2020

Investigating stutter characteristics via isoalleles in massively parallel sequencing of a family pedigree

BOSTON UNIVERSITY

SCHOOL OF MEDICINE

Thesis

INVESTIGATING STUTTER CHARACTERISTICS VIA ISOALLELES IN MASSIVELY PARALLEL SEQUENCING OF A FAMILY PEDIGREE

by

PING YI WU

B.A., University of California, Berkeley, 2015

Submitted in partial fulfillment of the

requirements for the degree of

Master of Science

2020

Approved by

First Reader

Amy N. Brodeur, M.F.S.
Associate Director, Program in Biomedical Forensic Sciences
Assistant Professor, Department of Anatomy & Neurobiology

Second Reader

Robin W. Cotton, Ph.D.
Director, Program in Biomedical Forensic Sciences
Associate Professor, Department of Anatomy & Neurobiology

Third Reader

Fabio Oldoni, Ph.D.
Assistant Professor in Forensic Sciences
Arcadia University

ACKNOWLEDGMENTS

First and foremost, I would like to extend my appreciation to the amazing faculty of the Boston University Biomedical Forensic Sciences program for all their support during these difficult, yet rewarding, two years and their efforts in creating an unforgettable experience. I would especially like to express my gratitude for Professor Amy Brodeur and Dr. Robin Cotton's inexhaustible patience, encouragement and insights throughout the entire process. It was an honor to have the mentorship of two of Massachusetts's finest in the forensic DNA/Biology field, and the completion of this thesis would not have been possible without their invaluable guidance and flexibility during the COVID-19 pandemic.

Many thanks to the Verogen team for their expertise, thorough explanations, and efficient technical assistance. Their consideration for project deadlines and concerns went above and beyond the usual customer service, and we appreciate all their efforts in rectifying any issues we encountered throughout the process. I also want to thank alumni Andre Porto for donating his valuable time, resources, and experience to train the new "NGS team" on the technology.

To my parents and sisters, whom I love and cherish every day, motivating me on the opposite ends of the States – thank you for giving me nothing but strength, joy, and inspiration to pursue my dreams and ascending to heights I thought impossible. Warm musings of friends and family have helped me endure the brutal Boston winters and dissipated feelings of apprehension in an unfamiliar place.

INVESTIGATING STUTTER CHARACTERISTICS VIA ISOALLELES IN MASSIVELY PARALLEL SEQUENCING OF A FAMILY PEDIGREE

PING YI WU

ABSTRACT

Despite the prevalent utilization of capillary electrophoresis (CE) in the analysis of short tandem repeats (STRs) to generate deoxyribonucleic acid (DNA) profiles for forensic comparisons, the method is not without its inherent drawbacks. Massively parallel sequencing (MPS) is still a relatively novel technology in the forensics field, but contains the capacity to address current challenges faced by the traditional CE approach - such as degraded samples, low template DNA, and artifacts - while also providing additional information such as isoalleles, same-length alleles with sequence variation, and ancestry, mixture, and phenotyping-informative single nucleotide polymorphisms (SNPs).

One of the principal ongoing challenges faced by both technologies is the presence of artifacts such as stutter, a byproduct of slipped strand mispairing during amplification of STRs, which can further complicate interpretation of DNA profiles. Understanding and predicting the behavior of stutter is important in establishing appropriate thresholds to distinguish these artifacts from true alleles. With complex MPS data, new approaches in characterizing stutter have been established such as the BLMM and simplified sequence.

In this study, twenty-one oral samples from individuals belonging to the same family were constructed into libraries containing 58 STR regions and 98 identity SNPs using Verogen's Forenseq™ DNA Signature Prep Kit and sequenced on the MiSeq FGx™

Forensics Genomics System. Isoallele and stutter sequences were extracted from the data and simplified using the longest uninterrupted stretch (LUS), block length of missing motif (BLMM) and simplified sequence approaches. It was found that the stutter ratio for the 11 isoallele pairs at the D13S317 locus did not follow the linear correlation with increasing LUS. Instead, the isoallele with the higher LUS demonstrated equal or lower stutter ratios. Additionally, three different stutter patterns were identified for the same locus. Based on the provided pedigree, ten different relations were defined and the amount of allele sharing between the individuals in the pedigree was analyzed with and in the absence of isoallelic information to determine its impact on predicting relatedness. It was found that there was an average of 1.3% difference across the ten defined categories when isoalleles were taken into consideration. However, the difference in the percentage of shared alleles was not found to be significant for each of the relations category between the results before and after the consideration of isoallelelic data.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

AT	Analytical Threshold
BLMM	Block length of missing motif
BP	Base Pairs
CE	Capillary Electrophoresis
CODIS	Combined DNA Index System
DNA	Deoxyribonucleic Acid
DPMA	DNA Primer Mix A
FBI	Federal Bureau of Investigation
FEM	Forenseq™ Enzyme Mix
HSC	Human Sequencing Control
IT	Interpretation Threshold
LNB1	Library Normalization Beads 1
LNW1	Library Normalization Wash 1
LOD	Limit of Detection
LUS	Longest Uninterrupted Stretch
µL	Microliter(s)
MPS	Massively-Parallel Sequencing
NDIS	National DNA Index System
ng	Nanograms
NaOH	Sodium Hydroxide
NGS	Next-Generation Sequencing

PCR	Polymerase Chain Reaction
pg	Picograms
pM	Picomolar
RNA	Ribonucleic Acid
SNP	Single Nucleotide Polymorphism
STR	Short Tandem Repeat
TE	Tris-EDTA
UAS	Forenseq™ Universal Analysis Software

1. INTRODUCTION

1.1 STR Analysis

Although much of the human genome is identical across individuals, deoxyribonucleic acid (DNA) fingerprinting in forensic analyses and paternity testing relies on distinguishing the genetic variations, or polymorphisms, that exist from one person to another in order to establish linkage between a person and a criminal investigation [1]. Thus, regions containing higher mutation rates, such as repeating DNA motifs, or microsatellites, became areas of interest within the discipline due to their variability among individuals and compatibility with PCR-based methods. More specifically, the microsatellites investigated by forensic scientists are non-coding regions of repeated units typically two to five nucleotides in length known as short tandem repeats (STRs) [1-2]. The number of repeat units in the entire length of an STR comprise each allele, and can vary across individuals and loci. An individual inherits one of each pair of these alleles from their biological mother and father, and can be either heterozygous or homozygous at a specific locus. For instance, a mother and father with a 11,12 and 12,13 allele, respectively, at a given STR locus have the possibility of a child with a heterozygous genotype of 11,12, 11,13, or 12,13; or the child may have a homozygous genotype of 12,12.

STR analysis using capillary electrophoresis (CE) has been the method of choice for the interpretation of DNA evidence by forensic practitioners for more than 25 years. An individual's STR profile is essentially a DNA fingerprint consisting of multiple pairs of alleles that allows for the discrimination of individuals, making it an important tool for establishing a person's association with a specific case. This method involves the

amplification of extracted DNA, via the polymerase chain reaction (PCR), of the defined microsatellite repeat units with sequence-specific primers containing fluorescent dyes. The different dye colors are utilized in order to expand the number of loci that can be analyzed in a single reaction. After PCR, an internal DNA dye standard is added to the resulting DNA fragments before the samples are placed in a CE instrument. The standard is designed to determine the length of the unknown DNA fragments as they are separated in a polymer sieve based on their relative lengths. As the fragments reach the end of the gel, the assorted dyes are detected using a laser to produce an electropherogram, which can then be analyzed using genotyping software such as GeneMapper™ [1-3].

After the STR profile of an evidentiary sample is analyzed, it is then compared to a suspect's STR profile to determine whether the individual is included or excluded as a possible DNA contributor. An exclusion occurs when the STR alleles fail to match the compared profiles. If all investigated alleles match the profiles, the frequency of observing the same genotype in a population is calculated in order to establish statistical weight of the DNA evidence. STR profiles are compared with the Federal Bureau of Investigation's (FBI) Combined DNA Index System (CODIS) database. DNA profiles, consisting of 20 core loci, are stored within CODIS and can be matched to previous offenders as well as new and ongoing casework. The CODIS program is an attestation to the reliability and endurance of the STR method. [4-5].

Despite its long history, STR analysis using CE data is not without its drawbacks. Case samples are often heavily degraded or below the limit of detection (LOD) for CE methods [6-7]. Mixture interpretation can become difficult with increasing numbers of

contributors, while low template samples may result in significant loss of alleles in the ensuing STR profile [8-9]. Although new technologies and reagent kits have increased the level of sensitivity for low template DNA, including touch DNA, this can also introduce minor contributors with no connection to the crime through secondary or tertiary transfer [9]. Commonly-used STR kits such as the Globalfiler™ PCR Amplification Kit (Thermo Fisher Scientific, Waltham, MA) are also limited by the number of dyes and primer pairs to target additional STR regions - leading to the addition of workflows, such as gender-specific primer sets, to obtain supplementary information [11]. PCR-based artifacts, such as stutter and non-template additions, and instrument-based artifacts such as pull-up and spikes, can also arise from CE-based methods and further complicate the interpretation of STR profiles [12].

1.2 Stutter

One of the most common and highly-characterized artifacts studied in forensic STR analysis is stutter, defined as nonspecific byproducts of STR fragments created during PCR amplification. These artifacts are generally one repeat more (N+1) or, more commonly, one repeat less (N-1) than the true allele. Stutters with higher degrees of repeat variation such as two, three, or four additional repeats (N±2, N ±3, N ±4) have also been characterized, but at a lower intensity [12-15].

The analytical threshold (AT), a value that distinguishes interpretable DNA from background noise and artifacts, is important to establish empirically for allele detection and also to prevent over-interpretation [16]. The occurrence of stutter has been shown to be

more pronounced in smaller repeat units such as dinucleotides versus tetranucleotides, and the rate of producing stutter increases as the total STR length increases. It has also been found that the frequency of stutter can vary between different STR loci and alleles [17-19]. Thus, it is crucial for labs to determine stutter rates to assist in distinguishing these artifacts from true alleles.

Despite the extensive knowledge surrounding the interpretation and prediction of stutter occurrences for different alleles and conditions, very little is known about the specific mechanism of its formation. The common theory explaining stutter formation is DNA slippage or slipped strand mispairing, which is when a repeat unit of the template strand "slips" or loops out during amplification, causing the polymerase to pass the repeat and result in a N-1 product. A N+1 stutter artifact is posited to occur when the strand being extended experiences slippage instead of the template, causing the polymerase to copy the same repeat twice [20]. These stutter artifacts are usually identified as shorter peaks differing by one repeat from the true allele (Figure 1). Stutter interpretation can become challenging when these artifacts fall within the same repeat number as true alleles, or alleles that belong to the DNA contributor(s). The issue is compounded when there are multiple minor contributors with peaks that fall within the same peak height ratios as reported stutter values - leading to lower confidence in calling true peaks due to possible misinterpretation and confusion with stutter and vice versa [20-22].

Various efforts have been made to address slipped strand mispairing and reduce the occurrence of stutter such as directly altering PCR conditions. For instance, lowering the temperature of the extension step with a thermolabile polymerase is posited to decrease

DNA dissociation. However, because this method requires the addition of new polymerase after each denaturation step, it becomes impractical and time-consuming. A study found that isothermal amplification of DNA at temperatures between 37 and 42°C using recombinase polymerase amplification (RPA) reduced stutter ratios by a range of 21% to 67% [15]. Slipped strand mispairing has also been shown to be reduced in mononucleotide repeats by utilizing proofreading, fusion-based DNA polymerases such as Phusion (Finnzymes, Espoo, Finland) and Herculase II Fusion (Agilent, Santa Clara, CA), but has not been confirmed for other types of repeats. [21].

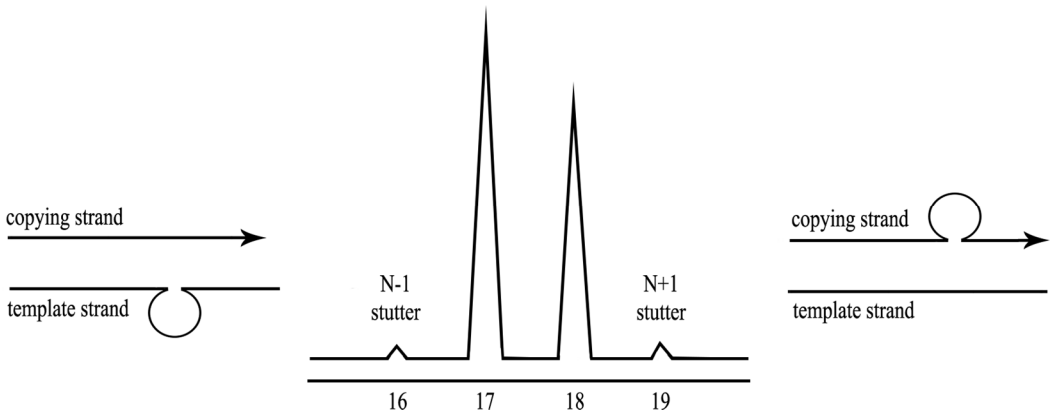

Figure 1. Slipped-strand mispairing. During replication, either the template or copying strand may dissociate or "loop out," causing DNA polymerase to misalign and create shorter or longer copies known as stutter.

1.3 Massively Parallel Sequencing (MPS)

1.3.1 MPS Beginnings

Prior to the advent of massively parallel sequencing (MPS) - also known as next-generation sequencing (NGS), deep, or second-generation sequencing - Sanger sequencing, or first-generation sequencing, dominated the genomics field from the late 1970s by

utilizing random dideoxynucleotide termination of PCR-amplified DNA fragments at the 3' end. This results in multiple DNA fragments from an identical sequence that have been truncated at different points along the entire DNA length. These DNA fragments are then injected and separated by molecular weight via CE, and the final output are the base calls (A, C, G, T) of the entire DNA sequence, which are depicted in an electropherogram. Since then, many efforts were made to automate the Sanger method, which ultimately led to the completion of the Human Genome Project [23]. However, first generation sequencing was still limited and unfeasible for sequencing whole human genomes at a rapid pace, and would be impractical for routine STR analysis due to its inability to multiplex primer pairs to target different regions of interest in the same reaction. Thus, efforts were made to develop more efficient, high-throughput, and accurate methods of sequencing.

Although there are many approaches to NGS, the term encompasses DNA sequencing that has evolved a few decades after the conventional Sanger method and is capable of sequencing millions of targeted DNA fragments in parallel. This is possible by ligating or amplifying adapter sequences to DNA fragments, which can then be directly amplified on a solid surface. Just as the CE enables base calls via a laser and camera to monitor fluorescence in Sanger sequencing, NGS also relies on the detection of fluorescent signals or changes in electrical current. Sequencing via second-generation sequencing generally involves gigabytes of data that are analyzed via bioinformatic pipelines.

The high-throughput and rapid turnover of NGS makes it possible to sequence the entire human genome in a matter of days in contrast to a decade with Sanger sequencing. Additionally, NGS offers high sensitivity, accurate detection of mutations occurring at low

levels and detection of various types of mutations such as small insertions/deletions and single nucleotide polymorphisms (SNPs), revolutionizing clinical studies and assisting in preliminary and individualized diagnostics [24].

At the turn of the 21st century, many companies made strides in different NGS technologies with distinct sequencing chemistries such as Roche's 454, Illumina's HiSeq platform, and Life Technologies' Ion Torrent. Roche's 454 sequencing system was one of the first methods to successfully enter the NGS field, and was originally developed by Life Sciences. The technology utilizes pyrosequencing, which involves the detection of luminescence arising from released pyrophosphates after nucleotide incorporation. The Illumina HiSeq platform, alternatively, relies on reversible fluorescently-labelled terminators and cluster generation of amplified DNA. Many varieties of the HiSeq have been produced, such as the MiSeq and NextSeq, to accommodate different project and industry scales. The Ion Torrent differs from both the 454 and HiSeq by detecting nucleotides based on changes in the pH in relation to the surrounding solution with electronic sensors, as opposed to optical detection of fluorescence or luminescence [23-27]. The proliferation of these NGS methods revolutionized sequencing methods by providing a high-throughput technology that allows the multiplexing of different samples in a single reaction, substantially increasing the amount of obtainable data and drastically reducing the overall sequencing costs per sample – making it the preferred method for large-scale genomic projects [28].

Although a diverse breadth of scientific disciplines has made strides in employing MPS since its inception, limited resources, strict guidelines, instrument expenses and the

time necessary for training and validation have made it difficult for crime labs to implement this technology in their workflow. Additionally, the large data output often requires highly-technical bioinformatics expertise for efficient analysis. Hence, the adoption of MPS and the admissibility of the technology in courts have not been widely used and accepted until more recently [23-29].

1.3.2 Verogen's Approach to Forensic MPS

Verogen's MiSeq FGx® Forensic Genomics Solution was originally released as the MiSeq™ Forensic Genomics Solution by Illumina© and was the first NGS method to be fully validated and approved by the FBI's National DNA Index System (NDIS) for DNA profiling in May 2019 [30]. Similar to standard NGS workflows, the MiSeq FGx® Forensic Genomics Solution involves DNA extraction, quantification, library preparation, normalization, and library-pooling. The prepared libraries are then sequenced on the MiSeq FGx® Forensic Genomics System and analyzed with the accompanying Forenseq Universal Analysis Software (UAS) (Figure 2). The reagents and consumables are available in kits such as the ForenSeq DNA Signature Prep kit or the ForenSeq mtDNA Whole Genome kit - specifically for mitochondrial genome analysis. [31-33].

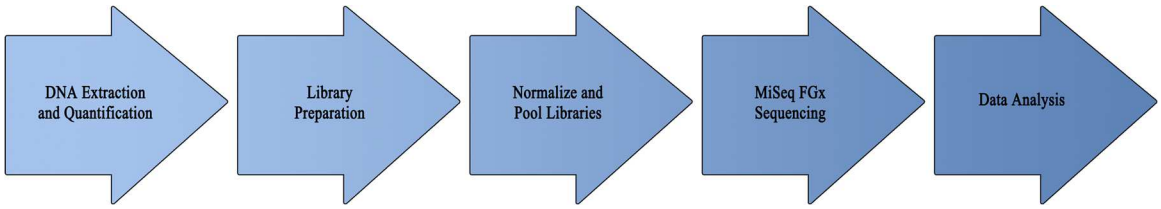

Figure 2. Schematic of MPS workflow. After DNA samples are extracted and quantified, target regions are amplified and tagged with adapter and index sequences. These completed libraries are then normalized and combined, or pooled, into a single reaction for sequencing. The generated data is then analyzed via bioinformatics software.

1.3.2.1 Library Construction

The Forenseq™ DNA Signature Prep utilizes a targeted approach to library construction by employing primer pairs to amplify regions of interest along stretches of DNA. In NGS, a library is defined as a collection of DNA fragments from a single sample containing two key components: 1) regions of interest or target sequences and 2) adapter sequences that allow the region of interest to be sequenced on an NGS instrument such as the MiSeq FGx® [23-29]. Multiple libraries can be pooled and be sequenced in a single analytical run. NGS also resolves some of the limitations of CE by introducing the ability to multiplex different autosomal as well as X-specific and Y-specific STR loci in a single reaction.

The kit includes two primer sets targeting more than 150 forensically-relevant markers: DNA Primer Mix A, which includes 27 autosomal STRs, 24 Y-STRs, 7 X-STRs, Amelogenin, and 94 identity-informative SNPs; and DNA Primer Mix B, which has an additional 54 ancestry and phenotype-informative SNPs. After the extracted DNA is PCR-amplified using one of the primer sets available, the resulting amplicons undergo a second enrichment step. The key to multiplexing and ability to discriminate different samples originates from the tagging of adapter sequences to both the 5' and 3' ends of the DNA fragments during this second enrichment. The ends of both adapters contain complementary sequences that hybridize to oligos attached to the flow cell's glass surface – effectively capturing successfully amplified products and washing away non-specific artifacts. The adapters also contain indexes or barcode sequences, that allow the instrument to parse out each sample, based on a unique combination of indices from the combined

pool. For instance, Sample 1 may have the barcode combination 501, 701 while Sample 2 contains 501, 702. Despite having the same 501 barcode, the addition of the i7 barcodes create a unique combination, which enables demultiplexing (Figure 3) [34-37].

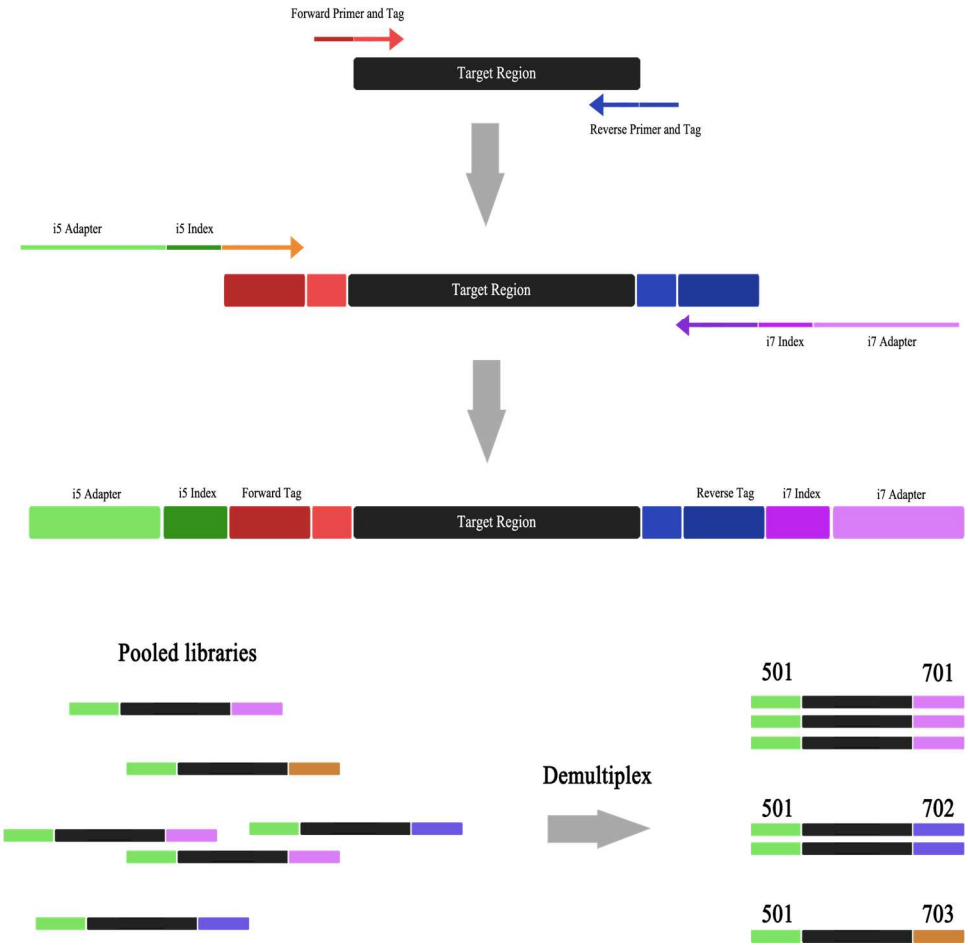

Figure 3. Library Preparation by PCR amplification. After multiple copies of the target region is amplified from a panel of primers, the i5 and i7 indices and adapters are tagged via an additional PCR step, resulting in a complete sequence-compatible library. Each sample is tagged with a unique combination of i5 and i7 barcodes, or indices, which allows the software to demultiplex, or distinguish, the samples after pooling.

1.3.2.2 Sequencing Process

After purification and normalization, an equimolar amount of each library is loaded onto a cartridge to be sequenced using Illumina-derived sequencing technology. The denatured libraries populate and hybridize onto the flow cell's oligo lawn and are extended with DNA Polymerase to produce a double-stranded product before the original strand is denatured and removed from the subsequent reaction. The strands then bend over and are hybridized to neighboring complementary P5 or P7 oligos, extended via DNA polymerase, and then denatured in an entire process termed "bridge amplification" (Figure 4). This process is repeated simultaneously for all library fragments to form clonal clusters containing thousands of copies of the same sequence, which allows for a large enough fluorescent signal to be detected by the sequencer's camera. The amount of clustering affects the quality and quantity of the sequencing output. Low cluster formation under-utilizes the imaging area of the flow cell and may not yield significant coverage of the libraries while high cluster density will overload the camera and lower the accuracy or confidence of base calls. Thus, it is important to either normalize or accurately quantify the libraries and dilute to the recommended concentration prior to sequencing.

For the forward read, the reverse strands are cleaved from the reaction and the resulting free 3' ends of the oligos are blocked to prevent non-specific priming. After the forward primer anneals to the tethered strands on the flow cell, a polymerase and four types of fluorescently-tagged nucleotide bases are washed over the flow cell. The complementary base is incorporated onto the template DNA and the fluorescence is detected by a camera that captures four images, one for each of the nucleotides per cycle. Each nucleotide

contains a 3' reversible terminator that prevents additional bases from being incorporated and moving ahead of the cycle. After the camera completes image capturing, the 3' reversible terminator is removed and the process is repeated for a user-defined number of cycles. After the i7 index is sequenced, the process repeats for the reverse strand and i5 index. This entire process is known as sequencing-by-synthesis (SBS) and the data is processed by the instrument's internal computer and the UAS (Figure 5) [34-40].

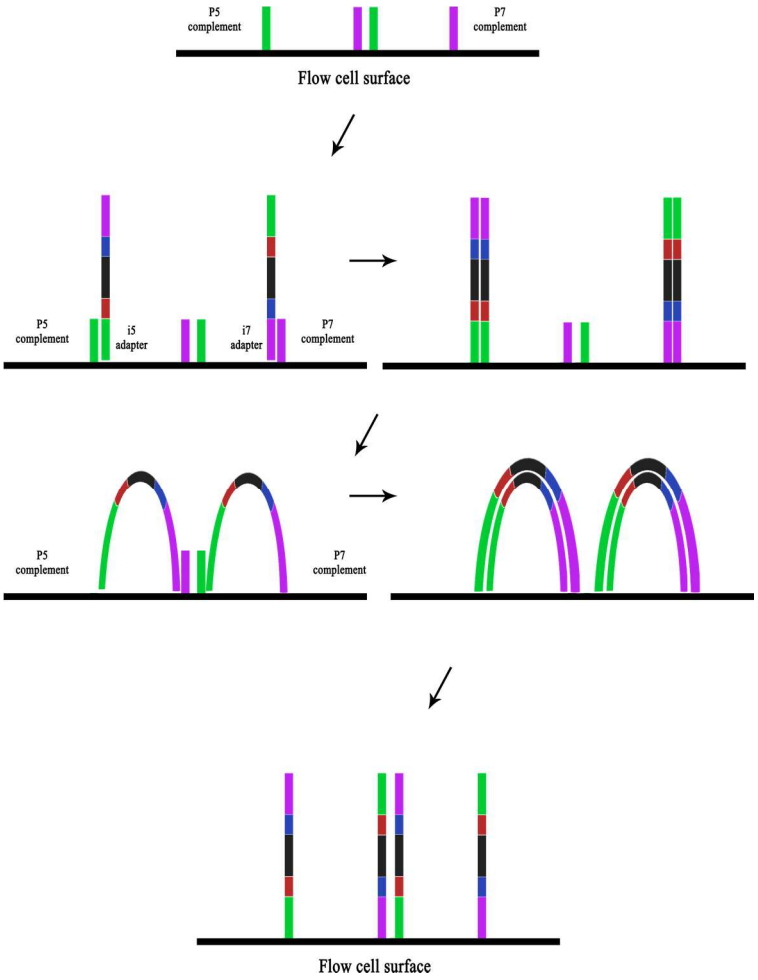

Figure 4. Schematic of bridge amplification. The library pool is washed over the glass surface of the flow cell, which contains tethered oligos complementary to the adapters. Hybridized libraries are then extended via DNA polymerase and the unattached template is removed. The extended strands bend over to anneal to neighboring oligos and are extended once again with DNA polymerase before being denatured to form multiple clusters of the same library fragment.

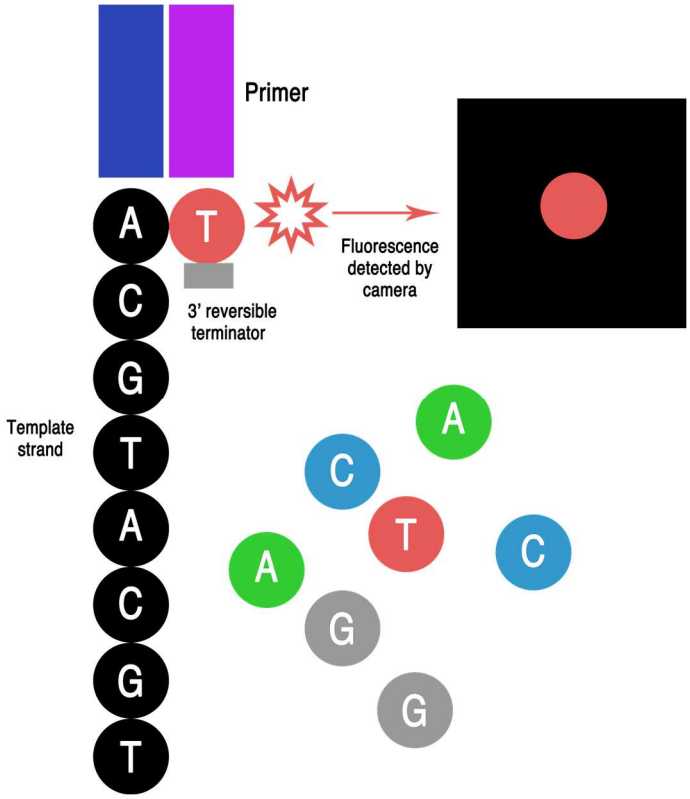

Figure 5. Sequencing by synthesis. Nucleotide bases (adenine, cytosine, guanine, and thymine) with fluorescent tags and a 3' reversible terminators are washed over the flow cell's surface. One complementary base is added to the template strand per cycle. The fluorescence attached to the base is released and detected by the camera before the 3' reversible terminator is cleaved, allowing the next complementary base to be incorporated.

1.3.3 Advantages of MPS

MPS has demonstrated higher sensitivity sequencing for skeletal remains and shorter fragments – producing complete detection rates in more samples compared to traditional CE [41]. Additionally, MPS has been able to successfully generate full profiles for CODIS loci from severely degraded samples [42-43]. MPS technology also offers additional factors of discrimination such as the ability to analyze more biomarkers in parallel, identify SNP data, and most importantly, provide the actual sequence output for

relevant markers. This information can be made accessible to forensic scientists and addresses the aforementioned challenges present in traditional CE-based STR analysis. A collaborative effort, known as the STR Sequencing Project (STRSeq), was initiated to establish an international database of STR sequences for more uniform communication, reporting, and sequence exchange amongst labs [44-45].

1.3.3.1 Isoalleles

Because MPS allows access the specific base calls of the STRs being analyzed, information that was not available with traditional CE methods, it has the advantage of characterizing same-length fragments with different sequences. These isoalelles offer more power of discrimination by differentiating between allele pairs shared between individuals. For instance, two individuals may be homozygous (both alleles have the same STR length) at a locus using the CE method, but could differ in sequence - making them non-identical at that locus using MPS. Isoalleles may also assist in the deconvolution of complex mixtures where contributors share the same allele(s). Being able to identify the amount of contribution for the different isoalleles can guide DNA analysts in mixture interpretation (Figure 6). Current studies on the Human STR Sequence Diversity Database have characterized these sequence variants at 27 of the 44 STR loci amplified by the PowerSeq™ 46GY (Promega Corp, Madison, WI). Autosomal STRs such as SE33, D12S391, D21S11, D2S1338, and D8S1179 presented the highest occurrence of isoalleles, with loci D3S1338 and D8S1179 having the highest allelic diversity [42,44,46]

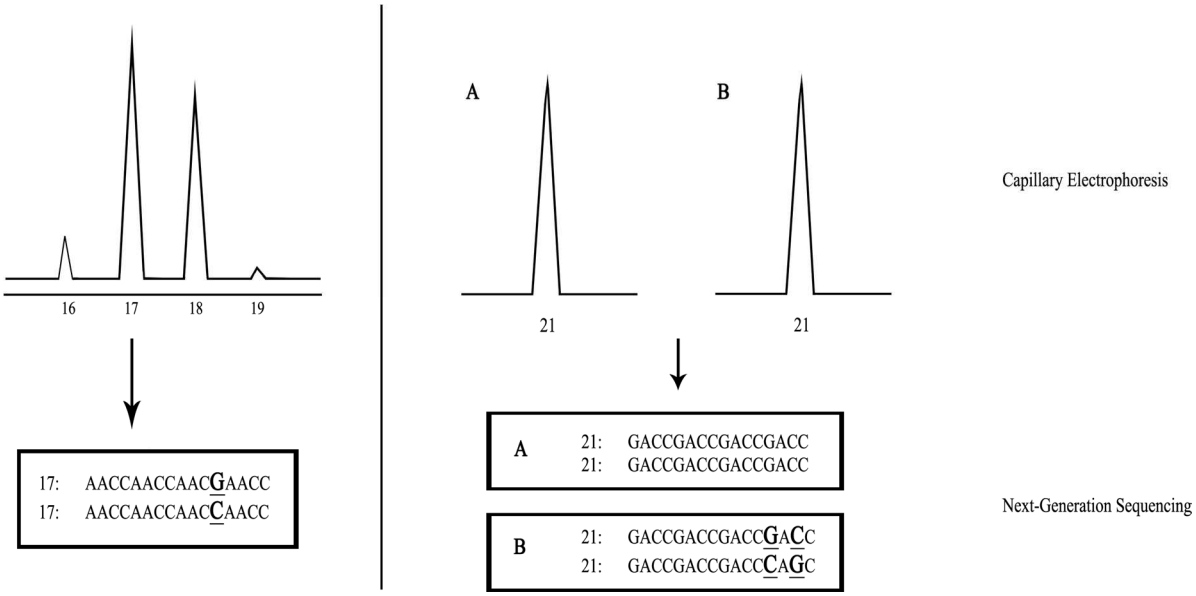

Figure 6. Isoalleles in mixture deconvolution and discrimination. Isoalleles, same-length alleles with sequence variants, can be obtained from NGS to help with mixture interpretation and add power of discrimination. For instance, when interpreting the locus for individual A and B using CE, they would both be reported as homozygotes for allele 21, and thus identical at that locus. With NGS data, however, individual B is observed to have an isoallele - whereas individual A does not – making them unidentical at that locus.

1.3.3.2 Methods for MPS-STR Analysis

The rate of stutter has been shown to have a strong, positive correlation with the longest uninterrupted stretch (LUS), which is defined as the number of identical repeat motifs [19]. However, this scenario is not always true at all loci, which indicates that alternative variables such as flanking and sequence variations may also affect stutter. STRs are categorized as either simple, compound, or complex depending on the number and type of repeat units, and the LUS can be determined based on which repeat motif occurs at the greatest length. Differences in stutter occurrence based on LUS, independent of the total STR length, can be compared between sequence variants of the same allele (Table 1) [19].

However, with sequence-specific MPS data, stutter occurrence can now be identified more specifically by using a method known as the block length of the missing

motif (BLMM) [46]. This concept defines uninterrupted stretches as "blocks," and can be used to define the origin of stutter within a block of an allele. It has been shown that the stutter ratio had a positive linear correlation with the BLMM. Another approach known as "sequence simplification" has been proposed, which uses a four-step process to identify repeat motifs before grouping the same motifs together, along with non-repeat sequences, to produce a condensed or simplified sequence (Table 2). This method can assist in establishing a more consistent approach in characterizing sequence variations and stutter products within alleles [48].

Table 1. Determining the LUS for different types of STRs.

STR Classification	STR Sequence Example	LUS
Simple	(ACGT)(ACGT)(ACGT)(ACGT) (ACGT) (ACGT)	$(ACGT)_6$
Compound	(ACGT)(ACGT)(ACGT)(ACGT)(GATA)(GATA)	$(ACGT)_4$
Complex	(ACGT)(ACGT)(GATA)(GT)(GACA)(GACC)	$(ACGT)_2$

Table 2. Different methods of characterizing stutter. Two methods that have been developed to characterize stutter from MPS data include the block length of the missing motif (BLMM) and sequence simplification approach and are expansions upon the LUS concept.

		BLMM	Sequence Simplification
True Allele Sequence	AATTAATTAATTACTA	Block1 = $[AATT]_3$ Block 2 = $[ACTA]_1$	$(AATT)_3ACTA$
N-1 Stutter Sequence	AATTAATTACTA	Stutter occurs in Block 1	$(AATT)_2ACTA$

1.4 Aims of the Study

The purpose of this study is to observe and characterize isoalleles, and use sequence information to gain a better understanding of stutter patterns using samples with familial relationships, where the same alleles are found via inheritance in different individuals in the pedigree. By obtaining samples with direct inheritance, it is possible to analyze identical isoallele pairs within a small sample size and examine stutter behaviors.

A separate relatedness analysis was also performed to determine whether there is a significant difference in the amount of allele sharing across different relations such as child to parent and first cousins compared to unrelated individuals. The study also investigates whether additional isoallelic information had a statistical impact on predicting relatedness from the amount of allele sharing and what kind of effect that additional sequencing information may have in forensic interpretations such as kinship analysis.

2. MATERIALS AND METHODS

2.1 Sample Preparation

Anonymous human buccal samples, self-identified as Caucasian, were collected from twenty-one persons in the same family (Figure 7) using cotton swabs. The dry saliva samples were extracted and purified by employing the EZ1® Advanced Nucleic Acid Automated Purification System in conjunction with the EZ1® DNA Investigator Kit (Qiagen, Germantown, MD). A ¼ partition was obtained from each of the oral swabs using a sterile scalpel and pretreated with 10 µL of Proteinase K and 290 µL of a lysis buffer labeled - Buffer G2. The samples are then incubated on a thermomixer at 56°C for 15 minutes before being removed and adding 1 µL of carrier RNA. Reagent cartridges, containing silica-coated magnetic particles, and the lysed samples, were then loaded onto the EZ1® instrument according to the manufacturer's instructions for dried saliva. The instrument performs a "tip-dance" protocol for efficient processing and purification of the swabs before the samples are eluted with 40 µL of TE Buffer. DNA quantification used the Quantifiler™ Duo DNA Quantification Kit (Thermo Fisher Scientific, Waltham, MA) on the Applied Biosystems™ 7500 Real-Time PCR System (Thermo Fisher Scientific, Waltham, MA) using a validated external standard curve [49-52]. Each of the purified eluates were diluted to 0.2 ng/µL with nuclease-free water prior to targeted amplification.

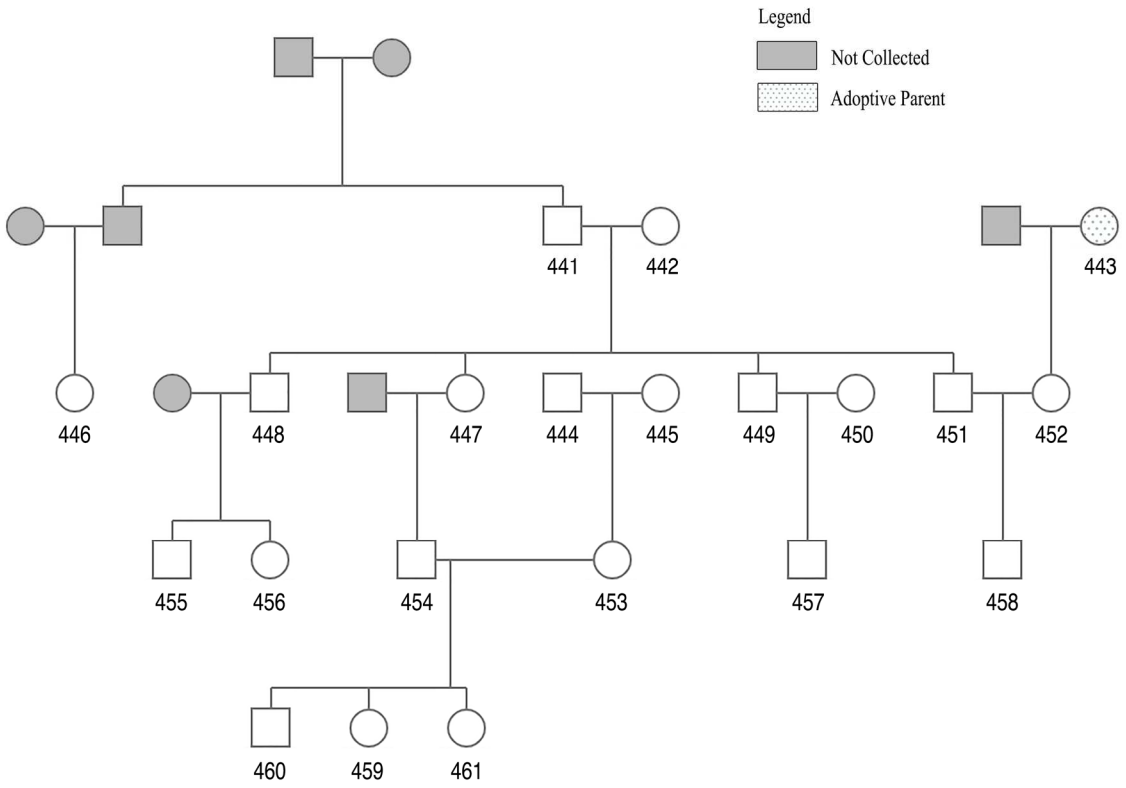

Figure 7. The human family pedigree associated with the twenty-one oral samples extracted in this study.

2.2 Next Generation Sequencing

The pedigree samples, a single-source male genomic 2800M DNA (gDNA) control (Promega, Madison, WI), and a nuclease-free water negative control were then constructed into libraries with the Forenseq™ DNA Signature Prep (Verogen, San Diego, CA) for sequencing on the MiSeq FGx® Forensic Genomics System (Verogen, San Diego, CA) according to the manufacturer's instructions [27-29]. The 2800M DNA is a positive control that is prepared alongside the experimental samples to help assess library preparation.

2.3.1 Targeted Enrichment and Tagging

Enrichment of 1 ng extracted DNA was targeted using DNA Primer Mix A (DPMA), which contains oligonucleotide primer pairs that anneal to upstream and downstream forensically relevant STR and SNP regions (Table 3). The first master mix consisted of 4.7 µL PCR Mix 1 (PCR1), 0.3 µL Forenseq™ Enzyme Mix (FEM), and 5 µL DPMA per reaction. The samples were then transferred to a thermal cycler using the protocol described in Table 4. Upon completion of the PCR program, a unique combination of Index 1 (i7) and Index 2 (i5) adapters were then tagged to the resulting amplicons and enriched to allow for cluster generation and bridge amplification on the provided flow cell. A total of 27 µL of PCR Mix 2 (PCR2) and 4 µL of each index was added to each reaction before running the PCR protocol in Table 5. Both PCR amplification steps were performed in an Applied Biosystems (ABI) GeneAmp® 9700 thermal cycler using the provided settings in the kit's reference guide.

Table 3. DNA Primer Mix A forensic loci. Forensic loci targeted by DNA Primer Mix A from the ForenSeq™ DNA Signature Prep Kit. DNA Primer Mix A targets a total of 58 autosomal and gender-specific STR regions and 94 identity SNPs ranging from 60-500 bps.

Feature	Number of Markers	Amplicon Size Range (bp)
Autosomal STRs	27	61-467
Y-STRs	24	119-390
X-STRs	7	157-462
Identity SNPs	94	63-231

Table 4. Thermal cycler conditions for the first PCR target enrichment.

Step	Temperature (°C)	Time	Cycles
Initial Denaturation	96	45 sec	1
Denaturation	80	30 sec	8
Annealing	54	2 min	
Extension	68	2 min	
Denaturation	96	30 sec	10
Annealing and Extension	68	3 min	
Final Extension	68	10 min	1
Hold	10	∞	1

Table 5. Thermal cycler conditions for the second PCR enrichment.

Step	Temperature (°C)	Time	Cycles
Initial Denaturation	98	30 sec	1
Denaturation	98	20 sec	15
Annealing	66	30 sec	
Extension	68	90 sec	
Final Extension	68	10 min	1
Hold	10	∞	1

2.3.2 Purification and Normalization

Amplified and tagged libraries were then purified using solid-phase reversible immobilization (SPRI) chemistry, which also size-selects for a range of library fragments from approximately 60-460 base pairs (bp). Sample Purification Beads (SPB) were added to the amplified libraries at a 0.9X bead to sample ratio. The reaction was incubated for 5 minutes at room temperature to allow the beads to selectively bind to the desired DNA size range before being placed in a magnetic field to separate the libraries from miscellaneous PCR reagents and products. Two subsequent 200 µL 80% ethanol washes were performed

before the libraries were eluted off the magnetic beads with 52.5 µL of the provided Resuspension Buffer (RSB).

A master mix containing 48.8 µL Library Normalization Additives 1 (LNA1) and 8.5 µL Library Normalization Beads 1 (LNB1) was prepared for each reaction and added to ensure each purified library was present in the combined library pool in equimolar amounts prior to sequencing. LNB1 beads become saturated with libraries and any excess is removed by a series of two washes with 45 µL of Library Normalization Wash 1 (LNW1) solution. The washed beads are then treated and eluted with 32 µL of freshly prepared 0.1 N sodium hydroxide (NaOH), which is labelled as HP3 in the Forenseq™ DNA Signature Prep kit, before being transferred to a new reaction plate containing 30 µL of Library Normalization Storage Buffer 2 (LNS2). The combination of NaOH and formamide in HP3 and LNW1, respectively, allows the DNA to be cleaned and eluted while remaining in a single-stranded state, which is crucial for annealing to the flow cell. This normalization process eliminates varying yields arising from PCR biases and multiple purification steps and allows for a uniform representation of each purified library during cluster generation without further quantification.

2.3.3 Library Pooling and Loading onto the MiSeq FGx®

After normalization, 5 µL of each library was combined into a single reaction tube. From the pooled libraries, 7 µL was then diluted in 591 µL of Hybridization Buffer (HT1). A denaturation master mix containing 2 µL HP3, 2 µL Human Sequencing Control (HSC), and 36 µL of nuclease-free water was prepared and 2 µL of the mix was added to the diluted

library pool. The HSC acts as a positive control to assess sequencing performance and ensures the completion of the sequencing run in the presence of low library quality. The final denatured and diluted library pool was then placed on a heating block for 2 minutes at 96°C before being immediately transferred on ice for 5 minutes.

After the MiSeq FGx™ reagent cartridge was thoroughly thawed and mixed, the library pool was pipetted into the appropriate cartridge well before loading onto the sequencer. Verogen provides two different reagent kits – a standard MiSeq FGx™ Reagent Kit and a MiSeq FGx™ Micro Kit. Both versions are identical with the exception being the amount of imaging area available on the flow cell's surface. The flow cell provided in the MiSeq FGx™ Micro Kit used in this study has approximately 35% of the area of the standard flow cell and is capable of approximately five million paired reads, as opposed to 12.5 million paired reads in the other kit. The manufacturer suggests a maximum input of 36 single-source samples per run with the combination of the Micro Kit and DPMA. The micro flow cell was cleaned with an ethanol wipe before being placed on the sequencer.

2.3.4 Data Analysis using the UAS

The UAS is a pre-installed software that allows the setup of sequencing run criteria and automatically performs various analyses from sequencing output. Such analyses include calculating quality metrics such as cluster density (K/mm^2), clusters passing filter (%), and amount of phasing/pre-phasing (%). Instead of requiring an allelic ladder or size standard, such as in the case for CE, the UAS analyzes each read to determine the length, sequence, and stutter qualifications. In addition, it maps STR/SNP loci based on the

sequence of the primer utilized in the first targeted amplification and identifies the allele by dividing the repeat length from the total length of the locus. All similar sequences are binned together to determine the number of reads. Genotypes and stutter are determined based on locus-specific and percentage-based thresholds, which were established empirically by the manufacturer [32].

3. RESULTS and DISCUSSION

3.1 Sequencing Quality

A variety of quality metrics from the sequencing run, such as the number of called STRs and SNPs for controls, cluster density and phasing/pre-phasing, were assessed and all were in line with the manufacturer recommendations for reliable data (Tables 6 and 7).

Cluster density refers to the number of clusters on the flow cell after bridge amplification and should be optimized to prevent low-data output from under or over-clustering. Phasing refers to clusters falling behind the current sequencing cycle due to unsuccessful incorporation of a nucleotide, while pre-phasing occurs when clusters move ahead from the current sequencing cycle. These false reads are identified and removed prior to analysis. The AT and interpretation threshold (IT) for all loci are set to the instrument's default values at 1.5% and 4.5%, respectively. The exceptions to these AT and IT values are Y-STR loci DYS389II (>5.0% AT, >15% IT), DYS448 (>3.3% AT, >10% IT), and DYS635 (>3.3% AT, >10% IT) [25].

After reviewing the reported alleles in the sequencing data, all samples had successful allele calls with the exception of sample 445, which was found to contain a mixture of allele pairs, and thus was omitted for the rest of the analysis due to possible contamination with another female contributor during sample collection or extraction.

Table 6. Quality metrics for sequencing controls. The positive control is the 2800M DNA while the negative control is a DNA-free and nuclease-free water control.

# Samples	Type of samples	POS STRs	POS SNPs	NEG STRs	NEG SNPs
23	single-source	59/59	94/94	0	0

Table 7. Quality metrics for sequencing run. Optimal cluster density range, according to manufacturer, ranges from 400-1650 K/mm^2. Clusters passing filter should exceed 80%. Phasing should fall under 0.25% and 0.15% for pre-phasing events.

	Cluster Density (K/mm^2)	Clusters Passing Filter (%)	Phasing (%)	Pre-phasing (%)
Recommended	400-1650	> 80	< 0.25	< 0.15
Experimental Results	1125	89.31	0.167	0.131

3.2 Isoallele Frequency

Nine unique pairs of isoalleles were identified across the 23 samples, in nine individuals and the positive control for a total of 15 isoalleles in the entire sequencing pool (Table 8). The sequences of each of the isoalleles were extracted from the MPS data and organized based on the block and simplified-sequence approach. Nucleotide variants were also identified (Table 9). Isoalleles were found in two alleles at D2S1338, which has been previously shown to exhibit high levels of variation [53]. Another study reported isoalleles in D13S317 with a similar A-T substitution, but at the beginning of the repeat motif, as is the case here [54]. Among all DNA samples, nucleotide mutations ranged from 1-3 base pair differences between the isoalleles with 5/9 (55.6%) having a single mutation, 3/9 (33.3%) having two mutations, and 1/9 (11.1%) containing three mutations. For instance, in sample 442, the locus D13S317 allele 11 had a nucleotide variation of T>A occurring at

the end of the LUS. In sample 441, the locus D8S1179 allele 13 exhibited two polymorphisms G>A and A>G. Three polymorphisms were identified in locus DXS10135 allele 24 in sample 446. Overall, these SNPs appear to occur most frequently at either the beginning or end of a repeating motif. Multiple motifs that were not identified in the "NIST 1036" data set, a study that characterized sequence-based motifs from a 1036 samples using the same MPS method, were also identified such as at the DXS10135 locus allele 24. Although the Forenseq™ DNA Signature Prep kit was also used in this NIST study, the motifs identified for locus D2S1338 included [GGAA] and [GGCA] tetranucleotides whereas [TTGC] and [TTCC] motifs were found in this study [55].

Table 8. Isoalleles identified in this study. Nine pairs of unique isoalleles were identified in this study in nine individuals and the 2800M positive control.

#	Locus	Allele	Samples
1	D13S317	11	442, 443, 447, 448, 455, 456
2	D1S1656	16	454
3	D2S1338	21	442
4	D2S1338	20	458
5	D3S1358	16	455
6	D5S818	12	447
7	D8S1179	13	441, 454
8	D9S1122	12	2800M Positive Control
9	DXS10135	24	446

Table 9. Simplified sequences of identified isoalleles. Number of sequence variations identified in each isoallele pair (marked in red) and its corresponding full and simplified sequence.

Sample	Locus	Allele	# of SNPs	Origin	Simplified Sequence	Sequence
441	D8S1179	13	2	N/A	TCTATCTG [TCTA]$_{11}$‡	[TCTA] [TCT**G**] [TCT**A** TCTA TCTA TCTA TCTA TCTA TCTA TCTA TCTA TCTA]
				N/A	[TCTA]$_2$[TCTG] [TCTA]$_{10}$‡	[TCTA TCT**A**] [TCT**G**] [TCTA TCTA TCTA TCTA TCTA TCTA TCTA TCTA TCTA]
442	D2S1338	21	2	N/A	[TGCC]$_7$[TTCC]$_{11}$ [GTCC][TTCC]$_2$	[TGCC TGCC TGCC TGCC TGCC TGCC TGCC] [T**T**CC TTCC TTCC TTCC TTCC TTCC TTCC TTCC TTCC TTCC] [**G**TCC] [TTCC TTCC]
				N/A	[TGCC]$_8$[TTCC]$_{13}$	[TGCC TGCC TGCC TGCC TGCC TGCC TGCC T**G**CC] [TTCC TTCC TTCC TTCC TTCC TTCC TTCC TTCC TTCC **T**TCC TTCC TTCC]
442	D13S317	11	1	N/A	[TATC]$_{12}$[AATC] [ATCT]$_3$[TTCT] [GTCTGTC]	[TATC TATC TATC TATC TATC TATC TATC TATC TATC TATC TATC **T**ATC] [AATC] [ATCT ATCT ATCT] [TTCT] [GTCT GTC]
				N/A	[TATC]$_{11}$[AATC]$_2$ [ATCT]$_3$[TTCT] [GTCTGTC]	[TATC TATC TATC TATC TATC TATC TATC TATC TATC TATC TATC] [**A**ATC AATC] [ATCT ATCT ATCT] [TTCT] [GTCT GTC]
443	D13S317	11	1	N/A	[TATC]$_{12}$[AATC] [ATCT]$_3$[TTCT] [GTCTGTC]	[TATC TATC TATC TATC TATC TATC TATC TATC TATC TATC TATC **T**ATC] [AATC] [ATCT ATCT ATCT] [TTCT] [GTCT GTC]
				N/A	[TATC]$_{11}$[AATC]$_2$ [ATCT]$_3$[TTCT]	[TATC TATC TATC TATC TATC TATC

					[GTCTGTC]	TATC TATC TATC TATC TATC] [**A**ATC AATC] [ATCT ATCT ATCT] [TTCT] [GTCT GTC]
Sample	**Locus**	**Allele**	**# of SNPs**	**Origin**	**Simplified Sequence**	**Sequence**
446	DXS10135	24	3	N/A	$[AAGA]_3[GAAA][GGAA][AGAA]_{16}[AGGA]_3[AGAA][AAGA][GAAT][AGAA][AAGA][AGAG][AAGA][GAAA][AGAG][AAAA][GAAA][AAAG][AAAA][GAAA]$	[AAGA AAGA AAGA] [GAAA] [GGAA] [AGAA AGAA AGAA AGAA AGAA AGAA AGAA AGAA AGAA AGAA AGAA AGAA AG**A**A AGAA] [AG**G**A AG**G**A AGGA] [AGAA] [AAGA] [GAAT] [AGAA] [AAGA] [AGAG] [AAGA] [GAAA] [AGAG] [AAAA] [GAAA] [AAAG] [AAAA] [GAAA]
				N/A	$[AAGA]_3[GAAA][GGAA][AGAA]_{14}[AGGA][AGAA]_3[AGGA][AGAA] [AAGA] [GAAT] [AGAA] [AAGA] [AGAG] [AAGA] [GAAA] [AGAG] [AAAA] [GAAA] [AAAG] [AAAA] [GAAA]$	[AAGA AAGA AAGA] [GAAA] [GGAA] [AGAA AGAA AGAA AGAA AGAA AGAA AGAA AGAA AGAA AGAA AGAA AGAA] [AG**G**A] [AGAA AG**A**A AG**A**A] [AGGA] [AGAA] [AAGA] [GAAT] [AGAA] [AAGA] [AGAG] [AAGA] [GAAA] [AGAG] [AAAA] [GAAA] [AAAG] [AAAA] [GAAA]
447	D5S818	12	1	M*	$[AGAT]_{12}[AGAG]$	[AGAT AGAT AGAT AGAT AGAT AGAT AGAT AGAT AGAT AGAT AGAT AGAT] [AGA**G**]
				P†	$[AGAT]_{13}$	[AGAT AGAT AGAT AGAT AGAT AGAT AGAT AGAT AGAT AGAT AGAT AGAT] AGA**T**]
447	D13S317	11	1	M	$[TATC]_{11}[AATC]_2[ATCT]_3[TTCT][GTCTGTC]$	[TATC TATC TATC TATC TATC TATC TATC TATC TATC TATC TATC] [**A**ATC AATC] [ATCT ATCT

						ATCT] [TTCT] [GTCT GTC]
				P	[TATC]$_{12}$[AATC][ATCT]$_3$[TTCT][G TCTGTC]	[TATC TATC TATC TATC TATC TATC TATC TATC TATC TATC **T**ATC] [AATC] [ATCT ATCT ATCT] [TTCT] [GTCT GTC]
Sample	**Locus**	**Allele**	**# of SNPs**	**Origin**	**Simplified Sequence**	**Sequence**
448	D13S317	11	1	P	[TATC]$_{12}$[AATC] [ATCT]$_3$[TTCT] [GTCTGTC]	[TATC TATC TATC TATC TATC TATC TATC TATC TATC TATC **T**ATC] [AATC] [ATCT ATCT ATCT] [TTCT] [GTCT GTC]
				M	[TATC]$_{11}$[AATC]$_2$ [ATCT]$_3$[TTCT][G TCTGTC]	[TATC TATC TATC TATC TATC TATC TATC TATC TATC TATC] [**A**ATC AATC] [ATCT ATCT ATCT] [TTCT] [GTCT GTC]
454	D1S1656	16	1	M	[TAGA]$_{15}$[TAGG] [TG]$_5$	[TAGA TAGA TAGA TAGA TAGA TAGA TAGA TAGA TAGA TAGA TAGA TAGA TAGA TAGA TAGA] [TAG**G**] [TG TG TG TG TG]
				P	[TAGA]$_{16}$[TG]$_5$	[TAGA TAGA TAGA TAGA TAGA TAGA TAGA TAGA TAGA TAGA TAGA TAGA TAGA TAGA TAG**A**] [TG TG TG TG TG]
454	D8S1179	13	2	P	TCTATCTG[TCT A]$_{11}$‡	[TCTA] [TCT**G**] [TCT**A** TCTA TCTA TCTA TCTA TCTA TCTA TCTA TCTA TCTA TCTA]
				M	[TCTA]$_2$[TCTG][T CTA]$_{10}$‡	[TCTA TCT**A**] [TCT**G**] [TCTA TCTA TCTA TCTA TCTA TCTA TCTA TCTA TCTA TCTA]

Sample	Locus	Allele	# of SNPs	Origin	Simplified Sequence	Sequence
455	D3S1358	16	2	M	[TCTA][TCTG]$_3$[TCTA][TCTG][TCTA]$_{10}$	[TCTA] [TCTG TCTG TCT**G**] [TCTA] [TCT**G**] [TCTA TCTA TCTA TCTA TCTA TCTA TCTA TCTA TCTA]
				P	[TCTA][TCTG]$_2$[TCTA]$_{13}$‡	[TCTA] [TCTG TCTG] [TCT**A** TCTA TCT**A** TCTA TCTA TCTA TCTA TCTA TCTA TCTA TCTA]
455	D13S317	11	1	N/A	[TATC]$_{11}$[AATC]$_2$ [ATCT]$_3$[TTCT][GTCTGTC]	[TATC TATC TATC TATC TATC TATC TATC TATC TATC TATC] [**A**ATC AATC] [ATCT ATCT ATCT] [TTCT] [GTCT GTC]
				N/A	[TATC]$_{12}$[AATC][ATCT]$_3$[TTCT][GTCTGTC]	[TATC TATC TATC TATC TATC TATC TATC TATC TATC TATC **T**ATC] [AATC] [ATCT ATCT ATCT] [TTCT] [GTCT GTC]
456	D13S317	11	1	N/A	[TATC]$_{12}$[AATC][ATCT]$_3$[TTCT][GTCTGTC]	[TATC TATC TATC TATC TATC TATC TATC TATC TATC TATC **T**ATC] [AATC] [ATCT ATCT ATCT] [TTCT] [GTCT GTC]
				N/A	[TATC]$_{11}$[AATC]$_2$ [ATCT]$_3$[TTCT][GTCTGTC]	[TATC TATC TATC TATC TATC TATC TATC TATC TATC TATC] [**A**ATC AATC] [ATCT ATCT ATCT] [TTCT] [GTCT GTC]
458	D2S1338	20	1	P	[TGCC]$_7$[TTCC]$_{10}$ [GTCC][TTCC]$_2$	[TGCC TGCC TGCC TGCC TGCC TGCC TGCC] [TTCC TTCC TTCC TTCC TTCC TTCC TTCC TTCC TTCC TTCC] [**G**TCC] [TTCC TTCC]
				M	[TGCC]$_7$[TTCC]$_{13}$	[TGCC TGCC TGCC TGCC TGCC TGCC TGCC] [TTCC TTCC

						TTCC TTCC TTCC TTCC TTCC TTCC TTCC TTCC TTCC TTCC]
Sample	**Locus**	**Allele**	**# of SNPs**	**Origin**	**Simplified Sequence**	**Sequence**
POS CNTRL	D9S1122	12	1	N/A	$[TAGA]_{12}$‡	[TAGA TAGA TAGA TAGA TAGA TAGA TAGA TAGA TAGA TAGA TAGA TAGA]
				N/A	[TAGA][TCGA][TAGA]$_{10}$‡	[TAGA] [TCGA] [TAGA TAGA TAGA TAGA TAGA TAGA TAGA TAGA TAGA TAGA]

* Isoallele was contributed by the maternal line
† Isoallele was contributed by the paternal line
‡ Isoallele was previously characterized by the NIST 1036 data set

3.3 Variation Effects on Stutter Ratio

The effect of increasing LUS and stutter ratio and the degree of change for increasing LUS between isoallele pairs were compared (Figures 8 and 9). The stutter sequences of each isoallele pair were also simplified and the LUS and BLMM identified. With isoalleles, SNPs in the sequence for same-length alleles and their effect on stutter ratio can be observed. The non-linear relation between LUS and stutter ratio is consistent with previous findings on multiple variables impacting stutter rate such as nucleotide content, STR type, and flanking variations [17-19]. Although it has been found that there is high predictability between stutter ratio and LUS [19], some exceptions were reported in the pedigree MPS data. One example of this deviation occurs for the homozygous 11 isoalleles in D13S317. The allele with the shorter LUS $[TATC]_{11}$ exhibited lower or comparable stutter ratios compared to $[TATC]_{12}$. Additionally, despite being two repeats shorter, $[TCTA]_{11}$ had a stutter ratio of 14.24% while $[TCTA]_{13}$ had a ratio of 13.45%. After identifying the block in which the stutters occurred for each of the isoallele pairs, it

was found that 92.3% (24/26) of stutter occurred within the LUS of that allele. The exceptions were the D13S317 loci allele 11 in samples 442 and 443 where the stutter occurred in the [AATC] repeat as opposed to the [TATC] LUS.

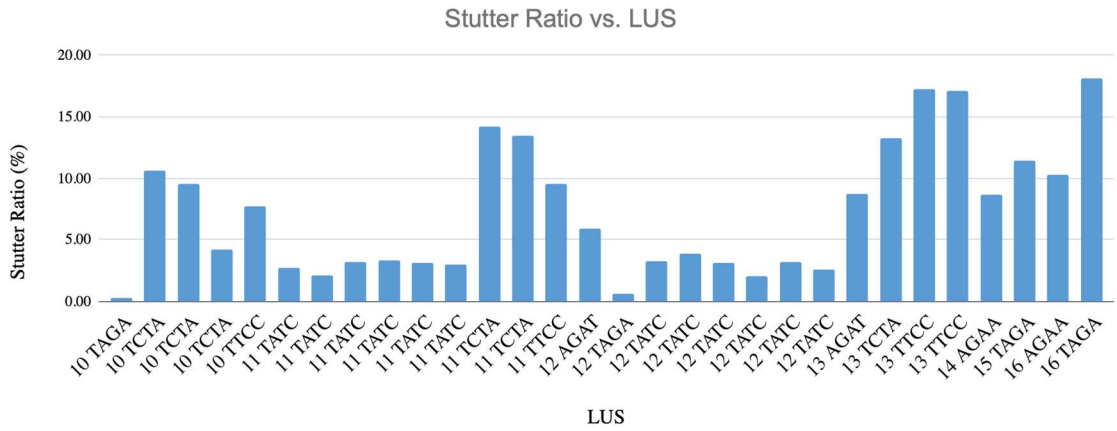

Figure 8. Effect on stutter ratio with increasing longest uninterrupted stretch (LUS).

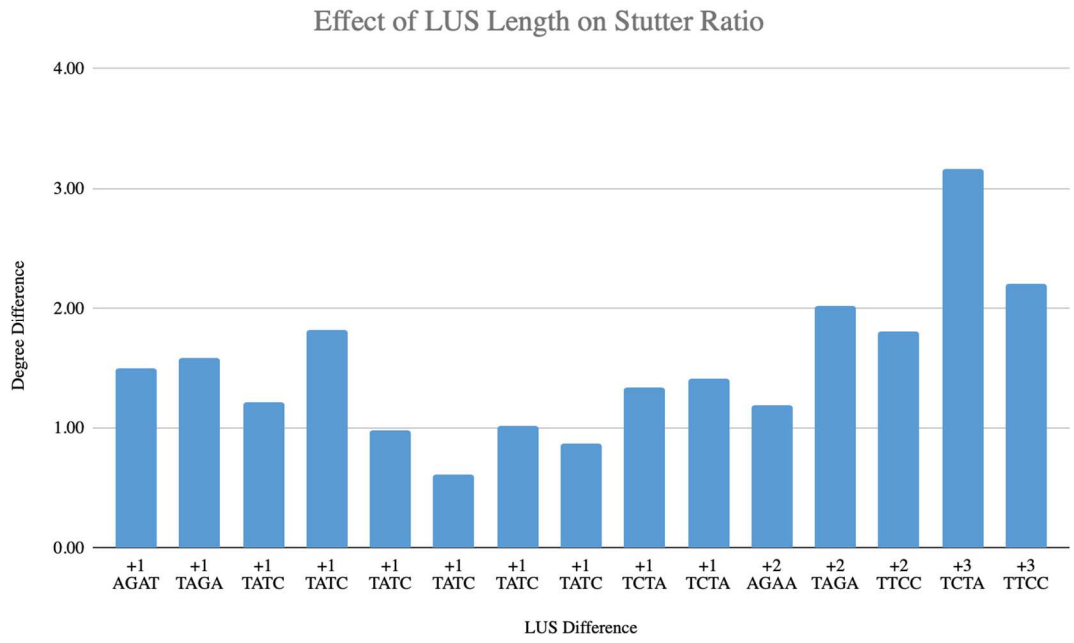

Figure 9. Degree of change in stutter ratio based on different repeat increases in longest uninterrupted stretch (LUS).

3.4 Different Stutter Behaviors in Same Isoalleles

Since there were multiple occurrences of the D13S317 isoallele 11, it was possible to observe differences in stutter behavior among the samples. Three different N-1 stutter patterns were observed in D13S317 for allele 11. The first, and predictably the most common amongst the samples, was a missing repeat in the first block $[TATC]_{12}$ or $[TATC]_{11}$, which was also the LUS in both isoalleles. This stutter pattern was found in both isoalleles in samples 448 and in one of the pair of isoalleles in samples 447, 455, and 456. The second type of stutter was a missing repeat in the second block $[AATC]_2$ and occurred in two unrelated individuals – samples 442 and 443. The third type was a stutter with two repeats missing in the LUS $[TATC]_2$ and the addition of a [AATC] motif, which only occurred once in one of the isoalleles in sample 456. The sequences of the stutters are simplified in Table 10, and not all stutter behaviors were observed in the two isoalleles. Considering the small sample size, this suggests that the behavior of slipped-strand mispairing can differ significantly even in inherited same-sequence alleles.

Table 10. Different stutter patterns in D13S317. Two sequence variants were identified for D13S317 Allele 11. Among the two isoalleles, three types of stutter behaviors were observed. The stutters are characterized by their simplified sequences.

D13S317 Allele 11	**$(TATC)_{12}AATC(ATCT)_3TTCTGTCTGTC$**	**Occurrence**	**$(TATC)_{11}(AATC)_2(ATCT)_3TTCTGTCTGTC$**	**Occurrence**
Stutter 1	$(TATC)_{11}AATC(ATCT)_3TTCTGTCTGTC$	1	$(TATC)_{10}(AATC)_2(ATCT)_3TTCTGTCTGTC$	4
Stutter 2	Not observed	n/a	$(TATC)_{11}AATC(ATCT)_3TTCTGTCTGTC$	2
Stutter 3	$(TATC)_{10}(AATC)_2(ATCT)_3TTCTGTCTGTC$	1	Not observed	n/a

3.5 Relatedness Analysis

An analysis on the percent of total allele sharing based on familial relation for the 20 pedigree samples was performed for 27 autosomal STRs (Table 11). From the pedigree chart, ten different types of relation were defined (Table 12). Alleles may be shared based on biological relationships as well as general sharing of alleles within a population. As shown in Figure 10, the amount of allele sharing can introduce difficulties in distinguishing specific relationship in kinship analyses once one goes beyond immediate relations (biological parents and full siblings). For the grandparent and uncle/aunt relations, the highest data point for unrelated individuals based on the samples studied from the pedigree (44.44%) is still equal to the average of both types of relations (44.44% and 46.06% respectively). When isoalleles were taken into consideration, the average percent of allele sharing for all categories decreased by an average of 1.3% and two previously reported outliers within the first cousin and niece/nephew/uncle/aunt relation are now corrected and fall within the predicted range (Figure 11). However, after assessing the statistical difference of each type of the relationship for the test group (with isoalleles) versus the control group (without isoalleles) using the t-test, it was found that none of the groups demonstrated a statistically significant difference at an alpha level of 0.05 (Table 13).

Table 11. Percent relatedness for all individuals in pedigree. Comparison of all individuals, except for 445, in the pedigree to determine percentage of allele sharing based on the type of biological relationship. The upper half contains the data without taking isoalleles into consideration while the lower half are the corrected percentages with sequence variation.

without isoalleles

with isoalleles

	441	442	443	444	446	447	448	449	450	451	452	453	454	455	456	457	458	459	460	461
441		25.93%	27.78%	25.93%	46.30%	57.41%	61.11%	57.41%	40.74%	62.96%	29.63%	22.22%	53.70%	48.15%	44.44%	44.44%	48.15%	33.33%	40.74%	42.59%
442	24.07%		38.89%	24.07%	35.19%	57.41%	59.26%	57.41%	44.44%	50.00%	38.89%	27.78%	38.89%	40.74%	64.81%	61.11%	44.44%	29.63%	29.63%	24.07%
443	27.78%	38.89%		29.63%	38.89%	33.33%	38.89%	25.93%	33.33%	33.33%	31.48%	33.33%	27.78%	27.78%	31.48%	35.19%	37.04%	37.04%	27.78%	27.78%
444	25.93%	24.07%	29.63%		25.93%	20.37%	20.37%	20.37%	31.48%	29.63%	31.48%	66.67%	25.93%	18.52%	27.78%	25.93%	33.33%	53.70%	44.44%	44.44%
446	46.30%	35.19%	38.89%	24.07%		35.19%	37.04%	37.04%	37.04%	31.48%	24.07%	20.37%	33.33%	33.33%	31.48%	29.63%	22.22%	27.78%	25.93%	31.48%
447	57.41%	57.41%	33.33%	18.52%	35.19%		61.11%	59.26%	42.59%	57.41%	35.19%	24.07%	55.56%	50.00%	57.41%	53.70%	40.74%	33.33%	40.74%	31.48%
448	61.11%	57.41%	38.89%	18.52%	37.04%	61.11%		55.56%	40.74%	61.11%	31.48%	31.48%	48.15%	61.11%	70.37%	51.85%	42.59%	40.74%	35.19%	38.89%
449	57.41%	55.56%	24.07%	18.52%	37.04%	55.56%	51.85%		35.19%	46.30%	31.48%	24.07%	50.00%	46.30%	50.00%	64.81%	38.89%	24.07%	35.19%	35.19%
450	38.89%	44.44%	33.33%	31.48%	37.04%	42.59%	40.74%	35.19%		35.19%	44.44%	29.63%	33.33%	25.93%	33.33%	62.96%	40.74%	29.63%	29.63%	37.04%
451	62.96%	50.00%	33.33%	27.78%	31.48%	55.56%	59.26%	46.30%	35.19%		33.33%	25.93%	46.30%	33.33%	50.00%	40.74%	57.41%	38.89%	35.19%	35.19%
452	27.78%	38.89%	31.48%	31.48%	22.22%	31.48%	29.63%	29.63%	44.44%	29.63%		33.33%	24.07%	27.78%	35.19%	42.59%	64.81%	29.63%	31.48%	25.93%
453	22.22%	27.78%	33.33%	66.67%	20.37%	24.07%	31.48%	24.07%	29.63%	25.93%	33.33%		27.78%	24.07%	31.48%	24.07%	33.33%	61.11%	55.56%	57.41%
454	51.85%	37.04%	27.78%	25.93%	31.48%	55.56%	46.30%	44.44%	33.33%	42.59%	24.07%	27.78%		38.89%	42.59%	37.04%	35.19%	57.41%	61.11%	55.56%
455	48.15%	40.74%	27.78%	16.67%	33.33%	50.00%	61.11%	42.59%	25.93%	31.48%	25.93%	24.07%	38.89%		57.41%	35.19%	27.78%	27.78%	33.33%	31.48%
456	42.59%	62.96%	31.48%	25.93%	29.63%	55.56%	68.52%	44.44%	31.48%	46.30%	31.48%	31.48%	40.74%	57.41%		51.85%	35.19%	33.33%	38.89%	35.19%
457	42.59%	61.11%	35.19%	24.07%	29.63%	51.85%	48.15%	64.81%	62.96%	40.74%	40.74%	24.07%	35.19%	35.19%	48.15%		44.44%	27.78%	33.33%	33.33%
458	46.30%	44.44%	37.04%	33.33%	20.37%	38.89%	40.74%	35.19%	40.74%	53.70%	64.81%	33.33%	33.33%	25.93%	33.33%	40.74%		37.04%	35.19%	29.63%
459	37.04%	29.63%	37.04%	53.70%	25.93%	31.48%	38.89%	22.22%	29.63%	37.04%	29.63%	61.11%	57.41%	25.93%	31.48%	24.07%	35.19%		64.81%	68.52%
460	37.04%	27.78%	27.78%	42.59%	24.07%	38.89%	33.33%	33.33%	29.63%	33.33%	29.63%	55.56%	61.11%	33.33%	37.04%	29.63%	35.19%	61.11%		68.52%
461	42.59%	24.07%	27.78%	44.44%	29.63%	29.63%	37.04%	33.33%	37.04%	33.33%	25.93%	57.41%	55.56%	29.63%	33.33%	29.63%	29.63%	68.52%	68.52%	

	no relation		child to parent		first cousins once removed	niece/nephew to uncle/aunt		grand-child to grandparent
	siblings		first cousins		first cousins twice removed	great-niece/nephew to great uncle/aunt		great grandchild to great grandparent

Table 12. Average number of shared alleles, standard deviation, and number of comparisons made for each of the ten defined relationships present in the pedigree.

			Without Isoalleles		With Isoalleles	
Relationship	**# of Comparisons**	**Sample Pairs**	**Average shared alleles**	**Standard deviation**	**Average shared alleles**	**Standard deviation**
no relation	76	(441,442) (441,443) (441, 444) (441,450) (441,452) (441,453) ; (442,443) (442,444) (442,446) (442,450) (442,452) (442,453) ; (443,444) (443,446) (443,447) (443,448) (443,449) (443,450) (443,451) (443,452) (443,453) (443,454) (443,455) (443,456) (443,457) (443,458) (443,459) (443,460) (443,461) ; (444,446) (444,447) (444,448) (444,449) (444,450) (444,451) (444,452) (444,454) (444,455) (444,456) (444,457) (444,458); (446,450) (446,452) (446,453); (447,450) (447,452) (447,453); (448,450) (448,452); (449,450) (449,452); (450,451) (450,452) (450,453) (450,454) (450,455) (450,456) (450,458) (450,459) (450,460) (450,461); (451,452) (451,453); (452,453) (452,454) (452,455) (452,456) (452,457) (452,459) (452,460 (452,461)); (453,454) (453,455) (453,456) (453,457) (453,458)	30.97%	6.05	30.36%	6.27
first cousins twice removed	3	(446,459) (446,460) (446,461)	28.40%	2.31	26.54%	2.31
great niece/nephew to great uncle/aunt	9	(448,459) (448,460) (448,461) (449,459) (449,460) (449, 461) (451,459) (451,460) (451,461)	36.48%	4.50	34.81%	4.51
first cousins once removed	17	(446,454) (446,455) (446,456) (446,457) (446,458) (455,459) (455,460) (455,461) (456,459) (456,460) (456,461) (457,459) (457,460) (457,461) (458,459) (458,460) (458,461)	32.13%	3.81	30.50%	4.07

			With Isoalleles		Without Isoalleles	
Relationship	**# of Comparisons**	**Sample Pairs**	**Average shared alleles**	**Standard Deviation**	**Average shared alleles**	**Standard Deviation**
great-grandchild to great-grandparent	6	(441,459) (441,460) (441,461) (442,459) (442,460) (442,461)	33.33%	6.50	33.03%	6.36
first cousins	13	(446,447) (446,448) (446,449) (446,451) (454,455) (454,456) (454,457) (454,458) (455,457) (455,458) (456,457) (456,458) (457,458)	37.61%	5.78	36.33%	5.11
niece/nephew to uncle/aunt	17	(447,455) (447,456) (447,457) (447,458) (448,453) (448,454) (448,457) (448,458) (449,453) (449,454) (449,455) (449,456) (449,458) (451,454) (451,455) (451,456) (451,457)	44.44%	8.48	42.05%	7.80
grandchild to grandparent	16	(441,454) (441,455) (441,456) (441,457) (441,458) (442,454) (442,455) (442,456) (442,457) (442,458) (444,459) (444,460) (444,461) (447,459) (447,460) (447,461)	46.06%	8.66	44.91%	8.92
siblings	10	(447,448) (447,449) (447,451) (448,449) (448,451) (449,451); (459,460); (459,461); (460,461)	60.29%	6.26	58.64%	6.53
child to parent	22	(441,447) (441,448) (441,449) (441,451) (442,447) (442,448) (442,449) (442,451) (444,453); (447,454); (448,455) (448,456) (449,457); (450,457); (451,458); (452,458) (453,459) (453,460) (453,461) (454,459) (454,460) (454,461)	59.77%	4.40	59.34%	4.45

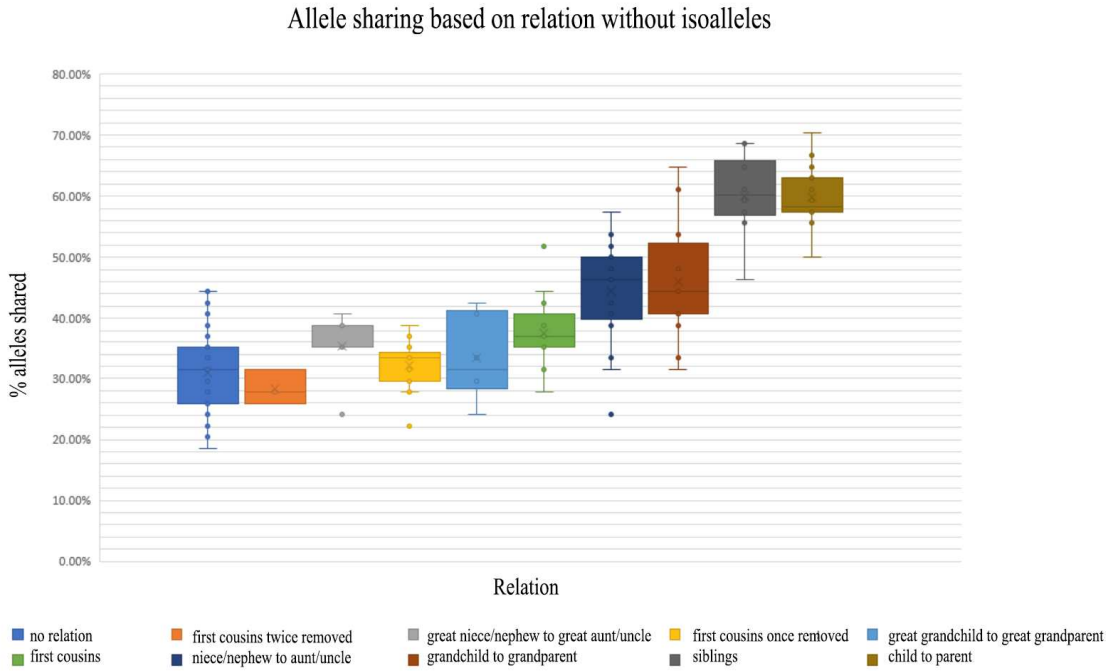

Figure 10. Box-and-whisker plot of allele sharing and sample variation without isoallele data. The median percent of allele sharing and range is based on family relation, as depicted in the pedigree used in this study.

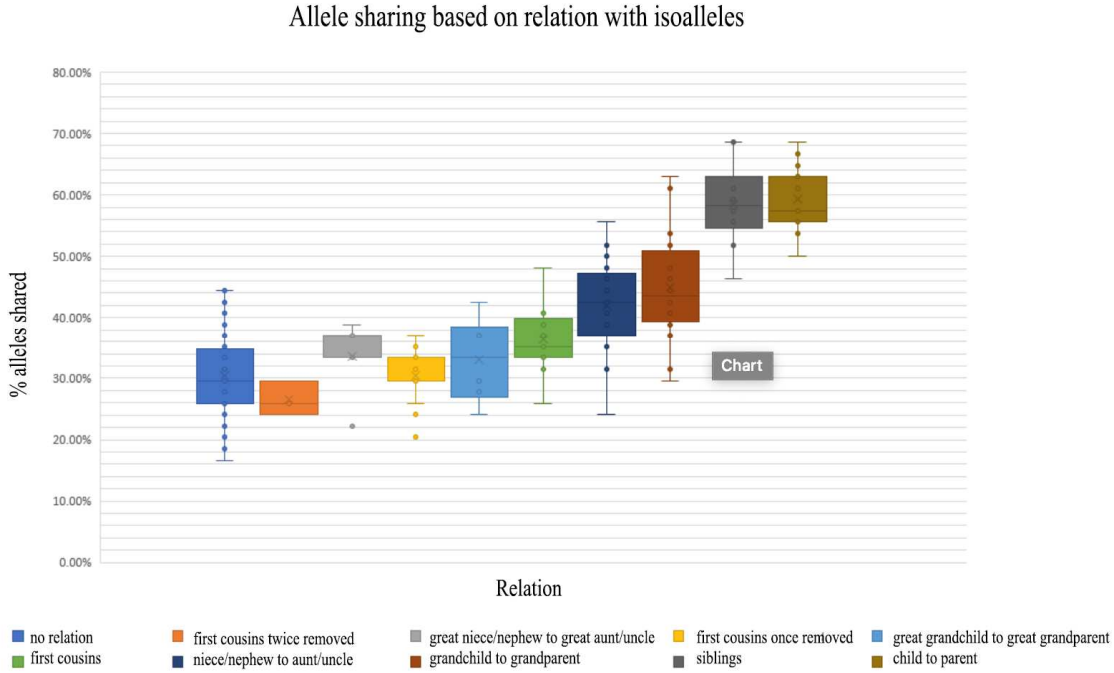

Figure 11. Box-and-whisker plot of allele sharing and sample variation with isoallele data. The median percent of allele sharing and range is based on family relation, as depicted in the pedigree used in this study.

Table 13. T-test for predicting relatedness with and without isoalleles. The statistical significance of including isoalleles in the amount of allele sharing for each of the ten relations was assessed using the t-test with a risk level of 0.5.

Relation	# of pairs	Average (w/o isoalleles)	Std Dev (w/o isoalleles)	Variance (w/o isoalleles)	Average (w/ isoalleles)	Std Dev (w/ isoalleles)	Variance (w/ isoalleles)	t-value	Degree of Freedom	Critical Value
no relation	76	30.97%	6.05%	24.60%	30.36%	6.27%	25.03%	-0.12	150	3.36
siblings	10	60.00%	6.26%	25.02%	58.64%	6.53%	25.56%	-0.26	18	3.92
child/parent	22	59.77%	4.40%	20.98%	59.34%	4.45%	21.09%	-0.09	42	3.54
first cousins	13	37.61%	5.78%	24.04%	36.33%	5.11%	22.61%	-0.26	24	3.75
first cousin once removed	17	32.13%	3.81%	19.52%	30.50%	4.07%	20.18%	-0.36	32	3.62
first cousin twice removed	3	28.40%	2.31%	15.19%	26.54%	2.31%	15.20%	-0.46	4	8.61
niece/nephew/uncle/aunt	17	44.44%	8.48%	29.11%	42.05%	7.80%	27.93%	-0.44	32	3.62
grand-niece/nephew/uncle/aunt	9	35.39%	4.49%	21.19%	34.81%	4.51%	21.23%	-0.12	16	4.02
grand-child/parent	16	46.06%	8.66%	29.42%	44.91%	8.92%	29.86%	-0.21	30	3.65
great grandchild/parent	6	33.33%	6.50%	25.50%	33.03%	6.36%	25.23%	-0.06	10	4.59

4. CONCLUSIONS

MPS technology has made crucial advancements in the field of forensics - increasing the threshold for multiplexing STRs and allowing more areas of discrimination such as identity-informative SNPs and sequence data to be analyzed. Sequencing data adds isoallele information that can assist in distinguishing individuals with similar genotypes while providing a better understanding of changes in stutter patterns with minor nucleotide variations. This study has shown that the stutter ratio for certain loci may not follow a linear correlation with increasing LUS, such as in the case for the pair of isoalleles in D13S317 allele 11. Additionally, three different stutter behaviors have been characterized in D13S317 allele 11, including an N-1 artifact with both an addition and subtraction of repeats within different blocks. A larger sample size of individuals with the D13S317 isoalleles would be needed to confirm the similar ratios and the frequency of the three stutter patterns observed in this study.

By comparing the shared alleles before and after the incorporation of data from isoalleles, it was found that there was an approximately 1.3% average difference in the reported percentages of allele sharing among the ten relation categories. The impact of including isoalleles was not found to have statistical significance in predicting relatedness in all relationship categories based on the pedigree examined. Since the ethnic details for the pedigree are not provided, a much larger sample size from different biogeographic ancestry origins would be helpful in elucidating more accurate averages and variation of stutter patterns for the different biological relationship categories.

Implementing a machine-learning algorithm to predict biological relationships based on the percent of allele sharing may be a useful forensic tool that MPS can expand upon with the increased multiplexing of forensically-relevant biomarkers. As the forensic community continues to implement and validate MPS methodologies into their workflow, it becomes increasingly important to establish a database of identified isoalleles and stutters with unambiguous nomenclature. Being able to repeatably identify these sequence variants and expand the current knowledge of stutter behavior would allow for higher levels of confidence in reporting sequencing results and ultimately, in interpreting STR profiles.

LIST OF JOURNAL ABBREVIATIONS

Biomed Rep	Biomedical Reports
BMC Clin Pathol	BioMed Clinical Pathology
BMC Genetics	BioMed Central Genetics
Bosn J Basic Med Sci	Bosnian Journal of Basic Medical Sciences
Curr Protoc Mol Biol	Current Protocols in Molecular Biology
Forensic Sci Int	Forensic Science International
Forensic Sci Int Genet	Forensic Science International: Genetics
Indian J Microbiol	Indian Journal of Microbiology
Int J Legal Med	International Journal of Legal Medicine
Investig Genet	Investigative Genetics
J Appl Genet	Journal of Applied Genetics
Malays J Med Sci	Malaysian Journal of Medical Sciences
Nucleic Acids Res	Nucleic Acids Research
Philos Trans R Soc Lond B Biol Sci	Philosophical Transactions of the Royal Society of London B: Biological
Sci Rep	Scientific Reports

BIBLIOGRAPHY

1. Butler JM. Forensic DNA Typing: Biology & Technology Behind STR Markers, Academic Press, 2001.

2. 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(1011):1397–412.

3. Lazaruk K, Walsh PS, Oaks F, Gilbert D, Rosenblum BB, Menchen S, et al. Genotyping of forensic short tandem repeat (STR) systems based on sizing precision in a capillary electrophoresis. Electrophoresis 1998;19(January):86-93.

4. Panneerchelvam S, Norazami MN. Forensic DNA profiling and database. Malays J Med Sci 2003;10(July):20-26.

5. Karkar S, Alfonse LE, Grgicak CM, Lun DS. Statistical modeling of STR capillary electrophoresis signal. BMC Bioinformatics 2019;20(December):584.

6. Gavazaj FQ, Mikerezi II, Morina VH, Fatmir AC, Ekrem MB, Gavazaj BB, et al. Optimization of DNA concentration to amplify short tandem repeats of human genomic DNA. Bosn J Basic Med Sci 2012;12(Nov): 236-239.

7. Putkonen MT, Palo JU, Cano JM, Hedman M, Sajantila A. Factors affecting the STR amplification success in poorly preserved bone samples. 2010 Investig Genet;1(October):9.

8. Bieber FR, Buckleton JS, Budowle B, Butler JM, Coble MD. Evaluation of forensic DNA mixture evidence: protocol for evaluation, interpretation, and statistical calculations using the combined probability of inclusion. BMC Genetics 2016;17(August):125.

9. Hu Na, Cong B, Li S, Ma C, Fu L Zhang X. Current developments in forensic interpretation of mixed DNA samples (review). Biomed Rep 2014;2(May):309-316.

10. Butler JM. The future of forensic DNA analysis. Philos Trans R Soc Lond B Biol Sci 2015;370(August): 20140252.

11. Gaag KJ, Leeuw RH, Hoogenboom J, Patel J, Storts DR, Laros JFJ, et al. Massively parallel sequencing of short tandem repeats – population data and

mixture analysis results for the Powerseq™ system. Forensic Sci Int Genet 2016;24(September):86-96.

12. Gill P, Sparkes B, Buckleton JS. Interpretation of simple mixtures of when artefacts such as stutters are present - With special reference to multiplex STRs used by the Forensic Science Service. Forensic Sci Int 1998;95(3):213–24.

13. Brookes C, Bright JA, Harbison S, Buckleton J. Characterizing stutter in forensic STR multiplexes. Forensic Sci Int Genet 2012;6(1):58–63.

14. Gill P, Sparkes R, Kimpton C. Development of guidelines to designate alleles using an STR multiplex system. Forensic Sci Int 1997;89:185–97.

15. Daunay A, Duval A, Laura GB, Buhard O, Renault V, Deleuze J, et al. Low temperature isothermal amplification of microsatellites drastically reduces stutter artifact formation and improves microsatellite instability detection in cancer. Nucleic Acids Research 2019;47(December):e141.

16. Young B, King JL, Budowle B, Armogida L. A technique for setting analytical thresholds in massively parallel sequencing-based forensic DNA analysis. PLoS One 2017;12(May):e0178005.

17. Walsh PS, Fildes NJ, Reynolds R. Sequence analysis and characterization of stutter products at the tetranucleotide repeat locus vWA. Nucleic Acids Res 1996;24(14):2807–12.

18. Schumm JW, Bacher JF, Hennes LF, Gu T, Micka KA, Sprecher C, et al. Pentanucleotide repeats: highly polymorphic genetic markers displaying minimal stutter artifact. 1998(Sept).

19. Woerner AE, King JL, Budowle B. Flanking variation influences rates of stutter in simple repeats. Genes (Basel) 2017;8(November):329.

20. Schlotterer C, Tautz D. Slippage synthesis of simple sequence DNA. Nucleic Acids Res 1992;20(2):211–5.

21. Fazekas AJ, Steeves R, Newmaster SG. Improving sequencing quality from PCR products containing long mononucleotide repeats. Biotechniques 2018;48(April).

22. Klintschar M, Wiegand P. Polymerase slippage in relation to the uniformity of tetrameric repeat stretches. Forensic Sci Int 2003;135:163–6.

23. Barba M, Czosnek H, Hadidi A. Historical perspective, development and applications of next-generation sequencing in plant virology. Viruses 2014;6(January):106-136.

24. Arsenic R, Treue D, Lehmann A, Hummel M, Dietel M, Denkert C, et al. Comparison of targeted next-generation sequencing and Sanger sequencing for detection of PIK3CA mutations in breast cancer. BMC Clin Pathol 2015;(15)(November):20.

25. Knijff P. From next generation sequencing to now generation sequencing in forensics. Forensic Sci Int Genet 2019;38(January):175-180.

26. Pareek CS, Smoczynski R, Tretyn A. Sequencing technologies and genome sequencing. J Appl Genet 2011;52(November):413-435.

27. Slatko BE, Gardner AF, Ausubel FM. Overview of next generation sequencing technologies. Curr Protoc Mol Biol 2018;122(April):e59.

28. Kulzki JK. Next-generation sequencing – an overview of the history, tools, and "omic" applications. Next Generation Sequencing – Advances, Applications and Challenges 2015.

29. Yang Y, Xie B, Yan J. Application of next-generation sequencing technology in forensic science. Genomics Proteomics Bioinformatics 2014;12(October):190-197.

30. Verogen Inc. FBI approves Verogen's Next-Gen forensic DNA technology for National DNA Index System. Business Wire. 2019.

31. Almalki N, Chow HY, Sharma V, Hart K, Siegel D, Wurmbach E. Systematic assessment of the performance of Illumina's MiSeq FGx ™ forensic genomics system. Electrophoresis 2017;38:846–54.

32. Jäger AC, Alvarez ML, Davis CP, Guzmán E, Han Y, Way L, et al. Developmental validation of the MiSeq FGx Forensic Genomics System for Targeted Next Generation Sequencing in Forensic DNA Casework and Database Laboratories. Forensic Sci Int Genet 2017;28:52–70.

33. Xavier C, Parson W. Evaluation of the Illumina ForenSeq^{TM} DNA Signature Prep Kit – MPS forensic application for the MiSeq FGx TM benchtop sequencer. Forensic Sci Int Genet 2017;28:188–94.

34. Verogen Inc. ForenSeq^{TM} DNA Signature Prep Reference Guide. 2018;(Document #VD2018005 Rev. A).

35. Verogen Inc. ForenSeq Universal Analysis Software Guide. 2018;(Document #VD2018007 Rev. A).

36. Verogen Inc. MiSeq FGx™ Instrument Reference Guide. 2018;(Document #VD2018006 Rev. A).

37. Illumina Inc. Indexed Sequencing Overview Guide. 2019;(Document #15057455 v05).

38. Köcher S, Müller P, Berger B, Bodner M, Parson W, Roewer L, et al. Inter-laboratory validation study of the ForenSeq™ DNA Signature Prep Kit. Forensic Sci Int Genet 2018;36(March):77–85.

39. Moreno LI, Galusha MB, Just R. A closer look at Verogen's Forenseq™ DNA Signature Prep kit autosomal and Y-STR data for streamlined analysis of routine reference samples. Electrophoresis 2018;39(21):2685–93.

40. Ambardar S, Gupta R, Trakroo D, Lal R, Vakhlu J. High throughput sequencing: an overview of sequencing chemistry. Indian J Microbiol 2016;56(December):394-404.

41. Liu Z, Gao L, Zhang J, Fan Q, Chen M, Cheng F, et al. DNA typing from skeletal remains: a comparison between capillary electrophoresis and massively parallel sequencing platforms. Int J Legal Med. 2020;(6).

42. Guo F, Yu J, Zhang L, Li J. Massively parallel sequencing of forensic STRs and SNPs using the Illumina® Forenseq™ DNA Signature Prep Kit on the MiSeq FGx™ Forensic Genomics System. Forensic Sci Int Genet 2017;31(November):135-148.

43. Sharma V, Plaat DA, Liu Y, Wurmbach E. Analyzing degraded DNA and challenging samples using the Forenseq™ DNA Signature Prep. Science & Justice 2020;60(May):243-252.

44. Gettings KB, Borsuk LA, Ballard D, Bodner M, Budowle B, Devesse L, et al. STRSeq: A catalog of sequencing diversity at human identification short tandem repeat loci. Forensic Sci Int Genet 2017;31(August):111-117.

45. Parson W, Ballard D, Budowle B, Butler JM, Gettings KB, Gill P, et al. Massively parallel sequencing of forensic STRs: Considerations of the DNA commission of the International Society for Forensic Genetics (ISFG) on minimal nomenclature requirements. Forensic Sci Int Genet 2016;22(May):54-63.

46. Zhang S, Niu Y, Bian, Y, Dong R, Liu X, Bao Y, et al. Sequence investigation of 34 forensic autosomal STRs with massively parallel sequencing. Sci Rep 2018;8(May): 6810.

47. Vilsen SB, Tvedebrink T, Eriksen PS, Bøsting C, Hussing C, Smidt H, et al. Stutter analysis of complex STR MPS data. Forensic Sci Int Genet 2018;35(April):107–12.

48. Li R, Wu R, Li H, Zhang Y, Peng D, Wang N, et al. Characterizing stutter variants in forensic STRs with massively parallel sequencing. Forensic Sci Int Genetic 2020;45(March):102225.

49. Qiagen. EZ1® DNA Investigator Handbook 2014. Available from: https://www.qiagen.com/us/resources/resourcedetail?id=46064856-1b88-4b27-a825-d3f616e06c08&lang=en. Access date: 1 Nov 2019.

50. Qiagen. EZ1® Advanced XL User Manual (9001874 Rev. R2) 2017;(Publication no. 9001874). Available from: https://www.qiagen.com/us/resources/resourcedetail?id=c9ecd500-147b-4a8e-ae71-3dc86cd3d17a&lang=en. Access date: 1 Nov 2019.

51. Applied Biosystems. Quantifiler™ Duo DNA Quantification Kit User Guide (4391294 Rev. F). 2018;(Publication no. 4391294). Available from: https://www.thermofisher.com/order/catalog/product/4387746?SID=srch-srp-4387746. Access date: 1 Nov 2019.

52. Grgicak CM, Urban ZM, Cotton RW. Investigation of reproducibility and error associated with qPCR methods using Quantifier™ Duo DNA Quantification Kit. J Forensic Sci 2010;55(September):1556.

53. Faith S. PopSeq: The human STR sequence diversity database [Internet]. Ishinews.com. 2017 [cited 2020 May 20]; Available from: https://www.ishinews.com/popseq-human-str-sequence-diversity-database/

54. Dai W, Pan Y, Sun X, Wu R, Li L, Yang D. High polymorphism detected by massively parallel sequencing of autosomal STRs using old blood samples from a Chinese Han population. Sci Rep 2019;9(December):18959.

55. Gettings KB, Borsuk LA, Steffen CR, Kiesler KM, Vallone PM. Sequence-based U.S. population data for 27 autosomal STR loci. Forensic Sci Int Genet 2018;37:106-115.

CURRICULUM VITAE