Earnshaw W, Lippincott Schwartz J. *Cell Biology* . 2nd ed. Saunders/Elsevier; 2008. Accessed August 16, 2020.

24. Watson J, Baker T, Bell S, Gann A, Levine M, Losick R. *Molecular Biology of the Gene*. 7th ed. Pearson Education, Inc.; 2017.

25. Mazzio EA, Soliman KFA. Basic concepts of epigenetics impact of environmental signals on gene expression. *Epigenetics*. 2012;7(2):119-130.

26. Rouault TA. The role of iron regulatory proteins in mammalian iron homeostasis and disease. *Nat Chem Biol*. 2006;2(8):406-414. doi:10.1038/nchembio807

27. Rakyan VK, Down TA, Balding DJ, Beck S. Epigenome-wide association studies for common human diseases. *Nat Rev Genet*. 2011;12(8):529-541. doi:10.1038/nrg3000

28. Zhu H, Wang G, Qian J. Transcription factors as readers and effectors of DNA methylation. *Nat Rev Genet*. 2016;17(9):551-565. doi:10.1038/nrg.2016.83

29. Han L, Su B, Li WH, Zhao Z. CpG island density and its correlations with genomic features in mammalian genomes. *Genome Biol*. 2008;9(5). doi:10.1186/gb-2008-9-5-r79

30. Tammen SA, Friso S, Choi SW. Epigenetics: The link between nature and nurture. *Mol Aspects Med*. 2013;34(4):753-764. doi:10.1016/j.mam.2012.07.018

31. Weber M, Schübeler D. Genomic patterns of DNA methylation: targets and function of an epigenetic mark. *Curr Opin Cell Biol*. 2007;19(3):273-280. doi:10.1016/j.ceb.2007.04.011

32. Genereux DP, Johnson WC, Burden AF, Stöger R, Laird CD. Errors in the bisulfite conversion of DNA: Modulating inappropriate- and failed-conversion frequencies. *Nucleic Acids Res*. 2008;36(22):e150. doi:10.1093/nar/gkn691

33. Kader F, Ghai M. DNA methylation and application in forensic sciences. *Forensic Sci Int*. 2015;249:255-265. doi:10.1016/j.forsciint.2015.01.037

34. 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. *Biochim Biophys Acta - Rev Cancer*. 2013;1835(2):155-163. doi:10.1016/j.bbcan.2012.12.003

35. Aguilera O, Fernández AF, Muñoz A, Fraga MF. Epigenetics and environment: A complex relationship. *J Appl Physiol*. 2010;109(1):243-251. doi:10.1152/japplphysiol.00068.2010

36. Malygin EG, Hattman S. DNA methyltransferases: Mechanistic models derived from kinetic analysis. *Crit Rev Biochem Mol Biol*. 2012;47(2):97-193. doi:10.3109/10409238.2011.620942

37. Chédin F. The DNMT3 family of mammalian de novo DNA methyltransferases. In: *Progress in Molecular Biology and Translational Science*. Vol 101. Elsevier B.V.; 2011:255-285. doi:10.1016/B978-0-12-387685-0.00007-X

38. Cheng X, Hashimoto H, Horton JR, Zhang X. Mechanisms of DNA methylation, methyl-CpG recognition, and demethylation in mammals. In: *Handbook of Epigenetics*. Elsevier Inc.; 2011:9-24. doi:10.1016/B978-0-12-375709-8.00002-2

39. Rivera RM, Ross JW. Epigenetics in fertilization and preimplantation embryo development. *Prog Biophys Mol Biol*. 2013;113(3):423-432. doi:10.1016/j.pbiomolbio.2013.02.001

40. Schübeler D. Function and information content of DNA methylation. *Nature*. 2015;517(7534):321-326. doi:10.1038/nature14192

41. Szyf M. The elusive role of 5′-hydroxymethylcytosine. *Epigenomics*. 2016;8(11):1539-1551. doi:10.2217/epi-2016-0076

42. Straussman R, Nejman D, Roberts D, et al. Developmental programming of CpG island methylation profiles in the human genome. *Nat Struct Mol Biol*. 2009;16(5):564-571. doi:10.1038/nsmb.1594

43. Cedar H, Bergman Y. Programming of DNA Methylation Patterns. *Annu Rev Biochem*. 2012;81(1):97-117. doi:10.1146/annurev-biochem-052610-091920

44. Jaenisch R, Bird A. Epigenetic regulation of gene expression: How the genome integrates intrinsic and environmental signals. *Nat Genet*. 2003;33(3S):245-254. doi:10.1038/ng1089

45. Nikolova YS, Hariri AR. Can we observe epigenetic effects on human brain function? *Trends Cogn Sci*. 2015;19(7):366-373. doi:10.1016/j.tics.2015.05.003

46. Frigola J, Song J, Stirzaker C, Hinshelwood RA, Peinado MA, Clark SJ. Epigenetic remodeling in colorectal cancer results in coordinate gene suppression across an entire chromosome band. *Nat Genet*. 2006;38(5):540-549. doi:10.1038/ng1781

47. Gu J, Stevens M, Xing X, et al. Mapping of variable DNA methylation across multiple cell types defines a dynamic regulatory landscape of the human genome. *G3 Genes, Genomes, Genet*. 2016;6(4):973-986. doi:10.1534/g3.115.025437

48. Rakyan VK, Hildmann T, Novik KL, et al. DNA methylation profiling of the human major histocompatibility complex: A pilot study for the Human Epigenome Project. *PLoS Biol*. 2004;2(12). doi:10.1371/journal.pbio.0020405

49. 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. doi:10.1093/ilar.53.3-4.289

50. Hou L, Zhang X, Wang D, Baccarelli A. Environmental chemical exposures and human epigenetics. *Int J Epidemiol*. 2012;41(1):79-105. doi:10.1093/ije/dyr154

51. Gluckman PD, Hanson MA, Pinal C. The developmental origins of adult disease. In: *Maternal and Child Nutrition*. Vol 1. Matern Child Nutr; 2005:130-141. doi:10.1111/j.1740-8709.2005.00020.x

52. Ehrlich M, Lacey M. DNA Hypomethylation and Hemimethylation in Cancer. In: *Advances in Experimental Medicine and Biology*. Vol 754. Adv Exp Med Biol; 2013:31-56. doi:10.1007/978-1-4419-9967-2_2

53. Fraga MF, Ballestar E, Paz MF, et al. Epigenetic differences arise during the lifetime of monozygotic twins. *Proc Natl Acad Sci U S A*. 2005;102(30):10604-10609. doi:10.1073/pnas.0500398102

54. Garrett-Bakelman FE, Darshi M, Green SJ, et al. The NASA twins study: A multidimensional analysis of a year-long human spaceflight. *Science (80-)*. 2019;364(6436). doi:10.1126/science.aau8650

55. Teschendorff AE, Menon U, Gentry-Maharaj A, et al. Age-dependent DNA methylation of genes that are suppressed in stem cells is a hallmark of cancer. *Genome Res*. 2010;20(4):440-446. doi:10.1101/gr.103606.109

56. Casillas MA, Lopatina N, Andrews LG, Tollefsbol TO. Transcriptional control of the DNA methyltransferases is altered in aging and neoplastically-transformed human fibroblasts. *Mol Cell Biochem*. 2003;252(1-2):33-43. doi:10.1023/A:1025548623524

57. Jung M, Pfeifer GP. Aging and DNA methylation. *BMC Biol*. 2015;13(1):1-8. doi:10.1186/s12915-015-0118-4

58. Lillycrop KA, Hoile SP, Grenfell L, Burdge GC. DNA methylation, ageing and the influence of early life nutrition. In: *Proceedings of the Nutrition Society*. Vol 73. Cambridge University Press; 2014:413-421. doi:10.1017/S0029665114000081

59. Vidaki A, Ballard D, Aliferi A, Miller TH, Barron LP, Syndercombe Court D. DNA methylation-based forensic age prediction using artificial neural networks and next generation sequencing. *Forensic Sci Int Genet*. 2017;28:225-236. doi:10.1016/j.fsigen.2017.02.009

60. Alghanim H, Antunes J, Silva DSBS, Alho CS, Balamurugan K, McCord B. Detection and evaluation of DNA methylation markers found at SCGN and KLF14 loci to estimate human age. *Forensic Sci Int Genet*. 2017;31:81-88. doi:10.1016/j.fsigen.2017.07.011

61. Silva DSBS, Antunes J, Balamurugan K, Duncan G, Alho CS, Mccord B. Evaluation of DNA methylation markers and their potential to predict human aging. *Electrophoresis*. 2015;36(15):1775-1780. doi:10.1002/elps.201500137

62. Boyd-Kirkup JD, Green CD, Wu G, Wang D, Han JDJ. Epigenomics and the regulation of aging. *Epigenomics*. 2013;5(2):205-227. doi:10.2217/epi.13.5

63. 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):1736-1745. doi:10.1002/elps.201100711

64. Lee HY, Jung SE, Lee EH, Yang WI, Shin KJ. DNA methylation profiling for a confirmatory test for blood, saliva, semen, vaginal fluid and menstrual blood. *Forensic Sci Int Genet*. 2016;24:75-82. doi:10.1016/j.fsigen.2016.06.007

65. McCord B, Gauthier Q, Alghanim H, et al. Applications of epigenetic methylation in body fluid identification, age determination and phenotyping. *Forensic Sci Int Genet Suppl Ser*. 2019;7(1):485-487. doi:10.1016/j.fsigss.2019.10.061

66. Illumina. ForenSeq™ DNA Signature Prep Reference Guide. 2015;(September):36. doi:# TG-450-9001DOC Material # 20000923 Document # 15049528 v01

67. 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(1):263-273. doi:10.1101/sqb.1986.051.01.032

68. Butler J. *Fundamentals of Forensic DNA Typing*. 1st ed. Elsevier, Inc.; 2009. Accessed August 18, 2020. https://www.elsevier.com/books/fundamentals-of-forensic-dna-typing/butler/978-0-12-374999-4

69. Benecke M. *Forensic DNA Samples—Collection and Handling*. Taylor & Francis; 2004.

70. Smith LM, Burgoyne LA. Collecting, archiving and processing DNA from wildlife samples using FTA® databasing paper. *BMC Ecol*. 2004;4:4. doi:10.1186/1472-6785-4-4

71. Scientific Working Group on DNA Analysis Methods (SWGDAM). Published 2015. www.swgdam.org

72. Ballantyne J. Serology: overview. In: *Encyclopedia of Forensic Sciences*. ; 2000:1322. Accessed August 19, 2020.

73. Hanson E, Lubenow H, Ballantyne J. Identification of forensically relevant body fluids using a panel of differentially expressed microRNAs. *Forensic Sci Int Genet Suppl Ser*. 2009;2(1):503-504. doi:10.1016/j.fsigss.2009.08.184

74. Old JB, Schweers BA, Boonlayangoor PW, Reich KA. Developmental Validation of RSID^{TM}-Saliva: A Lateral Flow Immunochromatographic Strip Test for the Forensic Detection of Saliva. *J Forensic Sci*. 2009;54(4):866-873. doi:10.1111/j.1556-4029.2009.01055.x

75. Pang BCM, Cheung BKK. Applicability of Two Commercially Available Kits for Forensic Identification of Saliva Stains. *J Forensic Sci*. 2008;53(5):1117-1122. doi:10.1111/j.1556-4029.2008.00814.x

76. Gaensslen R. *Sourcebook in Forensic Serology, Immunology, and Biochemistry*. US Department of Justice, National Institute of Justice; 1983. Accessed August 19, 2020. https://www.ncjrs.gov/app/abstractdb/AbstractDBDetails.aspx?id=91728

77. Kulstein G, Wiegand P. Comprehensive examination of conventional and innovative body fluid identification approaches and DNA profiling of laundered blood- and saliva-stained pieces of cloths. *Int J Legal Med*. 2018;132(1):67-81. doi:10.1007/s00414-017-1691-6

78. RSIDTM-Blood. http://www.ifi-test.com/rsidtm-blood/.

79. Schiff AF. Reliability of the Acid Phosphatase Test for the Identification of Seminal Fluid. *J Forensic Sci*. 1978;23(4):10745J. doi:10.1520/jfs10745j

80. Hochmeister MN, Budowle B, Rudin O, et al. Evaluation of Prostate-Specific Antigen (PSA) Membrane Test Assays for the Forensic Identification of Seminal Fluid. *J Forensic Sci*. 1999;44(5):12042J. doi:10.1520/jfs12042j

81. Allery J-P, Telmon N, Mieusset R, Blanc A, Rougé D. Cytological Detection of Spermatozoa: Comparison of Three Staining Methods. *J Forensic Sci*. 2001;46(2):14970J. doi:10.1520/jfs14970j

82. RSIDTM-Semen. http://www.ifi-test.com/rsidtm-blood/

83. Old J, Schweers BA, Boonlayangoor PW, Fischer B, Miller KWP, Reich K. Developmental Validation of RSID^{TM}-Semen: A Lateral Flow Immunochromatographic Strip Test for the Forensic Detection of Human Semen. *J Forensic Sci*. 2012;57(2):489-499. doi:10.1111/j.1556-4029.2011.01968.x

84. Randall B. Glycogenated Squamous Epithelial Cells as a Marker of Foreign Body Penetration in Sexual Assault. *J Forensic Sci*. 1988;33(2):11965J. doi:10.1520/jfs11965j

85. Divall GB, Ismail M. Lactate dehydrogenase isozymes in vaginal swab extracts: A problem for the identification of menstrual blood. *Forensic Sci Int*. 1983;21(2):139-147. doi:10.1016/0379-0738(83)90102-0

86. Kobilinsky L. Recovery and Stability of DNA in Samples of Forensic Science Significance. *Forensic Sci Rev*. 1992;4(1):67-87.

87. Lee SB, Shewale JG. DNA Extraction Methods in Forensic Analysis. In: *Encyclopedia of Analytical Chemistry*. John Wiley & Sons, Ltd; 2017:1-18. doi:10.1002/9780470027318.a1104m.pub2

88. Marmur J. [100] A procedure for the isolation of deoxyribonucleic acid from microorganisms. *Methods Enzymol*. 1963;6(C):726-738. doi:10.1016/0076-6879(63)06240-6

89. Wu J, Wang H, Zhu A, Long F. Adsorption Kinetics of Single-Stranded DNA on Functional Silica Surfaces and Its Influence Factors: An Evanescent-Wave Biosensor Study. *ACS Omega*. 2018;3(5):5605-5614. doi:10.1021/acsomega.7b02063

90. Raimondo TM, McCalla SE. Adsorption and desorption of DNA-functionalized beads in glass microfluidic channels. *Biomicrofluidics*. 2019;13(5). doi:10.1063/1.5115160

91. Bellete B, Flori P, Hafid J, Raberin H, Tran Manh Sung R. Influence of the quantity of nonspecific DNA and repeated freezing and thawing of samples on the quantification of DNA by the Light Cycler®. *J Microbiol Methods*. 2003;55(1):213-219. doi:10.1016/S0167-7012(03)00141-6

92. Plexor ® HY System. Published 2015. https://www.promega.com/products/forensic-dna-analysis-ce/human-specific-dna-quantitation/plexor-hy-system/?catNum=DC1001#protocols

93. Quantifiler™ HP and Trio DNA Quantification Kits User Guide. Published 2017. Accessed August 18, 2020. http://tools.thermofisher.com/content/sfs/manuals/4485354.pdf

94. Thompson RE, Duncan G, Mccord BR. An investigation of PCR inhibition using plexor®-based quantitative PCR and short tandem repeat amplification. *J Forensic Sci*. 2014;59(6):1517-1529. doi:10.1111/1556-4029.12556

95. 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):2002414. doi:10.1520/jfs2002414

96. Mandrekar MN, Erickson AM, Kopp K, et al. *Development of a Human DNA Quantitation System*. Vol 42.; 2001. Accessed August 18, 2020. www.cmj.hr

97. Product Sheet: GlobalFiler PCR Amplification Kit - 1,000 Reactions. Thermo Fisher Scientific. Published 2018. Accessed August 18, 2020.

98. 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;264(11):6427-6437. Accessed August 18, 2020. https://europepmc.org/article/med/2649500

99. Landgraf A, Wolfes H. Taq polymerase (EC 2.7.7.7): with particular emphasis on its use in PCR protocols. *Methods Mol Biol*. 1993;16:31-58. doi:10.1385/0-89603-234-5:31

100. Markoulatos P, Siafakas N, Moncany M. Multiplex polymerase chain reaction: A practical approach. *J Clin Lab Anal*. 2002;16(1):47-51. doi:10.1002/jcla.2058

101. Louwrier A, Van Der Valk A. Thermally reversible inactivation of Taq polymerase in an organic solvent for application in hot start PCR. *Enzyme Microb Technol*. 2005;36(7):947-952. doi:10.1016/j.enzmictec.2005.01.019

102. Brunstein J. PCR: the basics of the polymerase chain reaction. *Med Lab Obs*. 2013;45(4):32-35.

103. Blake R, Delcourt S. Thermodynamic effects of formamide on DNA stability. *Nucleic Acids Res*. 1996;24(11):2095-2103.

104. 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):1397-1412. doi:10.1002/elps.200305822

105. Wenz HM, Robertson JM, Menchen S, et al. High-precision genotyping by denaturing capillary electrophoresis. *Genome Res*. 1998;8(1):69-80. doi:10.1101/gr.8.1.69

106. Budowle B, Koons BW, Keys KM, Smerick JB. Methods for Typing the STR Triplex CSF1PO, TPOX, and HUMTHO1 That Enable Compatibility Among DNA Typing Laboratories. In: Springer, Berlin, Heidelberg; 1996:107-114. doi:10.1007/978-3-642-80029-0_27

107. Haas J, Katus HA, Meder B. Next-generation sequencing entering the clinical arena. *Mol Cell Probes*. 2011;25(5-6):206-211. doi:10.1016/j.mcp.2011.08.005

108. Ropers HH. On the future of genetic risk assessment. *J Community Genet*. 2012;3(3):229-236. doi:10.1007/s12687-012-0092-2

109. Gargis AS, Kalman L, Berry MW, et al. Assuring the quality of next-generation sequencing in clinical laboratory practice. *Nat Biotechnol*. 2012;30(11):1033-1036. doi:10.1038/nbt.2403

110. Illumina. Illumina Announces MiSeq(TM) Personal Sequencing System. Published 2011. Accessed August 18, 2020. https://emea.illumina.com/company/news-center/press-releases/2011/1515239.html

111. Illumina. System Specification Sheet: Forensic Genomics. www.illumina.com/systems/miseq-fgx.ilmn%0Ahttp://www.illumina.com/content/dam/illumina-marketing/documents/products/datasheets/miseq-fgx-system-spec-sheet-1470-2014-004.pdf

112. García-García G, Baux D, Faugère V, et al. Assessment of the latest NGS enrichment capture methods in clinical context. *Sci Rep*. 2016;6(1):1-8. doi:10.1038/srep20948

113. Hussing C, Kampmann ML, Mogensen HS, Børsting C, Morling N. Comparison of techniques for quantification of next-generation sequencing libraries. *Forensic Sci Int Genet Suppl Ser*. 2015;5:e276-e278. doi:10.1016/j.fsigss.2015.09.110

114. Broad Institute, Inc. I. Cluster Generation Module 2: Overview. Accessed August 18, 2020. https://www.broadinstitute.org/files/shared/illuminavids/clusterGenSlides.pdf

115. Voelkerding K V., Dames SA, Durtschi JD. Next-generation sequencing:from basic research to diagnostics. *Clin Chem*. 2009;55(4):641-658. doi:10.1373/clinchem.2008.112789

116. Campbell PJ, Stephens PJ, Pleasance ED, et al. Identification of somatically acquired rearrangements in cancer using genome-wide massively parallel paired-end sequencing. *Nat Genet*. 2008;40(6):722-729. doi:10.1038/ng.128

117. Adey A, Shendure J. Ultra-low-input, tagmentation-based whole-genome bisulfite sequencing. *Genome Res*. 2012;22(6):1139-1143. doi:10.1101/gr.136242.111

118. Buck-Koehntop BA, Defossez P-A. On how mammalian transcription factors recognize methylated DNA. *Epigenetics*. 2013;8(2):131-137. doi:10.4161/epi.23632

119. Shapiro R, Servis RE, Welcher M. Reactions of Uracil and Cytosine Derivatives with Sodium Bisulfite. A Specific Deamination Method. *J Am Chem Soc*. 1970;92(2):422-424. doi:10.1021/ja00705a626

120. Frommer M, McDonald LE, Millar DS, et al. A genomic sequencing protocol that yields a positive display of 5- methylcytosine residues in individual DNA strands. *Proc Natl Acad Sci U S A*. 1992;89(5):1827-1831. doi:10.1073/pnas.89.5.1827

121. Kristensen LS, Treppendahl MB, Grønbæk K. Analysis of Epigenetic Modifications of DNA in Human Cells. *Curr Protoc Hum Genet*. 2013;77(1):20.2.1-20.2.22. doi:10.1002/0471142905.hg2002s77

122. 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). doi:10.1093/nar/gkm013

123. Antunes J, Silva DSBS, Balamurugan K, Duncan G, Alho CS, Mccord B. High-resolution melt analysis of DNA methylation to discriminate semen in biological stains. *Anal Biochem*. 2016;494:40-45. doi:10.1016/j.ab.2015.10.002

124. Erali M, Voelkerding K V., Wittwer CT. High resolution melting applications for clinical laboratory medicine. *Exp Mol Pathol*. 2008;85(1):50-58. doi:10.1016/j.yexmp.2008.03.012

125. Hanson EK, Ballantyne J. Rapid and inexpensive body fluid identification by RNA profiling-based multiplex High Resolution Melt (HRM) analysis. *F1000Research*. 2013;2(May):281. doi:10.12688/f1000research.2-281.v1

126. Wojdacz TK. Methylation-sensitive high-resolution melting in the context of legislative requirements for validation of analytical procedures for diagnostic applications. *Expert Rev Mol Diagn*. 2012;12(1):39-47. doi:10.1586/erm.11.88

127. Seipp MT, Durtschi JD, Voelkerding K V., Wittwer CT. Multiplex amplicon genotyping by high-resolution melting. *J Biomol Tech*. 2009;20(3):160-164. Accessed August 20, 2020. /pmc/articles/PMC2700465/?report=abstract

128. Gonzalgo ML, Jones PA. Rapid quantitation of methylation differences at specific sites using methylation-sensitive single nucleotide primer extension (Ms-SNuPE). *Nucleic Acids Res*. 1997;25(12):2529-2531. doi:10.1093/nar/25.12.2529

129. Wong IHN. Qualitative and quantitative polymerase chain reaction-based methods for DNA methylation analyses. *Methods Mol Biol*. 2006;336:33-43. doi:10.1385/1-59745-074-x:33

130. Salas A, Quintáns B, Alvarez-Iglesias V. SNaPshot typing of mitochondrial DNA coding region variants. *Methods Mol Biol*. 2005;297:197-208. doi:10.1385/1-59259-867-6:197

131. Bertoncini S, Blanco-Rojo R, Baeza C, Arroyo-Pardo E, Vaquero MP, López-Parra AM. A novel snapshot assay to detect genetic mutations related to iron metabolism. *Genet Test Mol Biomarkers*. 2011;15(3):173-179. doi:10.1089/gtmb.2010.0140

132. Tost J, Gut IG. DNA Methylation Analysis by MALDI Mass Spectrometry. In: *Encyclopedia of Molecular Cell Biology and Molecular Medicine*. Wiley-VCH Verlag GmbH & Co. KGaA; 2012. doi:10.1002/3527600906.mcb.201100025

133. Suchiman HED, Slieker RC, Kremer D, Slagboom PE, Heijmans BT, Tobi EW. Design, measurement and processing of region-specific DNA methylation assays: The mass spectrometry-based method EpiTYPER. *Front Genet*. 2015;6(SEP):287. doi:10.3389/fgene.2015.00287

134. Kunze S. Quantitative region-specific dna methylation analysis by the EpiTYPER™ technology. In: *Methods in Molecular Biology*. Vol 1708. Humana Press Inc.; 2018:515-535. doi:10.1007/978-1-4939-7481-8_26

135. Nyrén P. Enzymatic method for continuous monitoring of DNA polymerase activity. *Anal Biochem*. 1987;167(2):235-238. doi:10.1016/0003-2697(87)90158-8

136. Hyman ED. A new method of sequencing DNA. *Anal Biochem*. 1988;174(2):423-436. doi:10.1016/0003-2697(88)90041-3

137. Diggle MA, Clarke SC. A Novel Method for Preparing Single-Stranded DNA for Pyrosequencing™. *Mol Biotechnol*. 2003;24.

138. Diggle MA, Clarke SC. *Pyrosequencing™ Pyrosequencing™ Sequence Typing at the Speed of Light*. Vol 28.; 2004.

139. Gharizadeh B, GhaderiHADERI M, Nyrén P. Pyrosequencing Technology for Short DNA Sequencing and Whole Genome Sequencing. *Seibutsu Butsuri*. 2007;47(2):129-132. doi:10.2142/biophys.47.129

140. Kuwahara M, Hagiwara K, Ozaki H. Polymerase Reactions that Involve Modified Nucleotides. In: Springer, Cham; 2016:429-453. doi:10.1007/978-3-319-34175-0_18

141. Qiagen. Pyromark Q48 Autoprep Product Profile.

142. Li C, Zhao S, Zhang N, Zhang S, Hou Y. Differences of DNA methylation profiles between monozygotic twins' blood samples. *Mol Biol Rep*. 2013;40(9):5275-5280. doi:10.1007/s11033-013-2627-y

143. Barros-Silva D, Marques CJ, Henrique R, Jerónimo C. Profiling DNA methylation based on next-generation sequencing approaches: New insights and clinical applications. *Genes (Basel)*. 2018;9(9). doi:10.3390/genes9090429

144. Richards R, Patel J, Stevenson K, Harbison S. Evaluation of massively parallel sequencing for forensic DNA methylation profiling. *Electrophoresis*. 2018;39(21):2798-2805. doi:10.1002/elps.201800086

145. Kint S, De Spiegelaere W, De Kesel J, Vandekerckhove L, Van Criekinge W. Evaluation of bisulfite kits for DNA methylation profiling in terms of DNA fragmentation and DNA recovery using digital PCR. Albertini E, ed. *PLoS One*. 2018;13(6):e0199091. doi:10.1371/journal.pone.0199091

146. Launen L. Illumina Sequencing (for Dummies) -An overview on how our samples are sequenced. – kscbioinformatics. kscbioinformatics. Published 2017. Accessed August 21, 2020. https://kscbioinformatics.wordpress.com/2017/02/13/illumina-sequencing-for-dummies-samples-are-sequenced/

147. QIAGEN. QIAseq ™ Targeted Methyl Panel Handbook. 2019;(October).

148. 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(1-3):1-17. doi:10.1016/j.forsciint.2009.02.013

149. Richard MLL, Harper KA, Craig RL, Onorato AJ, Robertson JM, Donfack J. Evaluation of mRNA marker specificity for the identification of five human body fluids by capillary electrophoresis. *Forensic Sci Int Genet*. 2012;6(4):452-460. doi:10.1016/j.fsigen.2011.09.007

150. 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. doi:10.1016/S0379-0738(03)00197-X

151. Bauer M. RNA in forensic science. *Forensic Sci Int Genet*. 2007;1(1):69-74. doi:10.1016/j.fsigen.2006.11.002

152. Bauer M, Patzelt D. Identification of menstrual blood by real time RT-PCR: Technical improvements and the practical value of negative test results. *Forensic Sci Int*. 2008;174(1):55-59. doi:10.1016/j.forsciint.2007.03.016

153. 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 Sci Int Genet*. 2009;3(2):80-88. doi:10.1016/j.fsigen.2008.11.003

154. Fleming RI, Harbison S. The development of a mRNA multiplex RT-PCR assay for the definitive identification of body fluids. *Forensic Sci Int Genet*. 2010;4(4):244-256. doi:10.1016/j.fsigen.2009.10.006

155. Karpetsky TP, Hieter PA, Frank JJ, Levy CC. Polyamines, ribonucleases, and the stability of RNA. *Mol Cell Biochem*. 1977;17(2):89-99. doi:10.1007/BF01743432

156. Thakur ML, Srivastava US, Majumdar PK, Ganguly PK, Radhakrishnamurty RK. m-RNA translatability in the liver, brain and kidney of rats: Effect of protein calorie malnutrition in early life. *Nutr Res*. 1987;7(3):307-318. doi:10.1016/S0271-5317(87)80020-9

157. Huang Z, Fasco MJ, Kaminsky LS. Optimization of DNase I removal of contaminating DNA from RNA for use in quantitative RNA-PCR. *Biotechniques*. 1996;20(6):1012-1020. doi:10.2144/96206st02

158. Bauer M, Patzelt D. A method for simultaneous RNA and DNA isolation from dried blood and semen stains. *Forensic Sci Int*. 2003;136(1-3):76-78. doi:10.1016/S0379-0738(03)00219-6

159. Freeman B, Smith N, Curtis C, Huckett L, Mill J, Craig IW. DNA from buccal swabs recruited by mail: Evaluation of storage effects on long-term stability and suitability for multiplex polymerase chain reaction genotyping. *Behav Genet*. 2003;33(1):67-72. doi:10.1023/A:1021055617738

160. Frumkin D, Wasserstrom A, Budowle B, Davidson A. DNA methylation-based forensic tissue identification. *Forensic Sci Int Genet*. 2011;5(5):517-524. doi:10.1016/j.fsigen.2010.12.001

161. Park JL, Kwon OH, Kim JH, et al. Identification of body fluid-specific DNA methylation markers for use in forensic science. *Forensic Sci Int Genet*. 2014;13:147-153. doi:10.1016/j.fsigen.2014.07.011

162. Alghanim H, Wu W, McCord B. DNA methylation assay based on pyrosequencing for determination of smoking status. *Electrophoresis*. 2018;39(21):2806-2814. doi:10.1002/elps.201800098

163. Zbieć-Piekarska R, Spólnicka M, Kupiec T, et al. Development of a forensically useful age prediction method based on DNA methylation analysis. *Forensic Sci Int Genet*. 2015;17:173-179. doi:10.1016/j.fsigen.2015.05.001

164. Zbieć-Piekarska R, Spólnicka M, Kupiec T, et al. Examination of DNA methylation status of the ELOVL2 marker may be useful for human age prediction in forensic science. *Forensic Sci Int Genet*. 2015;14:161-167. doi:10.1016/j.fsigen.2014.10.002

165. Hannum G, Guinney J, Zhao L, et al. Genome-wide Methylation Profiles Reveal Quantitative Views of Human Aging Rates. *Mol Cell*. 2013;49(2):359-367. doi:10.1016/j.molcel.2012.10.016

166. Lee HY, Lee SD, Shin KJ. Forensic DNA methylation profiling from evidence material for investigative leads. *BMB Rep*. 2016;49(7):359-369. doi:10.5483/BMBRep.2016.49.7.070

167. Verogen. ForenSeq DNA Signature Prep Kit. Published 2018. https://verogen.com/wp-content/uploads/2018/08/ForenSeq-prep-kit-data-sheet-VD2018002.pdf

168. Andreou I, Dugan B, Schaffer J, et al. Targeted Bisulfite Sequencing : An Efficient, Streamlined Approach to Maximize Throughput and Minimize Cost Performance. *GDNA Sample to Insight*.; 2019.

169. Slieker RC, Relton CL, Gaunt TR, Slagboom PE, Heijmans BT. Age-related DNA methylation changes are tissue-specific with ELOVL2 promoter methylation as exception. *Epigenetics and Chromatin*. 2018;11(1):25. doi:10.1186/s13072-018-0191-3

170. Bärlund M, Monni O, Weaver JD, et al. Cloning of BCAS3 (17q23) and BCAS4 (20q13) genes that undergo amplification, overexpression, and fusion in breast cancer. *Genes Chromosom Cancer*. 2002;35(4):311-317. doi:10.1002/gcc.10121

171. Eckhardt F, Lewin J, Cortese R, et al. DNA methylation profiling of human chromosomes 6, 20 and 22. *Nat Genet*. 2006;38(12):1378-1385. doi:10.1038/ng1909

172. Riffel AK, Schuenemann E, Vyhlidal CA. Regulation of the CYP3A4 and CYP3A7 promoters by members of the nuclear factor I transcription factor family. *Mol Pharmacol*. 2009;76(5):1104-1114. doi:10.1124/mol.109.055699

173. Silva DSBS, 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 Sci Int Genet*. 2016;23:55-63. doi:10.1016/j.fsigen.2016.01.017

174. Lee HY, Park MJ, Choi A, An JH, Yang WI, Shin KJ. Potential forensic application of DNA methylation profiling to body fluid identification. *Int J Legal Med*. 2012;126(1):55-62. doi:10.1007/s00414-011-0569-2

175. Antunes J, Silva DSBS, Balamurugan K, Duncan G, Alho CS, McCord B. Forensic discrimination of vaginal epithelia by DNA methylation analysis through pyrosequencing. *Electrophoresis*. 2016;37(21):2751-2758. doi:10.1002/elps.201600037

176. Behnen M, Murk K, Kursula P, et al. Testis-expressed profilins 3 and 4 show distinct functional characteristics and localize in the acroplaxome-manchette complex in spermatids. *BMC Cell Biol*. 2009;10. doi:10.1186/1471-2121-10-34

177. Liang J, Wang J, Azfer A, et al. A novel CCCH-zinc finger protein family regulates proinflammatory activation of macrophages. *J Biol Chem*. 2008;283(10):6337-6346. doi:10.1074/jbc.M707861200

178. Roy A, Kolattukudy PE. Monocyte chemotactic protein-induced protein (MCPIP) promotes inflammatory angiogenesis via sequential induction of oxidative stress, endoplasmic reticulum stress and autophagy. *Cell Signal*. 2012;24(11):2123-2131. doi:10.1016/j.cellsig.2012.07.014

179. Lin RJ, Chien HL, Lin SY, et al. MCPIP1 ribonuclease exhibits broad-spectrum antiviral effects through viral RNA binding and degradation. *Nucleic Acids Res*. 2013;41(5):3314-3326. doi:10.1093/nar/gkt019

180. Köchl S, Niederstätter H, Parson W. DNA extraction and quantitation of forensic samples using the phenol-chloroform method and real-time PCR. *Methods Mol Biol*. 2005;297:13-30. doi:10.1385/1-59259-867-6:013

181. Gauthier QT, Cho S, Carmel JH, McCord BR. Development of a body fluid identification multiplex via DNA methylation analysis. *Electrophoresis*. 2019;40(18-19):elps.201900118. doi:10.1002/elps.201900118

182. Chen JQ, Wu Y, Yang H, Bergelson J, Kreitman M, Tian D. Variation in the ratio of nucleotide substitution and indel rates across genomes in mammals and bacteria. *Mol Biol Evol*. 2009;26(7):1523-1531. doi:10.1093/molbev/msp063

183. 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):1631-1638. doi:10.1002/elps.201000694

184. Moskalev EA, Zavgorodnij MG, Majorova SP, et al. Correction of PCR-bias in quantitative DNA methylation studies by means of cubic polynomial regression. *Nucleic Acids Res*. 2011;39(11):e77-e77. doi:10.1093/nar/gkr213

185. Adey A, Shendure J. Ultra-low-input, tagmentation-based whole-genome bisulfite sequencing. *Genome Res*. 2012;22(6):1139-1143. doi:10.1101/gr.136242.111

186. Warnecke PM, Stirzaker C, Melki JR, Millar DS, Paul CL, Clark SJ. Detection and measurement of PCR bias in quantitative methylation analysis of bisulphite-treated DNA. *Nucleic Acids Res*. 1997;25(21):4422-4426. doi:10.1093/nar/25.21.4422

187. Benschop CCG, Van Der Beek CP, Meiland HC, Van Gorp AGM, Westen AA, Sijen T. Low template STR typing: Effect of replicate number and consensus method on genotyping reliability and DNA database search results. *Forensic Sci Int Genet*. 2011;5(4):316-328. doi:10.1016/j.fsigen.2010.06.006

188. Cornish R 2007. *Statistics: 3.1 Cluster Analysis*.; 2007.

189. Manly B, Navarro Alberto J. *Multivariate Statistical Methods: A Primer*. 4th ed. Chapman & Hall/CRC; 2016.

190. Everitt B, Landau S, Leese M, Stahl D. *Cluster Analysis*. 5th ed. Wiley; 2011. Accessed August 29, 2020. https://www.wiley.com/en-us/Cluster+Analysis%2C+5th+Edition-p-9780470749913

191. Ward JH. Hierarchical Grouping to Optimize an Objective Function. *J Am Stat Assoc*. 1963;58(301):236. doi:10.2307/2282967

192. Rencher AC. *Methods of Multivariate Analysis*. 2nd ed. John Wiley & Sons, Inc.; 2003. doi:10.1002/0471271357.fmatter

193. Burns R, Burns R. *Business Research Methods and Statistics Using SPSS*. 1st ed. SAGE Publications Ltd.; 2008. Accessed August 29, 2020.

194. Tipton E. Stratified Sampling Using Cluster Analysis: A Sample Selection Strategy for Improved Generalizations From Experiments. *Eval Rev*. 2013;37(2):109-139. doi:10.1177/0193841X13516324

195. Witherspoon DP, May EM, McDonald A, Boggs S, Bámaca-Colbert M. Parenting within residential neighborhoods: A pluralistic approach with African American and Latino families at the center. In: *Advances in Child Development and Behavior*. Vol 57. Academic Press Inc.; 2019:235-279. doi:10.1016/bs.acdb.2019.05.004

196. Biemer PP. *Latent Class Analysis of Survey Error*. John Wiley & Sons, Inc.; 2010. doi:10.1002/9780470891155

197. McCutcheon A. *Applied Latent Class Analysis*. (Hagenaars J, ed.). Cambridge University Press; 2002. doi:10.1017/cbo9780511499531

198. Oberski DL. *Mixture Models: Latent Profile and Latent Class Analysis*. Accessed August 30, 2020. https://daob.nl/wp-content/uploads/2015/06/oberski-LCA.pdf

199. Rosenberg J, Beymer P, Anderson D, van Lissa C j., Schmidt J. tidyLPA: An R Package to Easily Carry Out Latent Profile Analysis (LPA) Using Open-Source or Commercial Software. *J Open Source Softw*. 2018;3(30):978. doi:10.21105/joss.00978

200. Akaike H. Fitting autoregressive models for prediction. *Ann Inst Stat Math*. 1969;21(1):243-247. doi:10.1007/BF02532251

201. Schwarz G. Estimating the Dimension of a Model. *Ann Stat*. 1978;6(2):461-464. doi:10.1214/aos/1176344136

202. Nylund KL, Asparouhov T, Muthén BO. Deciding on the Number of Classes in Latent Class Analysis and Growth Mixture Modeling: A Monte Carlo Simulation Study. *Struct Equ Model A Multidiscip J*. 2007;14(4):535-569. doi:10.1080/10705510701575396

203. Henson JM, Reise SP, Kim KH. Detecting mixtures from structural model differences using latent variable mixture modeling: A comparison of relative model fit statistics. *Struct Equ Model*. 2007;14(2):202-226. doi:10.1080/10705510709336744

204. Celeux G, Soromenho G. An entropy criterion for assessing the number of clusters in a mixture model. *J Classif*. 1996;13(2):195-212. doi:10.1007/BF01246098

205. Jiang Z. Using the iterative latent-class analysis approach to improve attribute accuracy in diagnostic classification models. *Behav Res Methods*. 2019;51(3):1075-1084. doi:10.3758/s13428-018-01191-0

206. Horvath S. DNA methylation age of human tissues and cell types. *Genome Biol*. 2013;14(10):R115. doi:10.1186/gb-2013-14-10-r115

207. Berdyshev GD, Korotaev GK, Boiarskikh G V., Vaniushin BF. Nucleotide composition of DNA and RNA from somatic tissues of humpback and its changes during spawning. *Biokhimiya*. 1967;32(5):988-993. Accessed September 12, 2020. https://europepmc.org/article/med/5628601

208. Heyn H, Li N, Ferreira HJ, et al. Distinct DNA methylomes of newborns and centenarians. *Proc Natl Acad Sci*. 2012;109(26):10522-10527. doi:10.1073/pnas.1120658109

209. Illumina. Infinium™ MethylationEPIC BeadChip. Published 2015. Accessed September 12, 2020. https://science-docs.illumina.com/documents/Microarray/infinium-methylation-epic-data-sheet-1070-2015-008/infinium-methylation-epic-ds-1070-2015-008.pdf

210. Thompson RF, Fazzari MJ, Greally JM. Experimental approaches to the study of epigenomic dysregulation in ageing. *Exp Gerontol*. 2010;45(4):255-268. doi:10.1016/j.exger.2009.12.013

211. Freire-Aradas C, Phillips C, Lareu M. Forensic Individual Age Estimation with DNA: From Initial Approaches to Methylation Tests.pdf. *Forensic Sci Rev*. 2017;29(2):122-144.

212. Jung SE, Lim SM, Hong SR, Lee EH, Shin KJ, Lee HY. DNA methylation of the ELOVL2, FHL2, KLF14, C1orf132/MIR29B2C, and TRIM59 genes for age prediction from blood, saliva, and buccal swab samples. *Forensic Sci Int Genet*. 2019;38:1-8. doi:10.1016/j.fsigen.2018.09.010

213. Lee HY, Jung SE, Oh YN, Choi A, Yang WI, Shin KJ. Epigenetic age signatures in the forensically relevant body fluid of semen: A preliminary study. *Forensic Sci Int Genet*. 2015;19:28-34. doi:10.1016/j.fsigen.2015.05.014

214. Eipel M, Mayer F, Arent T, 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;8(5):1034-1048. doi:10.18632/aging.100972

215. Xu C, Qu H, Wang G, et al. A novel strategy for forensic age prediction by DNA methylation and support vector regression model. *Sci Rep*. 2015;5(1):17788. doi:10.1038/srep17788

216. Bekaert B, Kamalandua A, Zapico SC, Van De Voorde W, Decorte R. Improved age determination of blood and teeth samples using a selected set of DNA methylation markers. Published online 2015. doi:10.1080/15592294.2015.1080413

217. Roeh S, Wiechmann T, Sauer S, Ködel M, Binder EB, Provençal N. HAM-TBS: High-accuracy methylation measurements via targeted bisulfite sequencing. *Epigenetics and Chromatin*. 2018;11(1):39. doi:10.1186/s13072-018-0209-x

APPENDIX

APPENDIX I – METHODS AND PROTOCOLS

DNA Extraction Methods:

Manual Organic Extraction via Phenol:Chloroform:Isoamyl Alcohol

Materials:

Phenol:Chloroform:Isoamyl Alcohol (25:24:1), Invitrogen Cat#15593031
Proteinase K Solution, Invitrogen Cat#4333793
Spin Column, Invitrogen Cat#AM10065
Tris-Ethylenediamineacetic Acid Buffer (10mM. pH 8.5)
Glycogen (20 μg/μL)
Sodium Acetate (7.5M)
Ethanol (100%)

Protocol:

1. Place swab tip or liquid body fluid sample into 2.0mL tube with 10 μL of Proteinase K and 190 μL of TE buffer.

2. Incubate samples on heat mixer for at least 2 hours, or overnight, at 56 ºC.

3. If extracting from swab, transfer swab to spin column and place the spin column in the same tube. Spin at 10,000 RPM for 1 minute.

4. Adjust volume to 500 μL TE buffer, and then add 500 μL PCIA.

5. Centrifuge at room temperature at 10,000 RPM for 5 minutes. Transfer the upper aqueous layer to a new tube.

6. To each sample add 1 μL glycogen, 250 μL sodium acetate, and 750 μL ethanol.

7. Centrifuge at 12,000 RPM for 20 minutes to pellet the DNA.

8. Remove supernatant without disturbing the pellet.

9. Add 150 μL of 70% ethanol to wash the pellet.

10. Centrifuge for 2 minutes at 12,000 RPM, carefully remove the supernatant.

11. Repeat step 10 once. Allow sample to dry completely.

12. Resuspend sample in 50 μL of TE buffer by pipetting up and down until no pellet remains.

Automated Extraction using EZ1 Advanced (Qiagen, CA)

Materials:

EZ1 DNA Investigator Kit, Qiagen Cat#952034
Buffer MTL, Qiagen Cat#19112

Protocol:

1. Place swab tip or liquid body fluid sample into 2.0mL tube provided with the DNA Investigator Kit (Qiagen, CA) with 10 µL of Proteinase K and 190 µL of TE buffer.

2. Incubate samples on heat mixer for at least 2 hours, or overnight, at 56 ºC.

3. Adjust volume to 500 µL TE buffer and add 400 µL of Buffer MTL.

4. Following EZ1 Advanced BioRobot handbook, load sample tubes into the correct location.

5. Load reagent cartridge for each sample to be extracted.

6. Set the robot to Tip Dance Protocol and large volume with 50 µL final elution volume in TE buffer.

DNA Quantification Methods

Quantification using Alu markers

Materials:

RampTaq, Thomas Scientific Cat#C756P80
SybrGreen, Life Technologies Cat#S7563
Alu primers at 100 µM
$MgCl_2$ 25 mM Thermo Fisher Cat#AM9530G
Buffer 10x with 15mM $MgCl_2$
dNTP mix 2.5mM each, Invitrogen Cat#10297018
Bovine Serum Albumin 20 mg/mL, Sigma Aldrich Cat# A1933-1G
Triton X Sigma Aldrich Cat#T9284-100

Protocol:

1. Prepare a serial dilution of DNA standards from 50 ng/µL to 0.5 ng/µL.

2. Prepare a qPCR master mix for the number of samples and standards plus 2, as follows:

Water	16 µL
Buffer 10X	2.3 µL
dNTPs	1.9 µL
$MgCl_2$	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
RampTaq	0.4 µL

3. For each sample and standard, use 23 µL of master mix and 2 µL of DNA, standard, or water for a no template control

4. Run samples on Rotorgene Q following established protocol for qPCR with the following cycling conditions:

95 ºC for 10 minutes

45 cycles of 92 ºC 15 seconds

56 ºC 15 seconds

72 ºC 30 seconds, acquiring on Green channel

5. Using Rotorgene Q to automatically calculate concentration of samples using the regression model produced by the standards.

Quantification using Qubit ssDNA assay

Materials:

Qubit 4.0 Thermo Fisher Cat#Q33238
Qubit ssDNA Assay kit Thermo Fisher Cat#Q10212

Protocols:

1. Set up the required number of 0.5-mL tubes for standards and samples. The Qubit® ssDNA Assay requires 2 standards.

2. Prepare the Qubit® working solution by diluting the Qubit® ssDNA Reagent 1:200 in Qubit® ssDNA Buffer. Use a clean plastic tube each time you prepare Qubit® working solution.

3. Add 190 µL of Qubit® working solution to each of the tubes used for standards.

4. Add 10 µL of each Qubit® standard to the appropriate tube, then mix by vortexing 2–3 seconds. Read standards on the Qubit 4.0.

5. Add Qubit® working solution to individual assay tubes so that the final volume in each tube after adding sample is 200 µL.

6. Add each sample to the assay tubes containing the correct volume of Qubit® working solution, then mix by vortexing 2–3 seconds. Read samples on the Qubit 4.0

Polymerase Chain Reaction for Pyrosequencing

Materials:

PyroMark® PCR Kit, Qiagen Cat#978703
BCAS4 Primers at 100µM
cg06379435 Primers at 100µM
VE_8 Primers at 100µM
ZC3H12D Primers at 100µM

Protocols:

1. Create 25 µM aliquots of each primer. Prepare Primer Mix 10x as follows:

Primer	[Final] uM	Volume (uL)
BCAS4 F	0.2	16
BCAS4 R	0.15	12
CG06379435 F	0.175	14
CG06379435 R	0.135	10.8
VE_8 F	0.11	8.8
VE_8 R	0.105	8.4
ZC3H12D F	0.165	13.2
ZC3H12D R	0.165	13.2

2. Prepare a master mix for each sample, control, and no template control, plus 1, as follows:

PCR Master Mix	per sample (µL)
Pyromark Master Mix	22.5
Coral Load	4.5
10x Primer mix	4.5
MgCl2	1.08
H2O	10.42
DNA	2
Total	45µL

3. Vortex and centrifuge samples before loading on to the thermal cycler.

4. PCR cycling conditions are as follows:

95 ºC for 15 minutes

45 cycles - 94 ºC for 30 seconds

55 ºC for 30 seconds

72 ºC for 30 seconds

72 ºC for 10 minutes

4 ºC for infinite time.

5. After PCR, store samples at -20 ºC until pyrosequencing

Bisulfite Conversion of gDNA

Materials:

EpiTect Fast Bisulfite kit, Qiagen Cat#59826

Protocols:

1. Equilibrate samples to room temperature, aliquot 200ng of sample to 0.2mL tubes.

2. Bring each sample to a volume of 40 µL, then add 85 µL of Bisulfite solution and 15 µL DNA Protect buffer.

3. Vortex and Centrifuge samples. Place on thermal cycler for 5 minutes at 95 ºC, 20 minutes at 60 ºC, 5 minutes at 95 ºC, 20 minutes at 60 ºC, and hold for infinite at 20 ºC.

4. Transfer samples to 1.5mL tubes and add 310 µL of Buffer BL and 250 µL of ethanol.

5. Transfer each sample to a labeled MinElute DNA spin column. Centrifuge each column at 12,000 x g for 1 minute. Discard the flow through.

6. Add 500 μL of Buffer BW, centrifuge at 12,000 x g for 1 minute, discard the flow through.

7. Add 500 μL of Buffer BD to each tube and incubate for 15 minutes at room temperature. Centrifuge at 12,000 x g for 1 minute. Discard the flow through.

8. Add 500 μL of Buffer BW to each sample and centrifuge at 12,000 x g for 1 minute, discard the flow through.

9. Repeat step 8.

10. Add 250 μL of ethanol to each sample and centrifuge 12,000 x g for 1 minute.

11. Place the spin column in a new 2mL tube and centrifuge at 12,000 x g for 1 minute.

12. Incubate the sample at 60 ºC for 5 minutes with lid open to completely evaporate residual ethanol.

13. Place the spin column in a new 1.5mL. 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 at 12,000 x g for 1 minute.

14. Store converted samples at -20 ºC.

Pyrosequencing on the PyroMark Q48 Autoprep

Materials:

PyroMark Q48 Autoprep System, Qiagen Cat#9002470
PyroMark Q48 Software License, Qiagen Cat#9023425
PyroMark Q48 Advanced CpG Reagents, Qiagen Cat#974022
PyroMark Q48 Magnetic Beads, Qiagen Cat#974203
PyroMark Q48 Discs, Qiagen Cat#974901
PyroMark Q48 Absorber Strips, Qiagen Cat#974912
Hi-Di™ Formamide, Thermo Fisher Cat#4311320

Protocol:

1. Turn the PyroMark Q48 Autoprep on 30 minutes prior to use. Conduct a water wash prior to use following the instructions on the screen.

2. Thaw PCR product and sequencing primers and bring Advanced CpG reagents to room temperature.

3. Using the PyroMark Q48 Software License, set up a run for each sample specifying which sequencing primer should be dispensed to each well on the sample disc.

4. Transfer the run file to the PyroMark Q48 Autoprep and begin setup of the sequencing run.

5. Add the appropriate volume of each nucleotide to the nucleotide cartridge.

6. Add appropriate volumes of denaturation solution, enzyme solution, substrate solution, and annealing buffer to the reagent cartridge.

7. For the sequencing cartridge, add the appropriate sequencing primer to each well with the determined % of HiDi Formamide that was optimized for the sequencing primers of the assay.

8. Install the sample disc, aliquot 3 μL of magnetic beads, 5 μL of binding buffer, and 10 μL of sample to each well, as specified by the sample sheet.

9. After sequencing, transfer the results file to the computer with the Q48 Software for analysis.

10. The PyroMark Q48 Software will perform methylation percent analysis at each variable position in the assay in the resulting pyrograms. Any variable position flagged with a yellow or red warning should be evaluated for inclusion in final results.

Targeted Methyl Sequencing Library Preparation

Materials:

QIAseq Targeted Methyl Custom Panel, Qiagen Cat#335602
QIAseq Targeted Methyl 96 Index Set A, Qiagen Cat#335591

Protocol:

End repair of bisulfite converted DNA:

1. Thaw bisulfite converted DNA form previous step (20) and use the total volume of 20 μl for the End repair reaction.

2. Setup the bisulfite converted DNA repair reaction mix on ice according to Table 6. Mix by pulse vortexing (3-4 times) and spin down. Keep reaction on ice.

3. Program a thermal cycler with the protocols described in Table 7.

4. Transfer reaction mix from step 2 to the thermocycler and start the bisulfite

converted DNA repair cycling program (Table 7). Place samples on ice after cycling completion.

Adapter ligation:

5. During bisulfite converted DNA repair cycling, prepare the ligation mix according to Table 8. Mix thoroughly by pulse vortexing and spin down.

6. Add 55 µl ligation master mix to each 30 µl end-repaired DNA sample from the previous step and mix by pulse vortexing and spin down.

7. Add 5 µl of IL-Me-N7## adapter to the ligation mixes from the previous step and track the used adapters.

8. After adding the adapters, mix by short vortexing, spin down, and place samples on ice.

9. Program a thermal cycler with the protocol described in Table 9.

10. Place ligation mixes from step 8 in the thermocycler and run the ligation cycling program (Table 9).

11. After cycling is complete, proceed directly with cleanup of the ligated fragments.

Cleanup of ligated fragments:

12. For sample purification, mix 90 µl (1x) QIAseq Beads with each sample by pulse vortexing. Ensure that the beads are resuspended homogenously without any visual clumps.

13. Incubate for 5 min at room temperature. Pulse spin the tube to collect all liquid on the bottom, immobilize beads on a magnet for approximately 5 min, and discard the clear supernatant.

15. Discard the supernatant. Carefully remove all remaining ethanol droplets from the tube inner walls.

16. Incubate on the magnetic stand for 5–10 min until the beads are dry. Over-drying may result in lower DNA recovery. Remove from the magnetic stand.

17. Elute by carefully resuspending in 55 µl Nuclease-free water. Incubate for 5 min at room temperature. Immobilize beads and transfer 52 µl supernatant to a new tube.

18. Mix 52 µl (1x) QIAseq Beads with each sample by pulse-vortexing and repeat steps 13−16.

Target enrichment:

20. Thaw DNA from Step 17 if stored at −15 to −30°C and amplification reagents on ice. Mix all reagents gently, spin down, and place on ice.

21.Prepare a reaction mix according to Table 10. Add each component in the order listed in this table.

22. Mix carefully 17 µl of ligated and purified DNA from Step 19 with 23 µl target-enrichment reaction mix, spin down and place on ice.

23. Program a thermal cycler with the 8 cycles Table 1.1

24. Place the PCR tubes in the thermal cycler and start the preprogrammed target enrichment cycling with the conditions outlined in Table 1.1

25. After cycling is complete, **QUICKLY** transfer samples on ice.

26. Add 2 µl of ice-cold TM Stop Solution to the 40 µl sample mix and immediately place samples back on ice.

Cleanup of the target enrichment reaction:

27. For sample purification, mix carefully 42 µl (1x) QIAseq Beads with each sample by pulse-vortexing. Ensure that the beads are resuspended homogenously without any visual clumps.

28. Incubate for 5 min at room temperature. Shortly spin down and collect all liquid on the tube bottom and immobilize beads on a magnet for approximately 5 min and discard the clear supernatant.

29. Add 200 µl fresh 70% ethanol to each bead pellet immobilized on the magnet.

30. Discard the supernatant. Carefully remove all remaining ethanol droplets from the tube inner walls.

31. Incubate on the magnetic stand for 3–7 min until the beads are dry. Over-drying may result in lower DNA recovery, so visual control is strongly recommended. Remove from the magnetic stand.

32. Elute by resuspending in 55 µl Nuclease-free water. Incubate for 5 min at room temperature. Immobilize beads and transfer 52 µl supernatant to a new tube.

33. Mix 52 µl (1x) QIAseq Beads with each sample by pulse-vortexing and repeat steps 28–31.

34. Elute by resuspending beads in 20 µl Nuclease-free water. Incubate for 5 min at room temperature. Immobilize the beads and transfer 17 µl of supernatant into a new tube. Avoid any magnetic bead carry over. Store at −15 to −30°C.

Library amplification:

For library amplification use the number of 25 cycles.

35. Thaw DNA from Step 34 and amplification reagents on ice.

36. Prepare a reaction mix by adding the components in the order according to Table 13 if working with QIAseq Methyl DNA 8-index Kit and according to Table 14 if using the QIAseq Methyl DNA 96-index I Set A, B, C, or D. Track the number of the used indexes.

37. Mix by pulse vortexing and spin down and place on ice.

38. If working with QIAseq Methyl DNA 96-index I Set A, B, C, or D, add 13.4 µl of the DNA from Step 34 to one well of the QIAseq IL-S5 Index Primer Plate in Set A, B, C or D, as illustrated in Figure 4.
Add 6.6 µl of the universal PCR mix prepared according to Table 14 to each well of the adapter plate already including the DNA. Seal the plate, mix, spin down and place on ice.

40. Place the tubes or plates with the reaction mixes from step 37 and 39 in the cycler and start the cycling program as outlined in Table 15.

41. After cycling completion, proceed with library purification. Alternatively, the amplified library can be stored at −30 to −15°C.

Clean up of amplified library:

42. Add 80 µl of ice-cold nuclease-free water to the 20 µl sample from Step 41 and mix.

43. Add 100 µl (1x) QIAseq Beads to each sample and mix thoroughly by pulse vortexing.

44. Incubate for 5 min at room temperature. Immobilize beads on a magnet and discard the clear supernatant.

45. Add 200 µl fresh 70% ethanol to each bead pellet immobilized on the magnet.

46. Discard the supernatant. Carefully remove all remaining ethanol droplets from the tube inner walls.

47. Incubate on the magnetic stand for 5–10 min until the beads are dry. Over-drying may result in lower DNA recovery. Remove from the magnetic stand.

48. Elute by carefully resuspending in 25 µl Nuclease-free water. Incubate for 5 min at room temperature. Immobilize beads on a magnet and transfer 20 µl supernatant to a new LoBind tube.

APPENDIX II – TARGETED METHYL SEQUENCING FULL RESULTS

The following tables represent the full results of the targeted methyl sequencing assay. The X seen in each row for TRIM59 indicates that no data was obtained at this position.

Blood 1	Age: 34			
Marker	Genomic position	CpG coverage	Methylated coverage	Methylation level
BCAS4	20:49410865	63	1	0.016
BL_1	19:3344242	82	23	0.280
BL_2	19:3344251	88	24	0.273
VE_8	16:86398467	39	31	0.795
ZC3H12D	6:149778105	27	27	1.000
AHRR_1	5:373476	36	32	0.889
AHRR_2	5:373490	41	33	0.805
AHRR_3	5:373494	40	37	0.925
AHRR_4	5:373529	58	35	0.603
SCGN	6:25652606	59	5	0.085
KLF14_1	7:130418281	25	2	0.080
KLF14_2	7:130418311	36	8	0.222
ELOVL2	6:11044861	63	29	0.460
FHL2	2:106015739	36	11	0.306
KLF14	7:130419116	61	1	0.016
C1orf132	1:207997026	27	23	0.852
TRIM59	3:160167977	X	X	X
ASPA	17:3379567	90	79	0.878
PDE4C	19:18343915	80	6	0.075
ITGA2B	17:42467728	61	36	0.590
PDE4C_1	19:18343915	80	6	0.075
PDE4C_2	19:18343937	107	31	0.290
PDE4C_3	19:18343941	106	18	0.170
PDE4C_4	19:18343943	110	10	0.091
PDE4C_5	19:18344003	132	8	0.061
ITGA2B	17:42467780	90	57	0.633
ADAR_1	1:154582187	38	24	0.632
ADAR_2	1:154582288	74	57	0.770
ELOVL2	6:11044861	61	49	0.803
FHL2	2:106015739	39	12	0.308
KLF14	7:130419116	61	1	0.016
C1orf132	1:207997026	24	0	0.000
TRIM59	3:160167977	X	X	X
cg12837463	7:35300228	13	9	0.692
NOX4	11:89322851	153	119	0.778
TTC7B	14:91283606	86	81	0.942
ELOVL2_1	6:11044858	48	0	0.000
ELOVL2_2	6:11044861	63	29	0.460
ELOVL2_3	6:11044867	61	49	0.803
ELOVL2_4	6:11044873	63	29	0.460
ELOVL2_5	6:11044888	77	23	0.299
PDE4C	19:18343915	80	6	0.075
EDARADD	1:236557695	59	0	0.000
ELOVL2_CpG7	6:11044867	61	49	0.803
ELOVL2_CpG5	6:11044875	66	12	0.182
ASPA	17:3379567	90	79	0.878
ELOVL2	6:11044873	63	29	0.460
PDE4C	19:18343889	60	4	0.067
EDARADD	1:236557683	60	40	0.667

Blood 2	Age: 26			
Marker	Genomic position	CpG coverage	Methylated coverage	Methylation level
BCAS4	20:49410865	70	0	0.000
BL_1	19:3344242	90	19	0.211
BL_2	19:3344251	95	16	0.168
VE_8	16:86398467	74	55	0.743
ZC3H12D	6:149778105	25	21	0.840
AHRR_1	5:373476	34	30	0.882
AHRR_2	5:373490	36	32	0.889
AHRR_3	5:373494	36	31	0.861
AHRR_4	5:373529	53	30	0.566
SCGN	6:25652606	130	10	0.077
KLF14_1	7:130418281	42	12	0.286
KLF14_2	7:130418311	41	2	0.049
ELOVL2	6:11044861	38	10	0.263
FHL2	2:106015739	30	6	0.200
KLF14	7:130419116	72	3	0.042
C1orf132	1:207997026	28	22	0.786
TRIM59	3:160167977	X	X	X
ASPA	17:3379567	108	87	0.806
PDE4C	19:18343915	67	11	0.164
ITGA2B	17:42467728	44	29	0.659
PDE4C_1	19:18343915	67	11	0.164
PDE4C_2	19:18343937	80	14	0.175
PDE4C_3	19:18343941	84	11	0.131
PDE4C_4	19:18343943	84	4	0.048
PDE4C_5	19:18344003	117	0	0.000
ITGA2B	17:42467780	55	31	0.564
ADAR_1	1:154582187	61	32	0.525
ADAR_2	1:154582288	126	83	0.659
ELOVL2	6:11044861	42	26	0.619
FHL2	2:106015739	31	10	0.323
KLF14	7:130419116	72	3	0.042
C1orf132	1:207997026	26	0	0.000
TRIM59	3:160167977	X	X	X
cg12837463	7:35300228	17	2	0.118
NOX4	11:89322851	173	136	0.786
TTC7B	14:91283606	118	108	0.915
ELOVL2_1	6:11044858	36	0	0.000
ELOVL2_2	6:11044861	38	10	0.263
ELOVL2_3	6:11044867	42	26	0.619
ELOVL2_4	6:11044873	43	19	0.442
ELOVL2_5	6:11044888	57	7	0.123
PDE4C	19:18343915	67	11	0.164
EDARADD	1:236557695	96	0	0.000
ELOVL2_CpG7	6:11044867	42	26	0.619
ELOVL2_CpG5	6:11044875	45	7	0.156
ASPA	17:3379567	108	87	0.806
ELOVL2	6:11044873	43	19	0.442
PDE4C	19:18343889	57	4	0.070
EDARADD	1:236557683	100	45	0.450

Blood 3	Age: 47			
Marker	Genomic position	CpG coverage	Methylated coverage	Methylation level
BCAS4	20:49410865	54	5	0.093
BL_1	19:3344242	67	15	0.224
BL_2	19:3344251	69	15	0.217
VE_8	16:86398467	53	41	0.774
ZC3H12D	6:149778105	16	16	1.000
AHRR_1	5:373476	29	26	0.897
AHRR_2	5:373490	30	25	0.833
AHRR_3	5:373494	31	30	0.968
AHRR_4	5:373529	36	19	0.528
SCGN	6:25652606	95	8	0.084
KLF14_1	7:130418281	31	6	0.194
KLF14_2	7:130418311	18	3	0.167
ELOVL2	6:11044861	42	12	0.286
FHL2	2:106015739	37	8	0.216
KLF14	7:130419116	30	1	0.033
C1orf132	1:207997026	28	23	0.821
TRIM59	3:160167977	X	X	X
ASPA	17:3379567	54	45	0.833
PDE4C	19:18343915	66	24	0.364
ITGA2B	17:42467728	65	44	0.677
PDE4C_1	19:18343915	66	24	0.364
PDE4C_2	19:18343937	75	20	0.267
PDE4C_3	19:18343941	76	11	0.145
PDE4C_4	19:18343943	76	5	0.066
PDE4C_5	19:18344003	83	6	0.072
ITGA2B	17:42467780	68	44	0.647
ADAR_1	1:154582187	30	4	0.133
ADAR_2	1:154582288	46	21	0.457
ELOVL2	6:11044861	46	39	0.848
FHL2	2:106015739	38	14	0.368
KLF14	7:130419116	30	1	0.033
C1orf132	1:207997026	30	0	0.000
TRIM59	3:160167977	X	X	X
cg12837463	7:35300228	29	19	0.655
NOX4	11:89322851	102	70	0.686
TTC7B	14:91283606	58	53	0.914
ELOVL2_1	6:11044858	42	0	0.000
ELOVL2_2	6:11044861	42	12	0.286
ELOVL2_3	6:11044867	46	39	0.848
ELOVL2_4	6:11044873	46	29	0.630
ELOVL2_5	6:11044888	53	8	0.151
PDE4C	19:18343915	66	24	0.364
EDARADD	1:236557695	37	0	0.000
ELOVL2_CpG7	6:11044867	46	39	0.848
ELOVL2_CpG5	6:11044875	46	8	0.174
ASPA	17:3379567	54	45	0.833
ELOVL2	6:11044873	46	29	0.630
PDE4C	19:18343889	53	5	0.094
EDARADD	1:236557683	37	25	0.676

Blood 4	Age: 63			
Marker	Genomic position	CpG coverage	Methylated coverage	Methylation level
BCAS4	20:49410865	22	3	0.136
BL_1	19:3344242	54	1	0.019
BL_2	19:3344251	58	5	0.086
VE_8	16:86398467	25	21	0.840
ZC3H12D	6:149778105	5	5	1.000
AHRR_1	5:373476	15	12	0.800
AHRR_2	5:373490	16	10	0.625
AHRR_3	5:373494	16	14	0.875
AHRR_4	5:373529	23	9	0.391
SCGN	6:25652606	45	4	0.089
KLF14_1	7:130418281	9	1	0.111
KLF14_2	7:130418311	16	2	0.125
ELOVL2	6:11044861	23	7	0.304
FHL2	2:106015739	9	0	0.000
KLF14	7:130419116	7	0	0.000
C1orf132	1:207997026	6	3	0.500
TRIM59	3:160167977	X	X	X
ASPA	17:3379567	17	12	0.706
PDE4C	19:18343915	43	2	0.047
ITGA2B	17:42467728	19	10	0.526
PDE4C_1	19:18343915	43	2	0.047
PDE4C_2	19:18343937	68	25	0.368
PDE4C_3	19:18343941	68	10	0.147
PDE4C_4	19:18343943	67	9	0.134
PDE4C_5	19:18344003	74	11	0.149
ITGA2B	17:42467780	25	14	0.560
ADAR_1	1:154582187	27	8	0.296
ADAR_2	1:154582288	50	22	0.440
ELOVL2	6:11044861	24	20	0.833
FHL2	2:106015739	10	1	0.100
KLF14	7:130419116	7	0	0.000
C1orf132	1:207997026	5	0	0.000
TRIM59	3:160167977	X	X	X
cg12837463	7:35300228	5	5	1.000
NOX4	11:89322851	114	85	0.746
TTC7B	14:91283606	51	40	0.784
ELOVL2_1	6:11044858	18	0	0.000
ELOVL2_2	6:11044861	23	7	0.304
ELOVL2_3	6:11044867	24	20	0.833
ELOVL2_4	6:11044873	29	21	0.724
ELOVL2_5	6:11044888	39	13	0.333
PDE4C	19:18343915	43	2	0.047
EDARADD	1:236557695	32	0	0.000
ELOVL2_CpG7	6:11044867	24	20	0.833
ELOVL2_CpG5	6:11044875	25	9	0.360
ASPA	17:3379567	17	12	0.706
ELOVL2	6:11044873	29	21	0.724
PDE4C	19:18343889	42	29	0.690
EDARADD	1:236557683	33	14	0.424

Blood 5	Age: 19			
Marker	Genomic position	CpG coverage	Methylated coverage	Methylation level
BCAS4	20:49410865	97	7	0.072
BL_1	19:3344242	118	10	0.085
BL_2	19:3344251	130	11	0.085
VE_8	16:86398467	43	36	0.837
ZC3H12D	6:149778105	24	22	0.917
AHRR_1	5:373476	29	28	0.966
AHRR_2	5:373490	37	31	0.838
AHRR_3	5:373494	35	31	0.886
AHRR_4	5:373529	51	37	0.725
SCGN	6:25652606	125	11	0.088
KLF14_1	7:130418281	46	7	0.152
KLF14_2	7:130418311	40	0	0.000
ELOVL2	6:11044861	59	8	0.136
FHL2	2:106015739	40	16	0.400
KLF14	7:130419116	56	1	0.018
C1orf132	1:207997026	14	12	0.857
TRIM59	3:160167977	X	X	X
ASPA	17:3379567	101	88	0.871
PDE4C	19:18343915	159	36	0.226
ITGA2B	17:42467728	74	52	0.703
PDE4C_1	19:18343915	159	36	0.226
PDE4C_2	19:18343937	163	51	0.313
PDE4C_3	19:18343941	166	2	0.012
PDE4C_4	19:18343943	165	14	0.085
PDE4C_5	19:18344003	203	3	0.015
ITGA2B	17:42467780	82	44	0.537
ADAR_1	1:154582187	81	36	0.444
ADAR_2	1:154582288	109	70	0.642
ELOVL2	6:11044861	62	34	0.548
FHL2	2:106015739	38	14	0.368
KLF14	7:130419116	56	1	0.018
C1orf132	1:207997026	14	0	0.000
TRIM59	3:160167977	X	X	X
cg12837463	7:35300228	12	7	0.583
NOX4	11:89322851	126	75	0.595
TTC7B	14:91283606	52	46	0.885
ELOVL2_1	6:11044858	54	0	0.000
ELOVL2_2	6:11044861	59	8	0.136
ELOVL2_3	6:11044867	62	34	0.548
ELOVL2_4	6:11044873	59	24	0.407
ELOVL2_5	6:11044888	108	37	0.343
PDE4C	19:18343915	159	36	0.226
EDARADD	1:236557695	72	0	0.000
ELOVL2_CpG7	6:11044867	62	34	0.548
ELOVL2_CpG5	6:11044875	58	9	0.155
ASPA	17:3379567	101	88	0.871
ELOVL2	6:11044873	59	24	0.407
PDE4C	19:18343889	158	22	0.139
EDARADD	1:236557683	72	44	0.611

Saliva 2	Age: 28			
Marker	Genomic position	CpG coverage	Methylated coverage	Methylation level
BCAS4	20:49410865	52	41	0.788
BL_1	19:3344242	59	0	0.000
BL_2	19:3344251	62	0	0.000
VE_8	16:86398467	49	36	0.735
ZC3H12D	6:149778105	34	24	0.706
AHRR_1	5:373476	5	2	0.400
AHRR_2	5:373490	7	6	0.857
AHRR_3	5:373494	6	6	1.000
AHRR_4	5:373529	11	6	0.545
SCGN	6:25652606	24	0	0.000
KLF14_1	7:130418281	10	0	0.000
KLF14_2	7:130418311	8	1	0.125
ELOVL2	6:11044861	39	13	0.333
FHL2	2:106015739	26	0	0.000
KLF14	7:130419116	57	3	0.053
C1orf132	1:207997026	14	13	0.929
TRIM59	3:160167977	X	X	X
ASPA	17:3379567	58	2	0.034
PDE4C	19:18343915	75	33	0.440
ITGA2B	17:42467728	54	43	0.796
PDE4C_1	19:18343915	75	33	0.440
PDE4C_2	19:18343937	120	0	0.000
PDE4C_3	19:18343941	126	1	0.008
PDE4C_4	19:18343943	126	2	0.016
PDE4C_5	19:18344003	133	0	0.000
ITGA2B	17:42467780	59	53	0.898
ADAR_1	1:154582187	31	29	0.935
ADAR_2	1:154582288	60	46	0.767
ELOVL2	6:11044861	41	35	0.854
FHL2	2:106015739	27	1	0.037
KLF14	7:130419116	57	3	0.053
C1orf132	1:207997026	14	1	0.071
TRIM59	3:160167977	X	X	X
cg12837463	7:35300228	33	7	0.212
NOX4	11:89322851	74	6	0.081
TTC7B	14:91283606	60	34	0.567
ELOVL2_1	6:11044858	35	5	0.143
ELOVL2_2	6:11044861	39	13	0.333
ELOVL2_3	6:11044867	41	35	0.854
ELOVL2_4	6:11044873	41	14	0.341
ELOVL2_5	6:11044888	58	23	0.397
PDE4C	19:18343915	75	33	0.440
EDARADD	1:236557695	41	1	0.024
ELOVL2_CpG7	6:11044867	41	35	0.854
ELOVL2_CpG5	6:11044875	46	14	0.304
ASPA	17:3379567	58	2	0.034
ELOVL2	6:11044873	41	14	0.341
PDE4C	19:18343889	46	5	0.109
EDARADD	1:236557683	41	19	0.463

Saliva 4	Age: 41			
Marker	Genomic position	CpG coverage	Methylated coverage	Methylation level
BCAS4	20:49410865	25	15	0.600
BL_1	19:3344242	64	0	0.000
BL_2	19:3344251	70	0	0.000
VE_8	16:86398467	36	34	0.944
ZC3H12D	6:149778105	3	3	1.000
AHRR_1	5:373476	6	6	1.000
AHRR_2	5:373490	6	4	0.667
AHRR_3	5:373494	6	6	1.000
AHRR_4	5:373529	6	5	0.833
SCGN	6:25652606	17	2	0.118
KLF14_1	7:130418281	2	0	0.000
KLF14_2	7:130418311	2	0	0.000
ELOVL2	6:11044861	10	9	0.900
FHL2	2:106015739	4	0	0.000
KLF14	7:130419116	13	0	0.000
C1orf132	1:207997026	6	6	1.000
TRIM59	3:160167977	X	X	X
ASPA	17:3379567	46	16	0.348
PDE4C	19:18343915	9	2	0.222
ITGA2B	17:42467728	33	24	0.727
PDE4C_1	19:18343915	9	2	0.222
PDE4C_2	19:18343937	9	0	0.000
PDE4C_3	19:18343941	9	0	0.000
PDE4C_4	19:18343943	9	0	0.000
PDE4C_5	19:18344003	15	0	0.000
ITGA2B	17:42467780	62	60	0.968
ADAR_1	1:154582187	12	9	0.750
ADAR_2	1:154582288	25	18	0.720
ELOVL2	6:11044861	11	11	1.000
FHL2	2:106015739	4	1	0.250
KLF14	7:130419116	13	0	0.000
C1orf132	1:207997026	6	0	0.000
TRIM59	3:160167977	X	X	X
cg12837463	7:35300228	23	20	0.870
NOX4	11:89322851	79	21	0.266
TTC7B	14:91283606	47	45	0.957
ELOVL2_1	6:11044858	11	0	0.000
ELOVL2_2	6:11044861	10	9	0.900
ELOVL2_3	6:11044867	11	11	1.000
ELOVL2_4	6:11044873	10	10	1.000
ELOVL2_5	6:11044888	11	10	0.909
PDE4C	19:18343915	9	2	0.222
EDARADD	1:236557695	20	0	0.000
ELOVL2_CpG7	6:11044867	11	11	1.000
ELOVL2_CpG5	6:11044875	11	1	0.091
ASPA	17:3379567	46	16	0.348
ELOVL2	6:11044873	10	10	1.000
PDE4C	19:18343889	7	0	0.000
EDARADD	1:236557683	21	9	0.429

Saliva 5	Age: 54			
Marker	Genomic position	CpG coverage	Methylated coverage	Methylation level
BCAS4	20:49410865	49	40	0.816
BL_1	19:3344242	86	0	0.000
BL_2	19:3344251	100	0	0.000
VE_8	16:86398467	35	22	0.629
ZC3H12D	6:149778105	13	8	0.615
AHRR_1	5:373476	3	2	0.667
AHRR_2	5:373490	9	9	1.000
AHRR_3	5:373494	10	10	1.000
AHRR_4	5:373529	18	17	0.944
SCGN	6:25652606	52	8	0.154
KLF14_1	7:130418281	7	2	0.286
KLF14_2	7:130418311	6	2	0.333
ELOVL2	6:11044861	9	5	0.556
FHL2	2:106015739	19	0	0.000
KLF14	7:130419116	33	0	0.000
C1orf132	1:207997026	16	14	0.875
TRIM59	3:160167977	X	X	X
ASPA	17:3379567	100	9	0.090
PDE4C	19:18343915	66	13	0.197
ITGA2B	17:42467728	88	83	0.943
PDE4C_1	19:18343915	66	13	0.197
PDE4C_2	19:18343937	72	30	0.417
PDE4C_3	19:18343941	72	8	0.111
PDE4C_4	19:18343943	74	11	0.149
PDE4C_5	19:18344003	86	3	0.035
ITGA2B	17:42467780	98	93	0.949
ADAR_1	1:154582187	52	49	0.942
ADAR_2	1:154582288	113	107	0.947
ELOVL2	6:11044861	9	9	1.000
FHL2	2:106015739	17	1	0.059
KLF14	7:130419116	33	0	0.000
C1orf132	1:207997026	16	0	0.000
TRIM59	3:160167977	X	X	X
cg12837463	7:35300228	20	0	0.000
NOX4	11:89322851	160	23	0.144
TTC7B	14:91283606	161	129	0.801
ELOVL2_1	6:11044858	9	0	0.000
ELOVL2_2	6:11044861	9	5	0.556
ELOVL2_3	6:11044867	9	9	1.000
ELOVL2_4	6:11044873	9	7	0.778
ELOVL2_5	6:11044888	9	9	1.000
PDE4C	19:18343915	66	13	0.197
EDARADD	1:236557695	57	0	0.000
ELOVL2_CpG7	6:11044867	9	9	1.000
ELOVL2_CpG5	6:11044875	8	0	0.000
ASPA	17:3379567	100	9	0.090
ELOVL2	6:11044873	9	7	0.778
PDE4C	19:18343889	54	7	0.130
EDARADD	1:236557683	65	11	0.169

Saliva 7	Age: 47			
Marker	Genomic position	CpG coverage	Methylated coverage	Methylation level
BCAS4	20:49410865	57	21	0.368
BL_1	19:3344242	101	1	0.010
BL_2	19:3344251	107	0	0.000
VE_8	16:86398467	54	38	0.704
ZC3H12D	6:149778105	39	32	0.821
AHRR_1	5:373476	32	23	0.719
AHRR_2	5:373490	39	32	0.821
AHRR_3	5:373494	41	37	0.902
AHRR_4	5:373529	57	48	0.842
SCGN	6:25652606	60	5	0.083
KLF14_1	7:130418281	10	3	0.300
KLF14_2	7:130418311	10	1	0.100
ELOVL2	6:11044861	30	24	0.800
FHL2	2:106015739	36	6	0.167
KLF14	7:130419116	38	4	0.105
C1orf132	1:207997026	24	24	1.000
TRIM59	3:160167977	X	X	X
ASPA	17:3379567	35	2	0.057
PDE4C	19:18343915	48	13	0.271
ITGA2B	17:42467728	44	40	0.909
PDE4C_1	19:18343915	48	13	0.271
PDE4C_2	19:18343937	72	18	0.250
PDE4C_3	19:18343941	72	25	0.347
PDE4C_4	19:18343943	72	4	0.056
PDE4C_5	19:18344003	103	3	0.029
ITGA2B	17:42467780	54	51	0.944
ADAR_1	1:154582187	33	33	1.000
ADAR_2	1:154582288	62	60	0.968
ELOVL2	6:11044861	32	29	0.906
FHL2	2:106015739	36	7	0.194
KLF14	7:130419116	38	4	0.105
C1orf132	1:207997026	25	0	0.000
TRIM59	3:160167977	X	X	X
cg12837463	7:35300228	15	7	0.467
NOX4	11:89322851	129	50	0.388
TTC7B	14:91283606	67	60	0.896
ELOVL2_1	6:11044858	28	0	0.000
ELOVL2_2	6:11044861	30	24	0.800
ELOVL2_3	6:11044867	32	29	0.906
ELOVL2_4	6:11044873	32	30	0.938
ELOVL2_5	6:11044888	36	25	0.694
PDE4C	19:18343915	48	13	0.271
EDARADD	1:236557695	32	0	0.000
ELOVL2_CpG7	6:11044867	32	29	0.906
ELOVL2_CpG5	6:11044875	32	15	0.469
ASPA	17:3379567	35	2	0.057
ELOVL2	6:11044873	32	30	0.938
PDE4C	19:18343889	43	8	0.186
EDARADD	1:236557683	36	7	0.194

Saliva 8	Age: 62			
Marker	Genomic position	CpG coverage	Methylated coverage	Methylation level
BCAS4	20:49410865	X	X	X
BL_1	19:3344242	19	0	0.000
BL_2	19:3344251	19	0	0.000
VE_8	16:86398467	44	34	0.773
ZC3H12D	6:149778105	X	X	X
AHRR_1	5:373476	X	X	X
AHRR_2	5:373490	X	X	X
AHRR_3	5:373494	X	X	X
AHRR_4	5:373529	42	0	0.000
SCGN	6:25652606	2	0	0.000
KLF14_1	7:130418281	X	X	X
KLF14_2	7:130418311	X	X	X
ELOVL2	6:11044861	X	X	X
FHL2	2:106015739	4	0	0.000
KLF14	7:130419116	2	0	0.000
C1orf132	1:207997026	X	X	X
TRIM59	3:160167977	X	X	X
ASPA	17:3379567	7	7	1.000
PDE4C	19:18343915	14	10	0.714
ITGA2B	17:42467728	18	18	1.000
PDE4C_1	19:18343915	14	10	0.714
PDE4C_2	19:18343937	13	0	0.000
PDE4C_3	19:18343941	14	0	0.000
PDE4C_4	19:18343943	14	0	0.000
PDE4C_5	19:18344003	19	0	0.000
ITGA2B	17:42467780	18	18	1.000
ADAR_1	1:154582187	22	22	1.000
ADAR_2	1:154582288	37	36	0.973
ELOVL2	6:11044861	X	X	X
FHL2	2:106015739	4	0	0.000
KLF14	7:130419116	2	0	0.000
C1orf132	1:207997026	X	X	X
TRIM59	3:160167977	X	X	X
cg12837463	7:35300228	2	0	0.000
NOX4	11:89322851	10	9	0.900
TTC7B	14:91283606	179	179	1.000
ELOVL2_1	6:11044858	1	1	1.000
ELOVL2_2	6:11044861	2	2	1.000
ELOVL2_3	6:11044867	X	X	X
ELOVL2_4	6:11044873	X	X	X
ELOVL2_5	6:11044888	X	X	X
PDE4C	19:18343915	14	10	0.714
EDARADD	1:236557695	12	0	0.000
ELOVL2_CpG7	6:11044867	X	X	X
ELOVL2_CpG5	6:11044875	1	0	0.000
ASPA	17:3379567	7	7	1.000
ELOVL2	6:11044873	1	1	1.000
PDE4C	19:18343889	10	0	0.000
EDARADD	1:236557683	13	10	0.769

Vag. Epi. 10	Age: 24			
Marker	Genomic position	CpG coverage	Methylated coverage	Methylation level
BCAS4	20:49410865	163	83	0.509
BL_1	19:3344242	163	1	0.006
BL_2	19:3344251	170	0	0.000
VE_8	16:86398467	94	3	0.032
ZC3H12D	6:149778105	38	30	0.789
AHRR_1	5:373476	7	5	0.714
AHRR_2	5:373490	7	7	1.000
AHRR_3	5:373494	7	7	1.000
AHRR_4	5:373529	15	9	0.600
SCGN	6:25652606	123	3	0.024
KLF14_1	7:130418281	8	0	0.000
KLF14_2	7:130418311	12	0	0.000
ELOVL2	6:11044861	26	6	0.231
FHL2	2:106015739	23	1	0.043
KLF14	7:130419116	65	0	0.000
C1orf132	1:207997026	32	31	0.969
TRIM59	3:160167977	X	X	X
ASPA	17:3379567	149	91	0.611
PDE4C	19:18343915	66	24	0.364
ITGA2B	17:42467728	96	89	0.927
PDE4C_1	19:18343915	66	24	0.364
PDE4C_2	19:18343937	79	6	0.076
PDE4C_3	19:18343941	78	0	0.000
PDE4C_4	19:18343943	80	2	0.025
PDE4C_5	19:18344003	94	4	0.043
ITGA2B	17:42467780	114	87	0.763
ADAR_1	1:154582187	107	56	0.523
ADAR_2	1:154582288	152	135	0.888
ELOVL2	6:11044861	45	38	0.844
FHL2	2:106015739	32	1	0.031
KLF14	7:130419116	65	0	0.000
C1orf132	1:207997026	34	0	0.000
TRIM59	3:160167977	X	X	X
cg12837463	7:35300228	38	29	0.763
NOX4	11:89322851	342	76	0.222
TTC7B	14:91283606	110	76	0.691
ELOVL2_1	6:11044858	13	0	0.000
ELOVL2_2	6:11044861	26	6	0.231
ELOVL2_3	6:11044867	45	38	0.844
ELOVL2_4	6:11044873	51	34	0.667
ELOVL2_5	6:11044888	67	29	0.433
PDE4C	19:18343915	66	24	0.364
EDARADD	1:236557695	98	0	0.000
ELOVL2_CpG7	6:11044867	45	38	0.844
ELOVL2_CpG5	6:11044875	50	34	0.680
ASPA	17:3379567	149	91	0.611
ELOVL2	6:11044873	51	34	0.667
PDE4C	19:18343889	55	1	0.018
EDARADD	1:236557683	100	96	0.960

Vag. Epi. 30	Age: 33			
Marker	Genomic position	CpG coverage	Methylated coverage	Methylation level
BCAS4	20:49410865	63	17	0.270
BL_1	19:3344242	34	0	0.000
BL_2	19:3344251	36	0	0.000
VE_8	16:86398467	31	6	0.194
ZC3H12D	6:149778105	21	20	0.952
AHRR_1	5:373476	7	7	1.000
AHRR_2	5:373490	7	6	0.857
AHRR_3	5:373494	7	7	1.000
AHRR_4	5:373529	7	5	0.714
SCGN	6:25652606	38	2	0.053
KLF14_1	7:130418281	X	X	X
KLF14_2	7:130418311	X	X	X
ELOVL2	6:11044861	1	0	0.000
FHL2	2:106015739	12	0	0.000
KLF14	7:130419116	10	0	0.000
C1orf132	1:207997026	3	3	1.000
TRIM59	3:160167977	X	X	X
ASPA	17:3379567	15	9	0.600
PDE4C	19:18343915	18	1	0.056
ITGA2B	17:42467728	27	24	0.889
PDE4C_1	19:18343915	18	1	0.056
PDE4C_2	19:18343937	21	0	0.000
PDE4C_3	19:18343941	21	12	0.571
PDE4C_4	19:18343943	22	12	0.545
PDE4C_5	19:18344003	52	1	0.019
ITGA2B	17:42467780	28	18	0.643
ADAR_1	1:154582187	25	14	0.560
ADAR_2	1:154582288	43	38	0.884
ELOVL2	6:11044861	5	0	0.000
FHL2	2:106015739	12	0	0.000
KLF14	7:130419116	10	0	0.000
C1orf132	1:207997026	3	0	0.000
TRIM59	3:160167977	X	X	X
cg12837463	7:35300228	33	24	0.727
NOX4	11:89322851	217	35	0.161
TTC7B	14:91283606	103	93	0.903
ELOVL2_1	6:11044858	1	0	0.000
ELOVL2_2	6:11044861	1	0	0.000
ELOVL2_3	6:11044867	5	0	0.000
ELOVL2_4	6:11044873	12	4	0.333
ELOVL2_5	6:11044888	12	4	0.333
PDE4C	19:18343915	18	1	0.056
EDARADD	1:236557695	47	0	0.000
ELOVL2_CpG7	6:11044867	5	0	0.000
ELOVL2_CpG5	6:11044875	11	4	0.364
ASPA	17:3379567	15	9	0.600
ELOVL2	6:11044873	12	4	0.333
PDE4C	19:18343889	17	1	0.059
EDARADD	1:236557683	47	39	0.830

Sem1_1	Age: 20			
Marker	Genomic position	CpG coverage	Methylated coverage	Methylation level
BCAS4	20:49410865	78	47	0.603
BL_1	19:3344242	115	0	0.000
BL_2	19:3344251	119	0	0.000
VE_8	16:86398467	62	52	0.839
ZC3H12D	6:149778105	37	22	0.595
AHRR_1	5:373476	20	20	1.000
AHRR_2	5:373490	22	21	0.955
AHRR_3	5:373494	23	22	0.957
AHRR_4	5:373529	25	25	1.000
SCGN	6:25652606	55	7	0.127
KLF14_1	7:130418281	30	5	0.167
KLF14_2	7:130418311	31	2	0.065
ELOVL2	6:11044861	40	21	0.525
FHL2	2:106015739	32	0	0.000
KLF14	7:130419116	47	4	0.085
C1orf132	1:207997026	10	9	0.900
TRIM59	3:160167977	X	X	X
ASPA	17:3379567	96	9	0.094
PDE4C	19:18343915	124	28	0.226
ITGA2B	17:42467728	52	45	0.865
PDE4C_1	19:18343915	124	28	0.226
PDE4C_2	19:18343937	132	23	0.174
PDE4C_3	19:18343941	134	15	0.112
PDE4C_4	19:18343943	136	25	0.184
PDE4C_5	19:18344003	152	1	0.007
ITGA2B	17:42467780	62	54	0.871
ADAR_1	1:154582187	56	49	0.875
ADAR_2	1:154582288	105	91	0.867
ELOVL2	6:11044861	42	29	0.690
FHL2	2:106015739	32	1	0.031
KLF14	7:130419116	47	4	0.085
C1orf132	1:207997026	10	0	0.000
TRIM59	3:160167977	X	X	X
cg12837463	7:35300228	14	6	0.429
NOX4	11:89322851	173	14	0.081
TTC7B	14:91283606	122	101	0.828
ELOVL2_1	6:11044858	35	0	0.000
ELOVL2_2	6:11044861	40	21	0.525
ELOVL2_3	6:11044867	42	29	0.690
ELOVL2_4	6:11044873	37	14	0.378
ELOVL2_5	6:11044888	52	34	0.654
PDE4C	19:18343915	124	28	0.226
EDARADD	1:236557695	97	0	0.000
ELOVL2_CpG7	6:11044867	42	29	0.690
ELOVL2_CpG5	6:11044875	42	10	0.238
ASPA	17:3379567	96	9	0.094
ELOVL2	6:11044873	37	14	0.378
PDE4C	19:18343889	116	23	0.198
EDARADD	1:236557683	97	33	0.340

Sem1_2	Age: 20			
Marker	Genomic position	CpG coverage	Methylated coverage	Methylation level
BCAS4	20:49410865	123	79	0.642
BL_1	19:3344242	169	0	0.000
BL_2	19:3344251	176	0	0.000
VE_8	16:86398467	84	57	0.679
ZC3H12D	6:149778105	39	31	0.795
AHRR_1	5:373476	69	62	0.899
AHRR_2	5:373490	78	75	0.962
AHRR_3	5:373494	81	73	0.901
AHRR_4	5:373529	106	73	0.689
SCGN	6:25652606	167	5	0.030
KLF14_1	7:130418281	99	4	0.040
KLF14_2	7:130418311	87	5	0.057
ELOVL2	6:11044861	64	26	0.406
FHL2	2:106015739	85	4	0.047
KLF14	7:130419116	132	8	0.061
C1orf132	1:207997026	52	48	0.923
TRIM59	3:160167977	X	X	X
ASPA	17:3379567	93	19	0.204
PDE4C	19:18343915	215	44	0.205
ITGA2B	17:42467728	112	99	0.884
PDE4C_1	19:18343915	215	44	0.205
PDE4C_2	19:18343937	232	40	0.172
PDE4C_3	19:18343941	236	26	0.110
PDE4C_4	19:18343943	236	43	0.182
PDE4C_5	19:18344003	274	5	0.018
ITGA2B	17:42467780	132	123	0.932
ADAR_1	1:154582187	67	55	0.821
ADAR_2	1:154582288	121	115	0.950
ELOVL2	6:11044861	68	54	0.794
FHL2	2:106015739	82	8	0.098
KLF14	7:130419116	132	8	0.061
C1orf132	1:207997026	52	1	0.019
TRIM59	3:160167977	X	X	X
cg12837463	7:35300228	46	23	0.500
NOX4	11:89322851	182	15	0.082
TTC7B	14:91283606	175	146	0.834
ELOVL2_1	6:11044858	61	1	0.016
ELOVL2_2	6:11044861	64	26	0.406
ELOVL2_3	6:11044867	68	54	0.794
ELOVL2_4	6:11044873	72	43	0.597
ELOVL2_5	6:11044888	103	41	0.398
PDE4C	19:18343915	215	44	0.205
EDARADD	1:236557695	85	1	0.012
ELOVL2_CpG7	6:11044867	68	54	0.794
ELOVL2_CpG5	6:11044875	73	25	0.342
ASPA	17:3379567	93	19	0.204
ELOVL2	6:11044873	72	43	0.597
PDE4C	19:18343889	176	19	0.108
EDARADD	1:236557683	88	27	0.307

Sem2_1	Age: 29			
Marker	Genomic position	CpG coverage	Methylated coverage	Methylation level
BCAS4	20:49410865	99.000	42.000	0.424
BL_1	19:3344242	86.000	2.000	0.023
BL_2	19:3344251	91.000	2.000	0.022
VE_8	16:86398467	66.000	48.000	0.727
ZC3H12D	6:149778105	18.000	18.000	1.000
AHRR_1	5:373476	31.000	27.000	0.871
AHRR_2	5:373490	35.000	30.000	0.857
AHRR_3	5:373494	34.000	33.000	0.971
AHRR_4	5:373529	42.000	35.000	0.833
SCGN	6:25652606	83.000	10.000	0.120
KLF14_1	7:130418281	52.000	10.000	0.192
KLF14_2	7:130418311	62.000	0.000	0.000
ELOVL2	6:11044861	41.000	34.000	0.829
FHL2	2:106015739	63.000	8.000	0.127
KLF14	7:130419116	71.000	5.000	0.070
C1orf132	1:207997026	16.000	14.000	0.875
TRIM59	3:160167977	X	X	X
ASPA	17:3379567	54.000	11.000	0.204
PDE4C	19:18343915	69.000	31.000	0.449
ITGA2B	17:42467728	51.000	48.000	0.941
PDE4C_1	19:18343915	69.000	31.000	0.449
PDE4C_2	19:18343937	81.000	26.000	0.321
PDE4C_3	19:18343941	81.000	16.000	0.198
PDE4C_4	19:18343943	83.000	26.000	0.313
PDE4C_5	19:18344003	89.000	2.000	0.022
ITGA2B	17:42467780	77.000	77.000	1.000
ADAR_1	1:154582187	60.000	51.000	0.850
ADAR_2	1:154582288	84.000	71.000	0.845
ELOVL2	6:11044861	50.000	50.000	1.000
FHL2	2:106015739	68.000	6.000	0.088
KLF14	7:130419116	71.000	5.000	0.070
C1orf132	1:207997026	17.000	0.000	0.000
TRIM59	3:160167977	X	X	X
cg12837463	7:35300228	18.000	7.000	0.389
NOX4	11:89322851	125.000	36.000	0.288
TTC7B	14:91283606	145.000	118.000	0.814
ELOVL2_1	6:11044858	37.000	0.000	0.000
ELOVL2_2	6:11044861	41.000	34.000	0.829
ELOVL2_3	6:11044867	50.000	50.000	1.000
ELOVL2_4	6:11044873	51.000	47.000	0.922
ELOVL2_5	6:11044888	69.000	31.000	0.449
PDE4C	19:18343915	69.000	31.000	0.449
EDARADD	1:236557695	79.000	0.000	0.000
ELOVL2_CpG7	6:11044867	50.000	50.000	1.000
ELOVL2_CpG5	6:11044875	50.000	30.000	0.600
ASPA	17:3379567	54.000	11.000	0.204
ELOVL2	6:11044873	51.000	47.000	0.922
PDE4C	19:18343889	58.000	8.000	0.138
EDARADD	1:236557683	81.000	43.000	0.531

Sem2_2	Age: 29			
Marker	Genomic position	CpG coverage	Methylated coverage	Methylation level
BCAS4	20:49410865	243	139	0.572
BL_1	19:3344242	321	3	0.009
BL_2	19:3344251	347	1	0.003
VE_8	16:86398467	194	139	0.716
ZC3H12D	6:149778105	108	83	0.769
AHRR_1	5:373476	142	122	0.859
AHRR_2	5:373490	166	150	0.904
AHRR_3	5:373494	164	149	0.909
AHRR_4	5:373529	199	141	0.709
SCGN	6:25652606	245	14	0.057
KLF14_1	7:130418281	163	13	0.080
KLF14_2	7:130418311	129	12	0.093
ELOVL2	6:11044861	152	92	0.605
FHL2	2:106015739	162	2	0.012
KLF14	7:130419116	184	9	0.049
C1orf132	1:207997026	99	96	0.970
TRIM59	3:160167977	X	X	X
ASPA	17:3379567	278	45	0.162
PDE4C	19:18343915	317	75	0.237
ITGA2B	17:42467728	233	197	0.845
PDE4C_1	19:18343915	317	75	0.237
PDE4C_2	19:18343937	367	99	0.270
PDE4C_3	19:18343941	371	69	0.186
PDE4C_4	19:18343943	372	59	0.159
PDE4C_5	19:18344003	405	19	0.047
ITGA2B	17:42467780	252	230	0.913
ADAR_1	1:154582187	155	123	0.794
ADAR_2	1:154582288	215	202	0.940
ELOVL2	6:11044861	155	139	0.897
FHL2	2:106015739	158	5	0.032
KLF14	7:130419116	184	9	0.049
C1orf132	1:207997026	94	2	0.021
TRIM59	3:160167977	X	X	X
cg12837463	7:35300228	72	30	0.417
NOX4	11:89322851	418	59	0.141
TTC7B	14:91283606	340	259	0.762
ELOVL2_1	6:11044858	146	2	0.014
ELOVL2_2	6:11044861	152	92	0.605
ELOVL2_3	6:11044867	155	139	0.897
ELOVL2_4	6:11044873	161	113	0.702
ELOVL2_5	6:11044888	216	95	0.440
PDE4C	19:18343915	317	75	0.237
EDARADD	1:236557695	221	1	0.005
ELOVL2_CpG7	6:11044867	155	139	0.897
ELOVL2_CpG5	6:11044875	161	65	0.404
ASPA	17:3379567	278	45	0.162
ELOVL2	6:11044873	161	113	0.702
PDE4C	19:18343889	257	56	0.218
EDARADD	1:236557683	229	60	0.262

Sem3_1	Age: 43			
Marker	Genomic position	CpG coverage	Methylated coverage	Methylation level
BCAS4	20:49410865	170	38	0.224
BL_1	19:3344242	154	0	0.000
BL_2	19:3344251	152	0	0.000
VE_8	16:86398467	75	56	0.747
ZC3H12D	6:149778105	44	41	0.932
AHRR_1	5:373476	82	28	0.341
AHRR_2	5:373490	93	28	0.301
AHRR_3	5:373494	87	44	0.506
AHRR_4	5:373529	104	20	0.192
SCGN	6:25652606	251	26	0.104
KLF14_1	7:130418281	111	29	0.261
KLF14_2	7:130418311	101	8	0.079
ELOVL2	6:11044861	95	41	0.432
FHL2	2:106015739	81	19	0.235
KLF14	7:130419116	134	14	0.104
C1orf132	1:207997026	43	34	0.791
TRIM59	3:160167977	X	X	X
ASPA	17:3379567	161	91	0.565
PDE4C	19:18343915	240	69	0.288
ITGA2B	17:42467728	131	60	0.458
PDE4C_1	19:18343915	240	69	0.288
PDE4C_2	19:18343937	255	103	0.404
PDE4C_3	19:18343941	261	54	0.207
PDE4C_4	19:18343943	262	52	0.198
PDE4C_5	19:18344003	284	17	0.060
ITGA2B	17:42467780	139	82	0.590
ADAR_1	1:154582187	101	74	0.733
ADAR_2	1:154582288	137	96	0.701
ELOVL2	6:11044861	98	70	0.714
FHL2	2:106015739	91	34	0.374
KLF14	7:130419116	134	14	0.104
C1orf132	1:207997026	46	2	0.043
TRIM59	3:160167977	X	X	X
cg12837463	7:35300228	58	31	0.534
NOX4	11:89322851	304	176	0.579
TTC7B	14:91283606	225	203	0.902
ELOVL2_1	6:11044858	92	5	0.054
ELOVL2_2	6:11044861	95	41	0.432
ELOVL2_3	6:11044867	98	70	0.714
ELOVL2_4	6:11044873	107	61	0.570
ELOVL2_5	6:11044888	180	81	0.450
PDE4C	19:18343915	240	69	0.288
EDARADD	1:236557695	120	0	0.000
ELOVL2_CpG7	6:11044867	98	70	0.714
ELOVL2_CpG5	6:11044875	105	41	0.390
ASPA	17:3379567	161	91	0.565
ELOVL2	6:11044873	107	61	0.570
PDE4C	19:18343889	196	22	0.112
EDARADD	1:236557683	122	35	0.287

Sem3_2	Age: 43			
Marker	Genomic position	CpG coverage	Methylated coverage	Methylation level
BCAS4	20:49410865	344	76	0.221
BL_1	19:3344242	322	2	0.006
BL_2	19:3344251	335	3	0.009
VE_8	16:86398467	151	108	0.715
ZC3H12D	6:149778105	91	83	0.912
AHRR_1	5:373476	164	59	0.360
AHRR_2	5:373490	184	60	0.326
AHRR_3	5:373494	188	95	0.505
AHRR_4	5:373529	260	72	0.277
SCGN	6:25652606	304	24	0.079
KLF14_1	7:130418281	15	3	0.200
KLF14_2	7:130418311	16	1	0.063
ELOVL2	6:11044861	225	98	0.436
FHL2	2:106015739	134	26	0.194
KLF14	7:130419116	151	11	0.073
C1orf132	1:207997026	90	67	0.744
TRIM59	3:160167977	X	X	X
ASPA	17:3379567	15	9	0.600
PDE4C	19:18343915	474	107	0.226
ITGA2B	17:42467728	69	32	0.464
PDE4C_1	19:18343915	474	107	0.226
PDE4C_2	19:18343937	577	205	0.355
PDE4C_3	19:18343941	599	123	0.205
PDE4C_4	19:18343943	614	90	0.147
PDE4C_5	19:18344003	811	30	0.037
ITGA2B	17:42467780	88	48	0.545
ADAR_1	1:154582187	78	47	0.603
ADAR_2	1:154582288	171	139	0.813
ELOVL2	6:11044861	237	188	0.793
FHL2	2:106015739	137	35	0.255
KLF14	7:130419116	151	11	0.073
C1orf132	1:207997026	87	0	0.000
TRIM59	3:160167977	X	X	X
cg12837463	7:35300228	23	9	0.391
NOX4	11:89322851	477	296	0.621
TTC7B	14:91283606	360	326	0.906
ELOVL2_1	6:11044858	206	2	0.010
ELOVL2_2	6:11044861	225	98	0.436
ELOVL2_3	6:11044867	237	188	0.793
ELOVL2_4	6:11044873	254	153	0.602
ELOVL2_5	6:11044888	332	118	0.355
PDE4C	19:18343915	474	107	0.226
EDARADD	1:236557695	141	1	0.007
ELOVL2_CpG7	6:11044867	237	188	0.793
ELOVL2_CpG5	6:11044875	257	102	0.397
ASPA	17:3379567	15	9	0.600
ELOVL2	6:11044873	254	153	0.602
PDE4C	19:18343889	349	51	0.146
EDARADD	1:236557683	148	63	0.426

Sem4_1	Age: 32			
Marker	Genomic position	CpG coverage	Methylated coverage	Methylation level
BCAS4	20:49410865	67	22	0.328
BL_1	19:3344242	92	1	0.011
BL_2	19:3344251	96	1	0.010
VE_8	16:86398467	44	31	0.705
ZC3H12D	6:149778105	19	15	0.789
AHRR_1	5:373476	23	21	0.913
AHRR_2	5:373490	23	19	0.826
AHRR_3	5:373494	23	23	1.000
AHRR_4	5:373529	30	19	0.633
SCGN	6:25652606	94	0	0.000
KLF14_1	7:130418281	26	5	0.192
KLF14_2	7:130418311	22	2	0.091
ELOVL2	6:11044861	13	7	0.538
FHL2	2:106015739	31	1	0.032
KLF14	7:130419116	36	2	0.056
C1orf132	1:207997026	17	13	0.765
TRIM59	3:160167977	X	X	X
ASPA	17:3379567	82	16	0.195
PDE4C	19:18343915	75	40	0.533
ITGA2B	17:42467728	47	38	0.809
PDE4C_1	19:18343915	75	40	0.533
PDE4C_2	19:18343937	80	37	0.463
PDE4C_3	19:18343941	80	30	0.375
PDE4C_4	19:18343943	80	28	0.350
PDE4C_5	19:18344003	81	8	0.099
ITGA2B	17:42467780	55	47	0.855
ADAR_1	1:154582187	61	51	0.836
ADAR_2	1:154582288	100	91	0.910
ELOVL2	6:11044861	15	13	0.867
FHL2	2:106015739	30	3	0.100
KLF14	7:130419116	36	2	0.056
C1orf132	1:207997026	17	0	0.000
TRIM59	3:160167977	X	X	X
cg12837463	7:35300228	14	5	0.357
NOX4	11:89322851	164	41	0.250
TTC7B	14:91283606	103	83	0.806
ELOVL2_1	6:11044858	11	0	0.000
ELOVL2_2	6:11044861	13	7	0.538
ELOVL2_3	6:11044867	15	13	0.867
ELOVL2_4	6:11044873	16	14	0.875
ELOVL2_5	6:11044888	43	27	0.628
PDE4C	19:18343915	75	40	0.533
EDARADD	1:236557695	89	0	0.000
ELOVL2_CpG7	6:11044867	15	13	0.867
ELOVL2_CpG5	6:11044875	16	4	0.250
ASPA	17:3379567	82	16	0.195
ELOVL2	6:11044873	16	14	0.875
PDE4C	19:18343889	65	7	0.108
EDARADD	1:236557683	90	18	0.200

Sem4_2	Age: 32			
Marker	Genomic position	CpG coverage	Methylated coverage	Methylation level
BCAS4	20:49410865	200	79	0.395
BL_1	19:3344242	193	2	0.010
BL_2	19:3344251	192	3	0.016
VE_8	16:86398467	140	91	0.650
ZC3H12D	6:149778105	51	41	0.804
AHRR_1	5:373476	81	62	0.765
AHRR_2	5:373490	86	69	0.802
AHRR_3	5:373494	89	84	0.944
AHRR_4	5:373529	98	75	0.765
SCGN	6:25652606	217	14	0.065
KLF14_1	7:130418281	20	2	0.100
KLF14_2	7:130418311	18	1	0.056
ELOVL2	6:11044861	125	96	0.768
FHL2	2:106015739	107	16	0.150
KLF14	7:130419116	39	4	0.103
C1orf132	1:207997026	46	40	0.870
TRIM59	3:160167977	X	X	X
ASPA	17:3379567	41	11	0.268
PDE4C	19:18343915	221	98	0.443
ITGA2B	17:42467728	76	60	0.789
PDE4C_1	19:18343915	221	98	0.443
PDE4C_2	19:18343937	244	108	0.443
PDE4C_3	19:18343941	254	71	0.280
PDE4C_4	19:18343943	254	95	0.374
PDE4C_5	19:18344003	285	20	0.070
ITGA2B	17:42467780	91	66	0.725
ADAR_1	1:154582187	101	82	0.812
ADAR_2	1:154582288	131	111	0.847
ELOVL2	6:11044861	126	106	0.841
FHL2	2:106015739	111	13	0.117
KLF14	7:130419116	39	4	0.103
C1orf132	1:207997026	46	0	0.000
TRIM59	3:160167977	X	X	X
cg12837463	7:35300228	27	9	0.333
NOX4	11:89322851	319	85	0.266
TTC7B	14:91283606	140	110	0.786
ELOVL2_1	6:11044858	109	0	0.000
ELOVL2_2	6:11044861	125	96	0.768
ELOVL2_3	6:11044867	126	106	0.841
ELOVL2_4	6:11044873	126	111	0.881
ELOVL2_5	6:11044888	193	111	0.575
PDE4C	19:18343915	221	98	0.443
EDARADD	1:236557695	94	0	0.000
ELOVL2_CpG7	6:11044867	126	106	0.841
ELOVL2_CpG5	6:11044875	126	74	0.587
ASPA	17:3379567	41	11	0.268
ELOVL2	6:11044873	126	111	0.881
PDE4C	19:18343889	202	38	0.188
EDARADD	1:236557683	101	28	0.277

Sem5_2	Age: 26			
Marker	Genomic position	CpG coverage	Methylated coverage	Methylation level
BCAS4	20:49410865	357	238	0.667
BL_1	19:3344242	376	7	0.019
BL_2	19:3344251	391	7	0.018
VE_8	16:86398467	138	99	0.717
ZC3H12D	6:149778105	146	94	0.644
AHRR_1	5:373476	141	118	0.837
AHRR_2	5:373490	159	134	0.843
AHRR_3	5:373494	160	146	0.913
AHRR_4	5:373529	181	119	0.657
SCGN	6:25652606	308	15	0.049
KLF14_1	7:130418281	58	6	0.103
KLF14_2	7:130418311	66	5	0.076
ELOVL2	6:11044861	209	114	0.545
FHL2	2:106015739	263	5	0.019
KLF14	7:130419116	170	7	0.041
C1orf132	1:207997026	121	116	0.959
TRIM59	3:160167977	X	X	X
ASPA	17:3379567	88	11	0.125
PDE4C	19:18343915	524	151	0.288
ITGA2B	17:42467728	138	133	0.964
PDE4C_1	19:18343915	524	151	0.288
PDE4C_2	19:18343937	606	151	0.249
PDE4C_3	19:18343941	614	82	0.134
PDE4C_4	19:18343943	618	81	0.131
PDE4C_5	19:18344003	741	25	0.034
ITGA2B	17:42467780	161	146	0.907
ADAR_1	1:154582187	91	76	0.835
ADAR_2	1:154582288	161	151	0.938
ELOVL2	6:11044861	219	182	0.831
FHL2	2:106015739	255	6	0.024
KLF14	7:130419116	170	7	0.041
C1orf132	1:207997026	117	0	0.000
TRIM59	3:160167977	X	X	X
cg12837463	7:35300228	43	18	0.419
NOX4	11:89322851	490	88	0.180
TTC7B	14:91283606	305	250	0.820
ELOVL2_1	6:11044858	192	0	0.000
ELOVL2_2	6:11044861	209	114	0.545
ELOVL2_3	6:11044867	219	182	0.831
ELOVL2_4	6:11044873	231	160	0.693
ELOVL2_5	6:11044888	329	150	0.456
PDE4C	19:18343915	524	151	0.288
EDARADD	1:236557695	136	2	0.015
ELOVL2_CpG7	6:11044867	219	182	0.831
ELOVL2_CpG5	6:11044875	234	83	0.355
ASPA	17:3379567	88	11	0.125
ELOVL2	6:11044873	231	160	0.693
PDE4C	19:18343889	406	51	0.126
EDARADD	1:236557683	141	37	0.262

Sem6_1	Age: 30			
Marker	Genomic position	CpG coverage	Methylated coverage	Methylation level
BCAS4	20:49410865	267	145	0.543
BL_1	19:3344242	250	1	0.004
BL_2	19:3344251	249	0	0.000
VE_8	16:86398467	184	116	0.630
ZC3H12D	6:149778105	82	66	0.805
AHRR_1	5:373476	121	92	0.760
AHRR_2	5:373490	124	91	0.734
AHRR_3	5:373494	129	113	0.876
AHRR_4	5:373529	139	84	0.604
SCGN	6:25652606	282	21	0.074
KLF14_1	7:130418281	82	7	0.085
KLF14_2	7:130418311	65	8	0.123
ELOVL2	6:11044861	81	52	0.642
FHL2	2:106015739	107	16	0.150
KLF14	7:130419116	165	13	0.079
C1orf132	1:207997026	83	53	0.639
TRIM59	3:160167977	X	X	X
ASPA	17:3379567	284	74	0.261
PDE4C	19:18343915	252	75	0.298
ITGA2B	17:42467728	182	142	0.780
PDE4C_1	19:18343915	252	75	0.298
PDE4C_2	19:18343937	283	70	0.247
PDE4C_3	19:18343941	287	59	0.206
PDE4C_4	19:18343943	285	48	0.168
PDE4C_5	19:18344003	310	11	0.035
ITGA2B	17:42467780	192	156	0.813
ADAR_1	1:154582187	161	131	0.814
ADAR_2	1:154582288	262	223	0.851
ELOVL2	6:11044861	91	87	0.956
FHL2	2:106015739	113	19	0.168
KLF14	7:130419116	165	13	0.079
C1orf132	1:207997026	83	0	0.000
TRIM59	3:160167977	X	X	X
cg12837463	7:35300228	76	42	0.553
NOX4	11:89322851	468	138	0.295
TTC7B	14:91283606	281	218	0.776
ELOVL2_1	6:11044858	68	0	0.000
ELOVL2_2	6:11044861	81	52	0.642
ELOVL2_3	6:11044867	91	87	0.956
ELOVL2_4	6:11044873	94	86	0.915
ELOVL2_5	6:11044888	150	83	0.553
PDE4C	19:18343915	252	75	0.298
EDARADD	1:236557695	194	0	0.000
ELOVL2_CpG7	6:11044867	91	87	0.956
ELOVL2_CpG5	6:11044875	93	42	0.452
ASPA	17:3379567	284	74	0.261
ELOVL2	6:11044873	94	86	0.915
PDE4C	19:18343889	236	48	0.203
EDARADD	1:236557683	190	38	0.200

Sem6_2	Age: 30			
Marker	Genomic position	CpG coverage	Methylated coverage	Methylation level
BCAS4	20:49410865	402	212	0.527
BL_1	19:3344242	364	0	0.000
BL_2	19:3344251	380	2	0.005
VE_8	16:86398467	242	187	0.773
ZC3H12D	6:149778105	110	94	0.855
AHRR_1	5:373476	181	140	0.773
AHRR_2	5:373490	211	162	0.768
AHRR_3	5:373494	213	177	0.831
AHRR_4	5:373529	235	155	0.660
SCGN	6:25652606	349	25	0.072
KLF14_1	7:130418281	145	19	0.131
KLF14_2	7:130418311	116	11	0.095
ELOVL2	6:11044861	166	101	0.608
FHL2	2:106015739	218	14	0.064
KLF14	7:130419116	157	12	0.076
C1orf132	1:207997026	88	76	0.864
TRIM59	3:160167977	X	X	X
ASPA	17:3379567	344	48	0.140
PDE4C	19:18343915	373	91	0.244
ITGA2B	17:42467728	241	189	0.784
PDE4C_1	19:18343915	373	91	0.244
PDE4C_2	19:18343937	429	147	0.343
PDE4C_3	19:18343941	433	91	0.210
PDE4C_4	19:18343943	433	107	0.247
PDE4C_5	19:18344003	485	12	0.025
ITGA2B	17:42467780	267	208	0.779
ADAR_1	1:154582187	249	180	0.723
ADAR_2	1:154582288	322	264	0.820
ELOVL2	6:11044861	179	154	0.860
FHL2	2:106015739	215	22	0.102
KLF14	7:130419116	157	12	0.076
C1orf132	1:207997026	86	0	0.000
TRIM59	3:160167977	X	X	X
cg12837463	7:35300228	114	43	0.377
NOX4	11:89322851	612	142	0.232
TTC7B	14:91283606	382	315	0.825
ELOVL2_1	6:11044858	152	0	0.000
ELOVL2_2	6:11044861	166	101	0.608
ELOVL2_3	6:11044867	179	154	0.860
ELOVL2_4	6:11044873	188	145	0.771
ELOVL2_5	6:11044888	292	155	0.531
PDE4C	19:18343915	373	91	0.244
EDARADD	1:236557695	208	0	0.000
ELOVL2_CpG7	6:11044867	179	154	0.860
ELOVL2_CpG5	6:11044875	190	102	0.537
ASPA	17:3379567	344	48	0.140
ELOVL2	6:11044873	188	145	0.771
PDE4C	19:18343889	299	55	0.184
EDARADD	1:236557683	210	47	0.224

Sem7_2	Age: 38			
Marker	Genomic position	CpG coverage	Methylated coverage	Methylation level
BCAS4	20:49410865	231	134	0.580
BL_1	19:3344242	279	4	0.014
BL_2	19:3344251	287	3	0.010
VE_8	16:86398467	162	114	0.704
ZC3H12D	6:149778105	96	68	0.708
AHRR_1	5:373476	153	126	0.824
AHRR_2	5:373490	180	147	0.817
AHRR_3	5:373494	178	164	0.921
AHRR_4	5:373529	202	104	0.515
SCGN	6:25652606	294	4	0.014
KLF14_1	7:130418281	136	11	0.081
KLF14_2	7:130418311	111	5	0.045
ELOVL2	6:11044861	168	55	0.327
FHL2	2:106015739	130	9	0.069
KLF14	7:130419116	159	3	0.019
C1orf132	1:207997026	69	62	0.899
TRIM59	3:160167977	X	X	X
ASPA	17:3379567	186	88	0.473
PDE4C	19:18343915	272	31	0.114
ITGA2B	17:42467728	184	159	0.864
PDE4C_1	19:18343915	272	31	0.114
PDE4C_2	19:18343937	302	54	0.179
PDE4C_3	19:18343941	303	14	0.046
PDE4C_4	19:18343943	305	16	0.052
PDE4C_5	19:18344003	331	4	0.012
ITGA2B	17:42467780	201	169	0.841
ADAR_1	1:154582187	155	115	0.742
ADAR_2	1:154582288	214	186	0.869
ELOVL2	6:11044861	182	105	0.577
FHL2	2:106015739	127	7	0.055
KLF14	7:130419116	159	3	0.019
C1orf132	1:207997026	72	2	0.028
TRIM59	3:160167977	X	X	X
cg12837463	7:35300228	58	15	0.259
NOX4	11:89322851	404	112	0.277
TTC7B	14:91283606	295	257	0.871
ELOVL2_1	6:11044858	159	3	0.019
ELOVL2_2	6:11044861	168	55	0.327
ELOVL2_3	6:11044867	182	105	0.577
ELOVL2_4	6:11044873	183	84	0.459
ELOVL2_5	6:11044888	255	62	0.243
PDE4C	19:18343915	272	31	0.114
EDARADD	1:236557695	187	1	0.005
ELOVL2_CpG7	6:11044867	182	105	0.577
ELOVL2_CpG5	6:11044875	184	31	0.168
ASPA	17:3379567	186	88	0.473
ELOVL2	6:11044873	183	84	0.459
PDE4C	19:18343889	242	27	0.112
EDARADD	1:236557683	189	108	0.571

Sem8_1	Age: 20			
Marker	Genomic position	CpG coverage	Methylated coverage	Methylation level
BCAS4	20:49410865	162	100	0.617
BL_1	19:3344242	179	0	0.000
BL_2	19:3344251	186	1	0.005
VE_8	16:86398467	94	49	0.521
ZC3H12D	6:149778105	46	43	0.935
AHRR_1	5:373476	84	53	0.631
AHRR_2	5:373490	88	63	0.716
AHRR_3	5:373494	85	70	0.824
AHRR_4	5:373529	97	62	0.639
SCGN	6:25652606	158	14	0.089
KLF14_1	7:130418281	44	4	0.091
KLF14_2	7:130418311	37	2	0.054
ELOVL2	6:11044861	104	79	0.760
FHL2	2:106015739	97	12	0.124
KLF14	7:130419116	95	16	0.168
C1orf132	1:207997026	31	25	0.806
TRIM59	3:160167977	X	X	X
ASPA	17:3379567	130	12	0.092
PDE4C	19:18343915	196	73	0.372
ITGA2B	17:42467728	84	63	0.750
PDE4C_1	19:18343915	196	73	0.372
PDE4C_2	19:18343937	200	119	0.595
PDE4C_3	19:18343941	201	60	0.299
PDE4C_4	19:18343943	201	52	0.259
PDE4C_5	19:18344003	221	33	0.149
ITGA2B	17:42467780	87	67	0.770
ADAR_1	1:154582187	91	69	0.758
ADAR_2	1:154582288	130	109	0.838
ELOVL2	6:11044861	104	97	0.933
FHL2	2:106015739	91	14	0.154
KLF14	7:130419116	95	16	0.168
C1orf132	1:207997026	28	0	0.000
TRIM59	3:160167977	X	X	X
cg12837463	7:35300228	26	11	0.423
NOX4	11:89322851	223	24	0.108
TTC7B	14:91283606	159	135	0.849
ELOVL2_1	6:11044858	100	0	0.000
ELOVL2_2	6:11044861	104	79	0.760
ELOVL2_3	6:11044867	104	97	0.933
ELOVL2_4	6:11044873	103	99	0.961
ELOVL2_5	6:11044888	142	86	0.606
PDE4C	19:18343915	196	73	0.372
EDARADD	1:236557695	109	0	0.000
ELOVL2_CpG7	6:11044867	104	97	0.933
ELOVL2_CpG5	6:11044875	98	52	0.531
ASPA	17:3379567	130	12	0.092
ELOVL2	6:11044873	103	99	0.961
PDE4C	19:18343889	186	35	0.188
EDARADD	1:236557683	117	23	0.197

Sem8_2	Age: 20			
Marker	Genomic position	CpG coverage	Methylated coverage	Methylation level
BCAS4	20:49410865	188	95	0.505
BL_1	19:3344242	180	0	0.000
BL_2	19:3344251	189	2	0.011
VE_8	16:86398467	111	83	0.748
ZC3H12D	6:149778105	23	22	0.957
AHRR_1	5:373476	86	57	0.663
AHRR_2	5:373490	104	74	0.712
AHRR_3	5:373494	114	95	0.833
AHRR_4	5:373529	134	92	0.687
SCGN	6:25652606	224	24	0.107
KLF14_1	7:130418281	13	0	0.000
KLF14_2	7:130418311	14	0	0.000
ELOVL2	6:11044861	160	129	0.806
FHL2	2:106015739	149	16	0.107
KLF14	7:130419116	95	4	0.042
C1orf132	1:207997026	41	37	0.902
TRIM59	3:160167977	X	X	X
ASPA	17:3379567	9	2	0.222
PDE4C	19:18343915	373	232	0.622
ITGA2B	17:42467728	56	44	0.786
PDE4C_1	19:18343915	373	232	0.622
PDE4C_2	19:18343937	403	225	0.558
PDE4C_3	19:18343941	410	108	0.263
PDE4C_4	19:18343943	406	121	0.298
PDE4C_5	19:18344003	458	22	0.048
ITGA2B	17:42467780	87	48	0.552
ADAR_1	1:154582187	67	40	0.597
ADAR_2	1:154582288	104	94	0.904
ELOVL2	6:11044861	169	155	0.917
FHL2	2:106015739	141	9	0.064
KLF14	7:130419116	95	4	0.042
C1orf132	1:207997026	41	1	0.024
TRIM59	3:160167977	X	X	X
cg12837463	7:35300228	11	3	0.273
NOX4	11:89322851	309	44	0.142
TTC7B	14:91283606	147	113	0.769
ELOVL2_1	6:11044858	118	0	0.000
ELOVL2_2	6:11044861	160	129	0.806
ELOVL2_3	6:11044867	169	155	0.917
ELOVL2_4	6:11044873	171	133	0.778
ELOVL2_5	6:11044888	253	154	0.609
PDE4C	19:18343915	373	232	0.622
EDARADD	1:236557695	84	0	0.000
ELOVL2_CpG7	6:11044867	169	155	0.917
ELOVL2_CpG5	6:11044875	172	129	0.750
ASPA	17:3379567	9	2	0.222
ELOVL2	6:11044873	171	133	0.778
PDE4C	19:18343889	338	116	0.343
EDARADD	1:236557683	85	24	0.282

MethControl	Age: --			
Marker	Genomic position	CpG coverage	Methylated coverage	Methylation level
BCAS4	20:49410865	45	45	1.000
BL_1	19:3344242	1	0	0.000
BL_2	19:3344251	1	0	0.000
VE_8	16:86398467	8	6	0.750
ZC3H12D	6:149778105	25	25	1.000
AHRR_1	5:373476	X	X	X
AHRR_2	5:373490	X	X	X
AHRR_3	5:373494	X	X	X
AHRR_4	5:373529	X	X	X
SCGN	6:25652606	72	70	0.972
KLF14_1	7:130418281	1	1	1.000
KLF14_2	7:130418311	1	1	1.000
ELOVL2	6:11044861	10	10	1.000
FHL2	2:106015739	1	0	0.000
KLF14	7:130419116	23	22	0.957
C1orf132	1:207997026	1	1	1.000
TRIM59	3:160167977	X	X	X
ASPA	17:3379567	36	28	0.778
PDE4C	19:18343915	1	0	0.000
ITGA2B	17:42467728	11	2	0.182
PDE4C_1	19:18343915	1	0	0.000
PDE4C_2	19:18343937	1	1	1.000
PDE4C_3	19:18343941	1	0	0.000
PDE4C_4	19:18343943	1	0	0.000
PDE4C_5	19:18344003	1	1	1.000
ITGA2B	17:42467780	12	2	0.167
ADAR_1	1:154582187	25	25	1.000
ADAR_2	1:154582288	29	24	0.828
ELOVL2	6:11044861	10	8	0.800
FHL2	2:106015739	1	0	0.000
KLF14	7:130419116	23	22	0.957
C1orf132	1:207997026	1	0	0.000
TRIM59	3:160167977	X	X	X
cg12837463	7:35300228	5	5	1.000
NOX4	11:89322851	172	160	0.930
TTC7B	14:91283606	4	4	1.000
ELOVL2_1	6:11044858	9	0	0.000
ELOVL2_2	6:11044861	10	10	1.000
ELOVL2_3	6:11044867	10	8	0.800
ELOVL2_4	6:11044873	10	8	0.800
ELOVL2_5	6:11044888	14	12	0.857
PDE4C	19:18343915	1	0	0.000
EDARADD	1:236557695	27	0	0.000
ELOVL2_CpG7	6:11044867	10	8	0.800
ELOVL2_CpG5	6:11044875	12	10	0.833
ASPA	17:3379567	36	28	0.778
ELOVL2	6:11044873	10	8	0.800
PDE4C	19:18343889	27	26	0.963
EDARADD	1:236557683	X	X	X

Unmeth Cont	Age: --			
Marker	Genomic position	CpG coverage	Methylated coverage	Methylation level
BCAS4	20:49410865	12	1	0.083
BL_1	19:3344242	1	0	0.000
BL_2	19:3344251	1	0	0.000
VE_8	16:86398467	15	0	0.000
ZC3H12D	6:149778105	4	0	0.000
AHRR_1	5:373476	5	2	0.400
AHRR_2	5:373490	5	1	0.200
AHRR_3	5:373494	5	2	0.400
AHRR_4	5:373529	5	2	0.400
SCGN	6:25652606	94	5	0.053
KLF14_1	7:130418281	3	0	0.000
KLF14_2	7:130418311	2	0	0.000
ELOVL2	6:11044861	16	2	0.125
FHL2	2:106015739	9	0	0.000
KLF14	7:130419116	9	0	0.000
C1orf132	1:207997026	2	1	0.500
TRIM59	3:160167977	X	X	X
ASPA	17:3379567	37	1	0.027
PDE4C	19:18343915	8	0	0.000
ITGA2B	17:42467728	X	X	X
PDE4C_1	19:18343915	8	0	0.000
PDE4C_2	19:18343937	8	0	0.000
PDE4C_3	19:18343941	8	0	0.000
PDE4C_4	19:18343943	8	0	0.000
PDE4C_5	19:18344003	8	0	0.000
ITGA2B	17:42467780	X	X	X
ADAR_1	1:154582187	39	0	0.000
ADAR_2	1:154582288	50	0	0.000
ELOVL2	6:11044861	17	4	0.235
FHL2	2:106015739	9	0	0.000
KLF14	7:130419116	9	0	0.000
C1orf132	1:207997026	2	0	0.000
TRIM59	3:160167977	X	X	X
cg12837463	7:35300228	2	0	0.000
NOX4	11:89322851	153	6	0.039
TTC7B	14:91283606	8	0	0.000
ELOVL2_1	6:11044858	10	1	0.100
ELOVL2_2	6:11044861	16	2	0.125
ELOVL2_3	6:11044867	17	4	0.235
ELOVL2_4	6:11044873	16	3	0.188
ELOVL2_5	6:11044888	18	4	0.222
PDE4C	19:18343915	8	0	0.000
EDARADD	1:236557695	19	0	0.000
ELOVL2_CpG7	6:11044867	17	4	0.235
ELOVL2_CpG5	6:11044875	16	3	0.188
ASPA	17:3379567	37	1	0.027
ELOVL2	6:11044873	16	3	0.188
PDE4C	19:18343889	8	0	0.000
EDARADD	1:236557683	19	0	0.000

VITA

QUENTIN GAUTHIER

Born, Wilmington, Delaware

2009-2013	B.S., Forensic and Investigative Sciences West Virginia University Morgantown, West Virginia
2012	Intern, Medico-Legal Death Investigations Dauphin County Coroner's Office Harrisbug, Pennsylvania
2013-2015	M.S.F.S., Forensic Sciences Marshall University Huntington, West Virginia
2014	Intern, Emerging Technologies Section Armed Forces DNA Identification Laboratory Dover, Delaware
2015 -2020	Doctoral Candidate Florida International University Miami, Florida

PUBLICATIONS AND PRESENTATIONS

Lee JE, Lee JM, Neubaur J, Naue J, Mills C, Cao Y, Pospiech E, Pisarek A, Vidaki A, Kalamara V, Fleckhaus J, Freire A, Conde A, Oh YN, Wang Z, Gauthier Q, Fernandez Tejero N, Phillips C, Schneider P, Hou Y, McCord B, Branicki W, Podini D, Haas C, Lee JY, Lee HY. A collaborate exercise on DNA methylation-based body fluid typing and age prediction. (2020) Manuscript in Progress.

Antunes J, Gauthier Q, Aguiar-Pulido V, Duncan G, McCord, B. A data-driven, high-throughput methodology to determine tissue-specific differentially methylated regions able to discriminate body fluids. *Electrophoresis*. (2020) In Review.

Quentin Gauthier, Bruce McCord. Latent Profile Analysis of DNA Methylation Markers for Body Fluid Identification. PittCon 2020, Chicago, IL March 2020

McCord B, Gauthier Q, Alghanim H, Antunes J, Fernandez Tejero N, Duncan G, Balamurugan K. Applications of epigenetic methylation in body fluid identification, age determination and phenotyping. *Forensic Sci. Int. Genet. Supp. Ser.* (2019) 1, 485-487.

Joana Antunes, Quentin Gauthier, George Duncan, Bruce McCord. Discovery of new loci of interest for body fluid identification through DNA methylation melt analysis. California Association of Criminalistics 2019 Meeting, Ontario, CA October 2019

Gauthier QT, Cho S, Carmel JH, McCord BR. Development of a Body Fluid Identification Multiplex via DNA Methylation Analysis. *Electrophoresis*. (2019) 40, 18-19, 2565-2574.

McCord B, Gauthier Q, Cho S, Roig M, Gibson-Daw G, Young B, Taglia F, Zapico S, Fogliatto Mariot R, Lee S, Duncan G. Forensic DNA Analysis. *Anal. Chem.* (2019) 91, 1, 673-688.

Gauthier, Q., McCord, B. Creating a DNA Methylation Multiplex Assay: DNA Methylation Analysis for Body Fluid, Age, and other Lifestyle Traits. QIAGEN Workshop at International Forensic Research Institute annual symposium, Miami, FL May 2019

Quentin Gauthier, Sohee Cho, Bruce McCord. Developmental Validation of a Body Fluid Identification Multiplex via DNA Methylation Analysis. PittCon 2019, Philadelphia, PA March 2019

Quentin Gauthier, Sohee Cho, Bruce McCord. Developmental Validation of a Body Fluid Identification Multiplex via DNA Methylation Analysis. 71st Annual Meeting of the American Academy of Forensic Sciences, Baltimore, MD February 2019

Gauthier, Q., McCord, B. Body Fluid Identification via Multiplex DNA Methylation Analysis. International Forensic Research Institute annual symposium, Miami, FL May 2018

Gauthier, Q., Antunes, J., McCord B. Identification of Body Fluid Using Multiplex Polymerase Chain Reaction (PCR). 70th Annual Meeting of the American Academy of Forensic Sciences, Seattle, WA February 2018

Quentin Gauthier, Joana Antunes, Vanessa Aguiar-Pulido, Kuppareddi Balamurugan, George Duncan, Giri Narasimhan and Bruce McCord. High-resolution melt analysis of DNA methylation patterns can discriminate body fluid of origin in crime scene samples. PittCon 2018, Orlando, FL February 2018.