

2021

# Analysis of multi-generational father-son pairs using a YFiler Plus PCR amplification kit and a ForenSeq DNA signature prep kit

---

<https://hdl.handle.net/2144/43353>

*"Downloaded from OpenBU. Boston University's institutional repository."*

BOSTON UNIVERSITY  
SCHOOL OF MEDICINE

Thesis

**ANALYSIS OF MULTI-GENERATIONAL FATHER-SON PAIRS USING A  
YFILER PLUS PCR AMPLIFICATION KIT AND A FORENSEQ DNA  
SIGNATURE PREP KIT**

by

**MARGO FOLWICK**

B.S., West Virginia University, 2017

Submitted in partial fulfillment of the  
requirements for the degree of  
Master of Science

2021

© 2021 by  
MARGO FOLWICK  
All rights reserved

Approved by

First Reader

---

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

Second Reader

---

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

Third Reader

---

Kathryne Hall, M.S.  
Criminalist, Boston Police Crime Laboratory

## **ACKNOWLEDGMENTS**

I would like to thank my amazing thesis advisor, Dr. Robin Cotton, for providing invaluable guidance and support throughout this research. Thesis research did not always go according to plan, but she was always there to offer suggestions and plans of attack. I would also like to thank the rest of the faculty of the Boston University Biomedical Forensic Sciences program. I am grateful to have become a more well-rounded student and individual, thanks to the passion of those teaching the courses and the diverse selection of forensic science courses.

I am extremely grateful to my parents for their love, patience, and sacrifices in preparing me for my future. Their encouragement and support have been invaluable when research setbacks would occur. I could always count on my mom to discuss ideas and protocols with me, and my dad to offer advice.

I would also like to thank my friends for their love and support during these past couple years. You all kept me sane.

**ANALYSIS OF MULTI-GENERATIONAL FATHER-SON PAIRS USING A  
YFILER PLUS PCR AMPLIFICATION KIT AND A FORENSEQ DNA  
SIGNATURE PREP KIT**

**MARGO FOLWICK**

**ABSTRACT**

Y-chromosome testing has become more prevalent in recent years as a means of identifying forensic samples using STRs or identifying biomarkers for disease or determining geographic origins of populations. Additionally, Y-chromosome analysis is especially useful in paternity testing as the Y chromosome is inherited paternally and the male-specific region of the Y chromosome does not undergo any recombination events, allowing the genotypic data of both the father and son to be identical. Though in most cases a father-son pair will have the same Y-allelic data, random mutations like allele insertions and deletions can occur, which can interfere and result in incorrect conclusions in regards to paternity testing, forensic analysis, or genealogy. Though the exact mechanism of Y loci mutability is unknown, postulations of factors that can cause mutations have been studied, as well as attempts to determine mutation rate specific to each locus.

A multi-generational pedigree consisting of 9 males was analyzed using two different methodologies: capillary electrophoresis and next-generation sequencing. The samples were amplified using either a ForenSeq™ Signature DNA Prep Kit (Verogen, San Diego, CA) or a YFiler™ Plus PCR Amplification Kit (Thermo Fisher Scientific, Waltham, MA). Between the two methods, five Y-STR loci were identified as being

discordant between a father-son pair. Next-generation sequencing identified an allele insertion at DYS385a/b, resulting in a potential tri-allelic locus, but was disproved after comparison with the capillary electrophoresis data of the sample. The capillary electrophoresis data identified four discordances between father-son pairs, one of which was an allele mutation with a gain of a repeat at DYS458. At DYS 389II, an allele insertion was identified, but was contradicted after comparison with the next-generation sequencing data. There was a potential null allele at DYS518 and either an OL variant allele or a 2 base pair deletion at DYS481. Following peak height ratio, stutter, and comparative analysis between the genotypic data of the two analysis methods, two of these discordances were proven to be errors, one was a definitive mutational event, and the other two could neither be confirmed nor denied due to differences in loci tested in each kit.

## TABLE OF CONTENTS

ACKNOWLEDGMENTS .....	iv
ABSTRACT.....	v
TABLE OF CONTENTS.....	vii
LIST OF TABLES .....	ix
LIST OF FIGURES .....	x
LIST OF ABBREVIATIONS.....	xi
1. INTRODUCTION .....	1
1.1 Structure and Function of DNA.....	1
1.2 Y Chromosome Structure and Function .....	3
1.3 History of DNA Analysis.....	5
1.3.1 Capillary Electrophoresis.....	8
1.3.2 Next-Generation Sequencing.....	11
1.4 Next-Generation Sequencing with MiSeq FGx™ System .....	13
1.5 Objective .....	17
2. METHODS .....	18
2.1 Capillary Electrophoresis Workflow .....	19
2.2 Next-Generation Sequencing Workflow.....	21
3. RESULTS .....	25
3.1 Next-Generation Sequencing Results .....	25

3.2 Capillary Electrophoresis Results .....	29
4. DISCUSSION .....	32
5. CONCLUSION.....	36
6. FUTURE DIRECTIONS .....	38
REFERENCES .....	39
CURRICULUM VITAE.....	44

## LIST OF TABLES

Table	Title	Page
1	Y-STR Genotype Results from Sequencing	26
2	Comparison of Allele Read Counts of DYS385a/b	28
3	Y-STR Genotype Results from Capillary Electrophoresis	31

## LIST OF FIGURES

Figure	Title	Page
1	The Y Chromosome	5
2	Multi-generational Family Pedigree	18
3	Capillary Electrophoresis Workflow	21
4	Sequencing Workflow	24
5	Alleles Detected at DYS385a/b	27
6	Distribution of Peak Height Ratios for DYS385a/b	28
7	Y-STR Allele Comparison	30

## LIST OF ABBREVIATIONS

A	Adenine
ARC	Allele Read Count
AT	Analytical threshold
bp	Base pairs
C	Cytosine
CCD	Charge-coupled device
CE	Capillary Electrophoresis
ddNTP	Dideoxynucleotide Triphosphates
DNA	Deoxyribonucleic Acid
dNTP	Deoxynucleotide Triphosphates
EPG	Electropherogram
G	Guanine
GBY	Gonadoblastoma
IT	Interpretation threshold
MPS	Massively parallel Sequencing
MSY	Male Specific Region
NDIS	National DNA Index System
NGS	Next Generation Sequencing
NRY	Non-recombining Y
OL	Off Ladder

PAR	Pseudoautosomal Region
PCR	Polymerase Chain Reaction
PHR	Peak Height Ratio
RFLP	Restriction Fragment Length Polymorphism
RNA	Ribonucleic Acid
SNP	Single Nucleotide Polymorphism
ST	Stochastic Threshold
STR	Short Tandem Repeat
SWGDM	Scientific Working Group on DNA Analysis Methods
T	Thymine
UAS	Universal Analysis Software
VNTR	Variable Number of Tandem Repeats

# 1. INTRODUCTION

## 1.1 Structure and Function of DNA

Deoxyribonucleic acid (DNA) is a molecule that acts as a blueprint for all living organisms, dictating everything from external appearance to cellular processes to propensity for diseases. It encodes this genetic information, which is passed from parents to offspring. Though it is often taught and believed that DNA was first discovered in the 1950s, Johann Friedrich Miescher was the one to first extract and purify human DNA from leukocytes [1]. He continued his DNA research with salmon spermatozoa, eventually beginning to theorize that DNA might be a contributor to fertilization [1]. However, it was not until 1953 when the structure of DNA was determined to be that of a double helix by James Watson and Francis Crick [2].

DNA's structure resembles that of a coiled ladder. The sides of the ladder are comprised of phosphate groups and deoxyribose sugars, while the rungs of the ladder are comprised of pairs of nucleotides using the nucleotides Adenine (A), Cytosine (C), Guanine (G), and Thymine (T). Nucleotides from each strand pair in a specific manner, A pairing with T and C pairing with G. These paired nucleotides are held together with hydrogen bonds, with A-T pairs utilizing two hydrogen bonds, and C-G pairs utilizing three [3].

The DNA molecules must be organized in order for DNA replication and cell division to occur. In eukaryotes, DNA molecules are condensed into thread-like structures called chromosomes, which are able to undergo replication and separation in mitosis or meiosis. The human genome consists of 23 chromosome pairs, with one

chromosome of every pair coming from each parent. Of these 23 pairs, 22 pairs are autosomes, and one pair is sex-determining chromosomes. The sex chromosomes are identified as being either “X” or “Y,” whereas the autosomes are labeled with a traditional numeric system (i.e. chromosome 1, chromosome 2, etc...).

Within DNA, there are non-coding and coding regions. Coding regions of DNA dictate and facilitate the production of various proteins through translation and transcription. Non-coding regions of DNA function as regulators of gene expression, or may produce functionally important ribonucleic acids (RNA) [4]. The human genome consists mainly of non-coding regions with coding regions interspersed throughout [5]. Of the five percent of the human genome that has been discovered to be highly conserved through time, four percent consists of non-coding DNA sequences [4].

Though they do not encode useful proteins, non-coding regions are extremely valuable for other reasons. They contain repeating nucleotide sequences that are highly polymorphic, known as tandem repeats. Although coding regions can also contain repetitive sequences, most of the tandem repeats occur within non-coding regions [4]. There are thousands of different tandem repeat polymorphisms, which differ by sequence length and complexity. Minisatellites, also known as variable number tandem repeat (VNTR) sequences have repeat unit lengths anywhere from 9-64 base pairs (bp) [4]. Microsatellites, known as short tandem repeats (STR), have repeat unit lengths between 2-6 bp [4]. Besides length polymorphisms, there are also sequence polymorphisms. Single nucleotide polymorphisms (SNP) occur when a single nucleotide differs between two or more individuals' sequences (i.e. AGT and AGA).

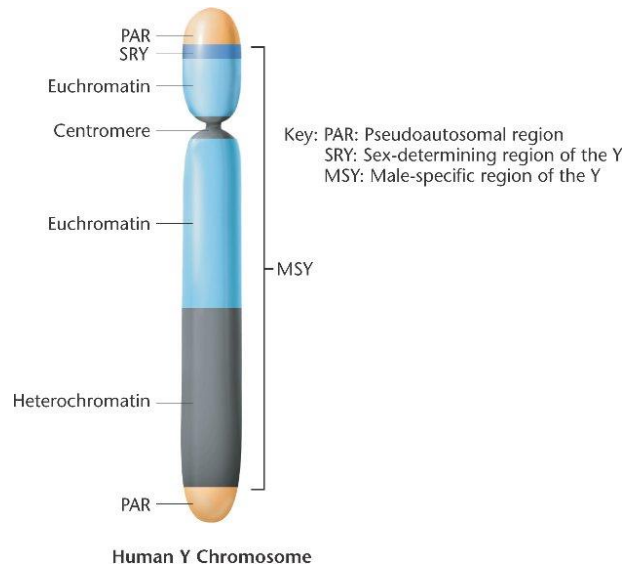
The specific chromosomal locations at which given polymorphisms occur are known as loci, and each variation is an allele. If a locus is highly polymorphic, many different alleles may be present in a population. For example, a repeat unit of (ATTA) may be repeated at a locus five times, [ATTA]<sub>5</sub>, designated as an allele 5 at that locus. However, there may be a number of alleles having different numbers of repeats, e.g. 3-12, in a population of individuals. The repeated sequence length of these alleles varies from dinucleotide repeats, e.g. CT, to pentanucleotide repeats, e.g. CTCAT, though many STR loci have repeat lengths of four nucleotides. In forensic DNA analysis, many highly polymorphic STR loci are analyzed to identify or exclude individuals based on the allele combination that comprise their genotype.

## **1.2 Y Chromosome Structure and Function**

The Y chromosome is one of the sex-determining chromosomes, as well as one of the smallest chromosomes in the human genome [5]. The Y chromosome's centromere divides the chromosome into two arms, with one being quite longer than the other. The short arm is designated Yp and the long arm is designated Yq. At either end of Yp and Yq are pseudoautosomal regions (PAR), PAR1 and PAR2, respectively [5]. Both of the PARs contain homologous genes that can recombine with the X chromosome during meiosis. The male specific region (MSY), also known as the non-recombining Y (NRY) region, is located between PAR1 and PAR2 [5,6]. In the MSY, there is a euchromatin region consisting of Yp, the centromere, and the proximal region of Yq [5]. The distal region of Yq contains heterochromatin, which is presumed to be polymorphic in length

due to its composition of two sequences with thousands of repeats each [5]. Despite its polymorphism, using VNTRs to analyze the heterochromatic region would be highly inefficient for identification purposes. However, the euchromatic region has over 400 STR loci that have been identified, though not all have been adequately studied [6]. Some of these STR regions are being used in current STR kits for human identification. However, these conventional STR kits cannot differentiate between members of the same paternal line. Certain Y-STRs have been identified with having high mutation rates, termed rapidly-mutating (RM) Y-STRs, which can differentiate samples within the same paternal line, or unrelated samples that happen to have an identical genotype due to mutation [7].

The Y chromosome has several functionally important biological roles. Firstly, the Y chromosome is inherited paternally, meaning that in an uninterrupted paternal line, the Y chromosome will be the same, barring mutation, insertion, or deletion. During meiosis, the PARs undergo recombination with the homologous sequences present on the X chromosome. In addition to this, the genes contained in the MSY are essential for male development. In 1990, the SRY gene was identified as encoding a protein responsible for gene expression of testis differentiation and development [8, 9]. In addition to the SRY gene, at least 50 other transcribed genes on the Y chromosome have been determined to be critical to male development.



**Figure 1. The Y Chromosome.** The human Y chromosome is one of the smallest chromosomes in the human genome. It is comprised of recombining regions PAR, heterochromatin which has two highly repeated sequences, and euchromatin, which is where over 400 STRs are located, including the SRY gene responsible for testis development. [8]

### 1.3 History of DNA Analysis

Since the discovery of the double-helical structure of DNA by Watson and Crick in 1953, various studies and experiments have been performed to elucidate the structure and sequence of the nucleotides in DNA, in addition to its potential applications [2]. The first widely adopted sequencing method was developed by Allan Maxam and Walter Gilbert, which employed chemically-treated radiolabeled DNA that would cleave the chain at certain bases [10, 11]. Once cleaved, these fragments were run on a polyacrylamide gel, which was used to determine the length of the fragments, and thus the sequence could be inferred [10]. In the late 1970s, Frederick Sanger introduced a chain termination method of sequencing DNA utilizing dideoxynucleotide triphosphates (ddNTPs), which became known as Sanger Sequencing [12]. In the 1980s, geneticist Sir Alec Jeffreys used restriction length polymorphisms (RFLPs) as a means of DNA

fingerprinting [13]. This technique utilized a restriction enzyme, which would cleave the DNA at a specific nucleotide sequence. Then, the resulting fragments would be separated using agarose gel electrophoresis. These separated DNA fragments were then transferred to a membrane using a Southern blot. The DNA fragments which were polymorphic, could be identified by hybridization to a  $^{32}\text{P}$  labeled DNA probe of the same sequence and the length of the resulting gel bands could be compared [13]. This technique was used shortly thereafter in 1986 to test semen and tissue samples from two victims, and a blood sample from the alleged killer [14]. Using the RFLP analysis, Jeffreys determined that the semen from both victims was identical; however, the resulting DNA fingerprint from the semen samples did not match that from the blood sample of the suspect [14]. This was the first use of DNA analysis on evidence samples. In 1985, a procedure was also developed by Gill et al. to separate epithelial and sperm cell mixtures into two fractions: a non-sperm-cell fraction and a sperm-cell fraction [15]. This technique became known as differential extraction and is extremely useful in sexual assault cases, where the profile of the victim and suspect can be separated and analyzed individually.

In 1985, the polymerase chain reaction (PCR) was invented and published by Kary Mullis as a means of amplifying specific targeted sequences of double-stranded DNA [16]. Prior to this, analysis of DNA in any context would require a larger sample amount in order to produce useful DNA related data. Using PCR, a small quantity of DNA could be analyzed and be used to produce a complete DNA profile. This method employs the use of a template DNA strand, a thermostable DNA polymerase, oligonucleotide primers, and deoxynucleotide triphosphates (dNTPs) [17]. The template

strand is first denatured at high heat before the forward and reverse primers identify the region that will be copied by the polymerase. The dNTPs supply the four nucleotide bases (A, C, T, G) for the DNA polymerase, which will incorporate the dNTPs when copying the template strand [17]. In the mid to late 1990s, PCR became more widely used for STR testing and began to replace VNTR testing, due to the fact that VNTR sequences were much longer than those of STRs and thus difficult to amplify [18]. Because of the smaller size of the PCR products when using STRs, DNA can be effectively recovered from degraded samples [19]. PCR amplification of multiple STR loci can occur simultaneously with the incorporation of different fluorescent dyes into the primers [19]. The amplification of multiple loci in a single PCR reaction is known as multiplexing.

In the early 1990s, capillary electrophoresis (CE) was introduced, which aimed to replace gel electrophoresis as the traditional method of separating and visualizing STR fragments [20]. With traditional gel electrophoresis, run time can take as long as a few hours to achieve visualization of separated bands, and depending on the staining solution, the investigator could be subject to carcinogenic or mutagenic chemicals. Capillary electrophoresis bypasses the use of a staining solution, and gives results quickly while providing a high resolution [21]. Since its emergence, CE has been used to separate and visualize PCR-amplified STR loci for forensic DNA analysis as well as non-forensic DNA sequencing applications.

With the growth of STR analysis for forensic application, the UK launched the first National DNA Database in 1995, and the USA developing its National DNA Index

System (NDIS) in 1998 with 13 core loci [22]. Soon after, commercial kits for autosomal STR testing became available, with STR loci chosen based on their chromosomal location, discriminating power, length of alleles, and low rates of stutter or other artifacts [23]. Despite the wide-spread use and reliability of CE for forensic DNA analysis, DNA sequencing is becoming relevant in forensic science applications.

### **1.3.1 Capillary Electrophoresis**

Before CE is performed, DNA analysis begins with extraction of DNA from samples, followed by quantification of the amount of human DNA in the sample using real-time PCR [24]. Quantification ensures that the target amount of human DNA is being added to the PCR reaction. Samples may have to be diluted or concentrated post-quantification if they contain too much or too little DNA, respectively. Targeting a specific amount of DNA for PCR is a preventative measure against artifacts that can result with low DNA, such as allelic dropout or imbalance, or high amounts of DNA, such as pull-up or off-scale peaks.

Capillary electrophoresis begins with denaturation of the amplified PCR products into single-stranded DNA, either with formamide or rapid heating and subsequent cooling on ice [23]. Following this, electrokinetic injection is performed: a voltage is applied for a certain amount of time, causing the DNA molecules to be drawn into the capillary. The aforementioned formamide must have a low conductivity, so as to not interfere with the electrokinetic injection; by-products of formamide decomposition can interfere with resolution and sensitivity [25]. Once the DNA is in the capillary, an electric current is

applied, causing the DNA fragments to travel from one end of the capillary to the other at different speeds dependent on their size. Small fragments of DNA travel faster, and thus will move past the CE's detection window before larger fragments.

The use of multiple dye colors attached to one of two PCR primers for each locus permits multiwavelength detection, and thus allows multiplexing in forensic DNA analysis. When detecting the DNA fragments as they pass through the capillary, an argon-ion or other type of laser excites the fluorescently-labelled molecules, which then emit light. A charged-coupled device (CCD) camera will then detect the emitted light and determine which dye color is present and the relative fluorescent intensity. As long as fragments of DNA of different loci which are similar in size are labelled with different dye colors, they can be separated and analyzed. The data captured by the CCD camera is presented visually in a plot of DNA fragment sizes known as an electropherogram (EPG). The fragments are visualized as peaks on the EPG, and the peak heights correspond to the light intensity detected by the CCD camera, which is referred to as relative fluorescent units (RFUs). High concentrations of fragments increase fluorescence, which increases the RFUs detected, leading to increased peak heights on the EPG.

In addition to RFUs being detected, the CCD camera records the amount of time a DNA fragment takes to travel through the capillary to the detection window. This time is compared to a size standard, which informs the computer of the time necessary for various-sized fragments to travel through the capillary. The elapsed times of the size standard fragments are compared to those of the sample fragments, in order to determine the size in base pairs (bp) of the sample DNA fragments. The time necessary for the size

standard fragments to travel through the capillary are also compared to the DNA fragments of the allelic ladder, which is a collection of known alleles. Through this comparison, the size in bp of the sample fragments is converted to allele repeat number.

Once the alleles are identified, the EPG needs to be analyzed to ensure genotyping is accurate. Thresholds such as stochastic threshold (ST) and analytical threshold (AT) have been established to ameliorate STR profile interpretation. When peaks are above AT, peaks are determined to be above baseline and not in the range of instrument noise. When peaks are above ST, it is assumed that drop-out of a sister allele of a heterozygous locus has not occurred; if one peak is present, homozygosity is assumed [26]. Even with thresholds in place, STR profile interpretation can be difficult due to the potential for multi-contributor mixture samples in forensic DNA analysis. With considerable differences in peak heights within alleles in the same locus, it may not be possible to differentiate the genotypes of contributors to the sample. In addition to this, when analyzing low-level DNA samples with new STR kits that have high sensitivity, the expected peak-height ratio (PHR) for single-source samples may not be satisfied [26].

There are many artifacts that can interfere with STR profile interpretation and cause genotyping to be difficult. Examples of artifacts typically found on an EPG include stutter, pull-up, spikes, PH imbalance, and allele drop-out. Stutter occurs during PCR, and has been proposed to be due to slipped strand mispairing, where the template DNA strand loops out and typically results in a deletion of one repeat unit on the new strand [27]. Stutter can also occur as an insertion of one repeat unit on the new strand, or as a deletion of two repeat units [27]. Stutter is one of the most difficult artifacts in STR

interpretation, as stutter products can mask true alleles either from a minor contributor in a mixture, or in low-level/quality samples [26]. Besides stutter, allele drop-out and PH imbalance are notorious for complicating EPG interpretation. When alleles drop-out, the correct genotype cannot be determined, so a heterozygous locus could be interpreted as a homozygous locus, leading to an incorrect exclusion of a contributor. When PH imbalance occurs at multiple loci, a single-source profile may be interpreted as a mixture sample. This could lead to a heterozygous locus being interpreted as a homozygous locus of a major contributor with an additional allele of a minor contributor. Both allele drop-out and PH imbalance can be attributed to PCR errors such as differential amplification across loci, or can occur with contaminated or low quality/quantity samples.

Spikes and pull-up are attributed to CE errors, either when different dye colors cannot be completely discriminated or when oversaturation leads to bleed-through of dyes into other colors. Though it can be more readily identified than other artifacts through analysis of peak position across the color spectrum, pull-up can resemble an allele of a minor contributor of a mixture. Despite numerous artifacts complicating STR profile interpretation, guidelines and protocols have been established by the Scientific Working Group on DNA Analysis Methods (SWGDM) to standardize and improve the analysis of EPGs.

### **1.3.2 Next Generation Sequencing**

Though CE has been known as the gold standard of forensic DNA analysis since the 1990s, the use of DNA sequencing is slowly entering into the forensic field. CE-based

STR analysis is a reliable and discriminating technique, allowing up to 24 STR loci to be multiplexed, but is limited in that it can only separate alleles by difference in length. Next-generation sequencing (NGS), also known as massively parallel sequencing (MPS), has a higher multiplexing potential than CE-based methods, as NGS allows sequencing of thousands of genomic regions in one reaction. Its predecessor, Sanger Sequencing, has all but been replaced by NGS due to its time efficiency and its sensitivity. In 2008, NGS accurately sequenced a whole human genome in about 8 weeks, from building the consensus sequence, to analyzing and ensuring accuracy, to determining genotypic data for the individual [28]. NGS technology has advanced in such a way that the time needed to sequence a whole human genome has been reduced from months to days [29].

For STR loci, NGS allows for a higher discrimination power than CE-based methods not only due to the fact that a larger number of markers can be processed in parallel, but also because alleles can be separated both by difference in length and difference in sequence [28]. NGS has resulted in numerous previously unidentified alleles being discovered [30]. The report by Gettings et al. identifies at least four STR loci as having sequence variants [30]. For example, the vWA locus has a sequence variant consisting of two SNPs occurring at either the 14 or 15 allele which interrupts the standard repeat pattern [30]. NGS analysis is not limited to only the STR repeat region of the PCR product, but also includes the surrounding flanking regions [31]. The report by Gettings et al. identifies 16 STR loci as having sequence variants in the flanking regions, with 32 SNPs and 8 insertion-deletions [30]. Previous data has indicated that the length of the longest uninterrupted repeat stretch of an allele is related to that allele's stutter

ratio [31]. Through NGS and the additional flanking region sequences, stutter behavior could be better described and predicted [32].

An additional advantage of NGS over CE-based methods is that sequencing is not as constrained. NGS is not reliant on fluorescent dye detection by electrophoretic systems and thus thousands of loci, both STRs and SNPs can be amplified simultaneously for forensic applications [31]. This could eventually lead to the adoption of co-amplification of both autosomal and Y STRs, which could provide higher discrimination power for mixture samples [31]. The analysis of SNPs allows for the target of a large number of markers with very low quantities of DNA. NGS can also identify microhaplotypes, sets of two or more SNPs in close proximity on a chromosome, with three or more allelic combinations [33]. These have shown promise in the forensic identification field and in analysis of DNA mixtures [31, 33].

#### **1.4 Next Generation Sequencing with MiSeq FGx™ System**

The MiSeq FGx™ system (Illumina, San Diego, CA) was released in 2015 as an instrument developed specifically for forensic genomics, and the first instrument to analyze both STRs and SNPs in a single run. The MiSeq FGx™ employs the sequencing by synthesis (SBS) method, which utilizes fluorescently labelled ddNTPs on clonally amplified DNA fragments [34]. The amplified DNA fragments are immobilized on a flow cell and bridge amplification occurs, generating hundreds of copies in close proximity, known as cluster generation [35]. With each cycle, the four ddNTPs are washed over the flow cell, but only one fluorescently labelled ddNTP is incorporated per cluster, per cycle [35]. A CCD captures four images after each cycle incorporation, and thus detects the ddNTP incorporated at each

cluster, since they all emit light at different wavelengths [35]. Even though an image is taken for each ddNTP, the emitted light is recorded as one signal, so each base-call is determinant on the emission wavelength and intensity [35].

The MiSeq FGx™ performs 4 reads for each sample (Read 1, Index 1, Index 2, Read 2) for a total of 398 sequencing cycles. Read 1 runs for 351 cycles, and is responsible for sequencing the first 351 nucleotides of the template strands. Index 1 and Index 2 are both 8 cycles and are responsible for sequencing the sample's i7 index and i5 index, respectively [35]. There are 12 different i7 index sequences and 8 different i5 index sequences which allow up to 96 samples to have a unique combination of indices. These index combinations allow the samples to be multiplexed and later differentiated in software. Read 2 is 31 cycles, and is responsible for sequencing 31 bases in the reverse direction to read 1 [35]. Having a short read in the reverse direction is useful for ensuring correct sequence alignment, while also being time efficient.

Following the sequencing process, the generated data is subsequently analyzed in Verogen's ForenSeq™ Universal Analysis Software (San Diego, CA). One of the most important functions of the UAS is to provide quality metrics for each run that was performed. Quality metrics that are provided include cluster density, clusters passing filter, phasing, pre-phasing, and information regarding positive and negative controls [35]. The cluster density is the average number of clusters per square millimeter of the flow cell, with the target density being in the range of 400-1650 thousand clusters [35]. Phasing and pre-phasing are the percentages of molecules in clusters that run behind the current cycle or run ahead of the current cycle, respectively. The chastity filter removes clusters of poor quality that are due to over clustering, poor amplification, or poor sequencing [35]. Chastity is the ratio of the

brightest base intensity divided by the sum of the brightest and second brightest base intensities. Clusters will pass the filter if less than 1 base call has a value of 0.6 in the first 25 cycles. Once clusters pass the chastity filter, they are converted into base calls and given quality scores. Some of the aforementioned quality metrics can still yield reliable results if the values are outside the desired parameters. In such instances, other indications of success can be useful, including the correct genotype calls for the human sequencing control (HSC) and recording the expected minimum intensity level [35]. In the case of the MiSeq FGx™ software, the intensity level of fluorescence is measured by the allele read count (ARC), the number of times the allele was detected by the CCD camera, rather than RFU in CE-based methods.

In addition to analyzing the images recorded by the CCD camera, calling bases, and providing quality metrics regarding each run, the UAS provides guidelines regarding thresholds. Similar to CE-based methods, the UAS has two thresholds: AT and an interpretation threshold (IT). The AT of the UAS is identical to that of CE, it represents the lower limit of detection, with anything below the AT being considered noise and thus not being called an allele. The IT of the UAS is likewise identical to the ST of CE; it represents a threshold above which drop-out of a sister allele of a heterozygous locus is considered unlikely. If an allele is above the IT, it is considered a homozygous locus, regardless of other alleles being present but below IT. The main difference between UAS thresholds and those of CE-based methods is that CE-based methods utilize specific RFU values across all loci, whereas the UAS thresholds are defined as percentages. The typical AT and ST of CE-based methods fall between 30-50 RFU and 150-200 RFU, respectively; these values are estimates, since each laboratory has specific validated threshold values that it uses to analyze STR

profiles and the values are dependent on the type of CE platform. Because the UAS utilizes percentages as its threshold values, the AT and IT will vary from sample to sample and between loci. For the UAS, most autosomal and X-STRs have defined the AT as 1.5% of reads and the IT as 4.5% of reads. For Y-STRs, however, although most thresholds are also 1.5% of reads for AT and 4.5% of reads for IT, a few thresholds are different. For locus DYS389II, the AT and IT are 5% of reads and 15% of reads, respectively, and for DYS448 and DYS635, the AT and IT are 3.3% of reads and 10% of reads, respectively. In addition to these thresholds, both CE-based methods and the UAS utilize percentages for filtering stutter, which vary across all loci. The aforementioned UAS thresholds and filters are preset on the software and have been internally validated by Verogen, though these thresholds can be manipulated according to a laboratory's specific validated UAS procedure.

In combination with the MiSeq FGx™ system, Verogen's ForenSeq DNA Signature Prep Kit can be used to provide a high amount of forensically relevant data, surpassing the robust Globalfiler™ amplification kit which targets 24 STR loci. Comparatively, the ForenSeq DNA Signature Prep Kit can target up to 231 STR and SNP loci, depending on which primer set is used: A or B. In primer set A, 27 autosomal STRs, 24 Y-STRs, 7 X-STRs, and 94 identity-informative SNPs are targeted in a single run [35]. In primer set B, the aforementioned loci are targeted, in addition to 54 ancestry-informative SNPs and 22 phenotype-informative SNPs [35]. With these additional targeted loci, the discriminating power is substantially higher than that of a typical STR-based amplification kit, thus it is extremely relevant in the forensic field. Though the bio-geographical and phenotypic SNPs are useful in identifying a contributor of a sample through deduction in their ancestry and appearance, the absence of these SNPs in primer set A is useful for those areas of the world

where their use is not authorized [35]. Nevertheless, the additional autosomal STRs included in the ForenSeq DNA Signature Prep Kit increases discriminating power and offers assistance in deconvoluting mixture samples.

### **1.5 Objective**

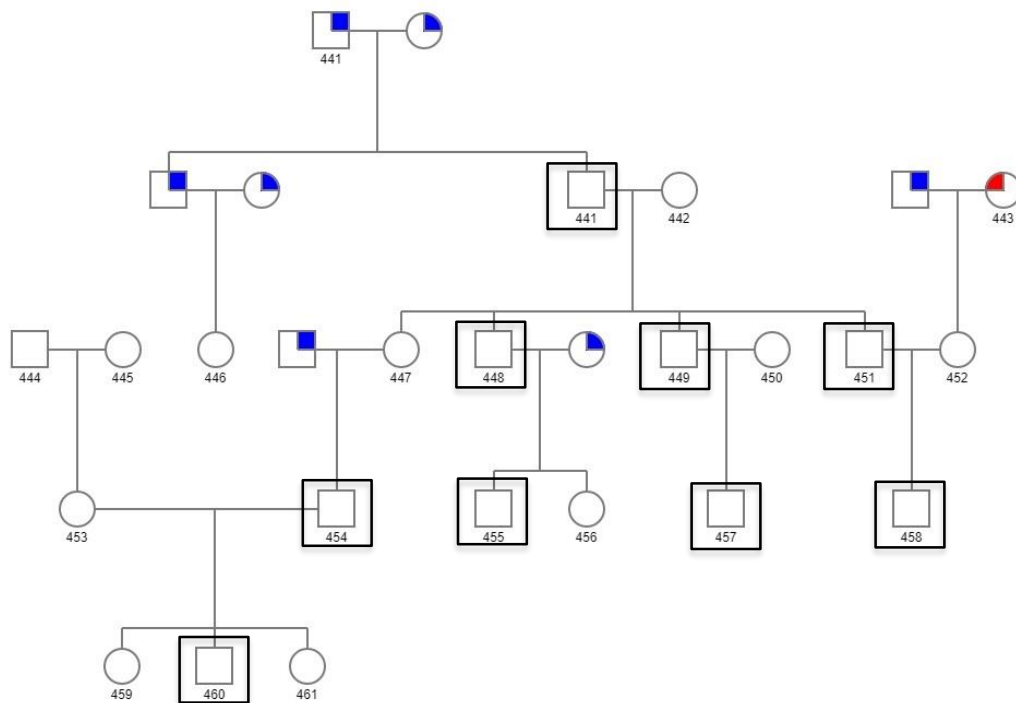
The inheritance of the Y chromosome occurs from father to son, and due to the lack of recombination of the MSY on the Y chromosome, provides a genetic history of a paternal line. Occasionally, random mutation events can occur that can alter the paternal line and lead to confusion regarding paternity or forensic casework samples. In this study, the MiSeq FGx™ system is compared against current CE-based methods to determine differences in detection of these mutational events, which can include insertions or deletions in alleles, as well as sequence mutations.

## 2. MATERIALS AND METHODS

Buccal swabs were obtained from 21 individuals, 10 male and 11 female, from a family pedigree spanning four generations. The samples were numbered 441 through 461, but only males numbered 441, 448, 449, 451, 455, 457, 458, 454, and 460 were analyzed for this study (Figure 2). The samples were previously extracted by David McEvoy, using the Qiagen EZ1 Advanced DNA Investigator Kit with the EZ1 Advanced DNA Investigator Kit Purification Protocol for Dried Saliva “Tip-Dance Protocol” [36].

Buccal Swabs Collected Dec. 2018

 Adopted Parent   Not Collected



**Figure 2. Multi-generational family pedigree.** A pedigree consisting of 21 individuals over four generations, labelled 441 through 461. The individuals analyzed in this study are outlined in black. Two paternal lines are analyzed in this study: a father and son (454 and 460) and a grandfather with three sons and three grandsons.

After extraction, quantification was performed in duplicate by David McEvoy, utilizing the Quantifiler Duo Kit (Applied Biosystems, Foster City, CA) with a 7500 Real-Time PCR Instrument (Applied Biosystems, Foster City, CA), according to the manufacturer's validated protocols (36). Due to lack of DNA in one sample, and a low concentration in the replicate, sample 455 was re-extracted and quantified by David McEvoy. He also performed re-quantification for Sample 444, which had large variation between the DNA concentrations of the replicates. A calibrated standard curve was used to determine the concentrations of DNA in each sample. For each sample, the two quantification values were averaged to determine the average concentration of DNA in the sample. For Sample 444, the values from the first sample of the duplicate and the re-quantification were averaged, since the two were closer together in concentration. For Sample 455, the re-quantification value was the only value taken into consideration, since the previous quantification values were both close to zero. Depending on the concentration, dilutions were performed using TE, or concentrations were performed with Microcon DNA Fast Flow filters and the "Concentration of DNA using Microcon DNA Fast Flow Filter" protocol.

## **2.1 Capillary Electrophoresis Workflow**

### **Amplification**

Following the protocols specified in User Guide Revision D, the samples were amplified using the YFiler™ Plus PCR Amplification Kit with an amplification target of 0.75 ng. DNA Control 007 served as a positive control, while TE served as a negative

control. After mixing the samples and the amplification components, the plate was centrifuged at 3000rpm for 30 seconds to ensure no air bubbles were present. The reaction plate was placed into a GeneAmp® PCR System9700 (Applied Biosystems, Foster City, CA) with amplification parameters as followed: 95°C for 1 minute, 30 cycles of [94°C for 4 seconds, 61.5°C for 1 minute], 60°C for 22 minutes, then 4°C until removal of the plate from the instrument.

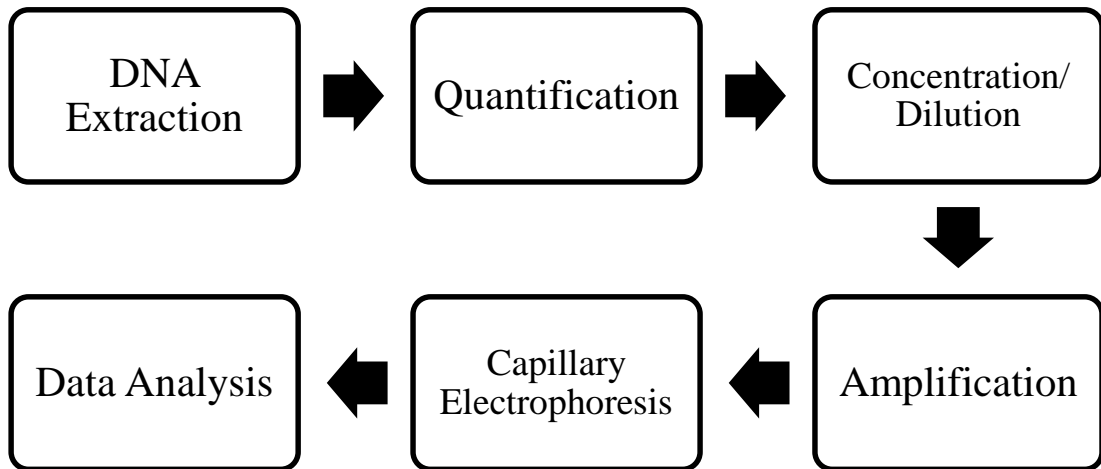
### **Capillary Electrophoresis**

A master mix of Hi-Di formamide and LIZ 600 size standard was prepared, with 9.4 µl of Hi-Di formamide per sample and 0.6 µl of LIZ 600 per sample. 10 µl of this master mix was pipetted into each appropriate well of a 96-well MicroAmp reaction plate. A Y allelic ladder, a positive control, and a negative control were used, and 1 µl of each was pipetted into the appropriate wells of the reaction plate. 1 µl of each sample was pipetted into the appropriate well of the reaction plate. The plate was then covered by a clean, dry septa and denatured in a heating block at 95°C for 3 minutes. The plate was subsequently chilled at 4°C for 3 minutes on an aluminum block in the freezer. A plate record was recorded on the computer linked to the ABI 3130 Genetic Analyzer. The amplified samples were separated using Pop-4™ Polymer and an electrokinetic injection of 5 seconds on the ABI 3130 Genetic Analyzer Capillary Electrophoresis (Applied Biosystems, Foster City, CA). Samples 451, 460, and 455 were reinjected to ensure optimal resolution.

## Analytical Software

Analysis of the resulting data from capillary electrophoresis was performed in GeneMapper ID-X V1.4 Software, with an AT of 30 RFU with the stutter filter on.

Analysis of EPGs in GeneMapper Software allowed STR genotypes to be determined and PHRs to be calculated, to determine if heterozygosity was present.



**Figure 3. Capillary electrophoresis workflow.** The workflow for DNA analysis with CE-based methods involves six discrete steps, beginning with a DNA sample, whether a liquid bodily fluid or a dried stain, and eventually leading to an EPG with genotypic data.

## 2.2 Sequencing Workflow

Sequencing data was previously produced by David McEvoy. These procedures are included here. The extracted and quantified samples were amplified using a primer mix from the ForenSeq DNA Signature Prep Kit and the PCR1 thermocycler procedure: 98°C for 3 minutes, 8 cycles of [96°C for 45 seconds, 80°C for 30 seconds, 54°C for 2 minutes, 68°C for 2 minutes], 10 cycles of [96°C for 30 seconds, 68°C for 3 minutes],

68°C for 10 minutes, then 10°C until removal of the plate from the instrument. Following this initial amplification, an additional amplification (PCR2) is performed to attach indices i5 and i7 to the target sequences.

Following both amplification runs, the tagged samples were purified using Sample Purification Beads along with a magnetic stand. The DNA binds to the Purification Beads, allowing the non-bound reaction components of the sample to be discarded. The magnetically-bound DNA is washed with ethanol multiple times, before the DNA is resuspended in buffer and transferred to a Purified Library Plate.

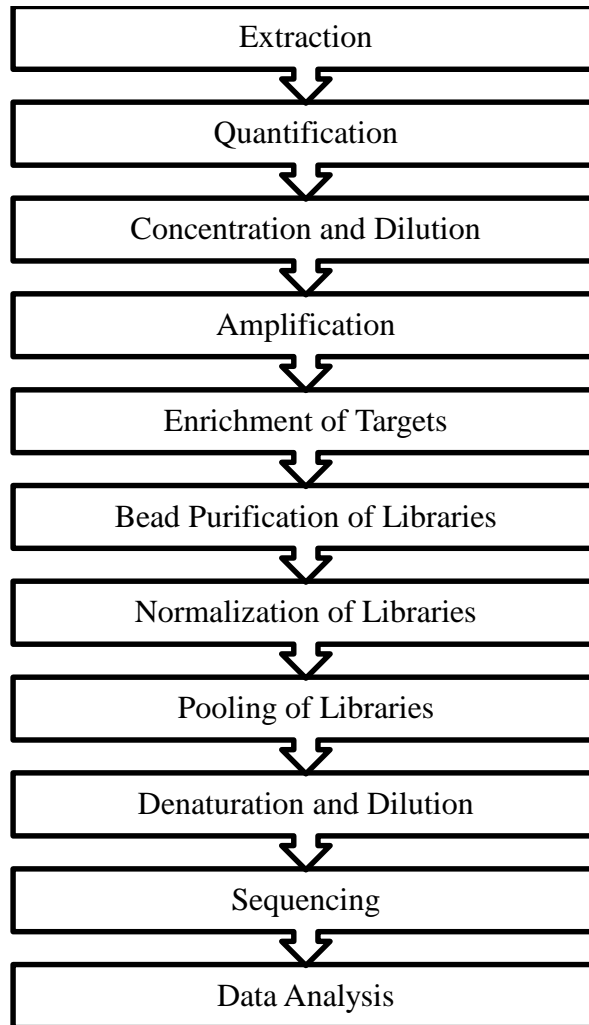
Following the purification of the DNA, the libraries are normalized, which ensures that each sample is equally represented when being sequenced, allowing for better resolution. This is performed by pipetting each of the samples into a Normalization Working Plate along with a master mix of LNA1 and LNB1. A series of wash steps with LNW1 are performed, before the DNA is resuspended in 0.1 N HP3 and transferred to a Normalization Library Plate. The samples were pooled by having 5 µl of each sample pipetted into a 1.5 microcentrifuge tube.

Immediately prior to sequencing, the pooled libraries must be denatured and diluted. The protocol-specified volumes of the pooled libraries, HT1 buffer, and the HSC were pipetted into a new microcentrifuge tube and heated for 2 minutes at 96°C before being placed in an ice-water bath for 5 minutes. A flow cell was cleaned with nuclease-free water and alcohol wipes before being loaded onto the instrument. The entire volume of the tube was loaded onto the reagent cartridge, placed on the instrument, and

sequencing was performed. These aforementioned steps were performed by David McEvoy (36).

### **Analytical Software**

Results obtained from the sequencing run were analyzed with the ForenSeq™ Universal Analysis Software version 1.3.6767 (Verogen, Inc., San Diego, CA) using Verogen's preset stutter percentages, AT, and IT for each individual Y-STR. In addition to determining genotypes of the samples, a sample comparison tool evaluated concordance of loci between samples. This tool graphically displays the length and intensity of typed STRs and SNPs in a scatter plot. In addition to visually depicting intersecting typed loci in a Venn diagram, disparity between two samples is illustrated in a table that shows the SNP or STR that is different.



**Figure 4. Sequencing Workflow.** The sequencing workflow has 10 distinct steps, with the first 4 steps being identical to that of the CE workflow. Following amplification, multiple steps are required to tag the DNA, purify, and normalize it in order to be run on the MiSeq FGx™.

### **3. RESULTS**

#### **3.1 Next Generation Sequencing Results**

The pedigree selected for this study had two distinct male lineages: a father and son (samples 454 and 460) and a grandfather with three sons and three grandsons.

Analysis of Miseq FGx™ data of these multi-generational paternal lines was conducted with the UAS to determine genotypic data of the samples, and to identify concordances and discordances present between samples. Upon comparison of genotypes of father-son pairs, a potential allele insertion at DYS385a/b was detected. At the DYS385a/b locus, instead of having an 11 allele and 14 allele, one of the sons had an additional 13 allele. The difference was only in length, with the 13 allele having one repeat less than the 14 allele; there was no difference in sequence of the allele or repeat. This was the only discordance observed between all samples at all loci (Table 1). A corresponding discrepancy in the CE data was not observed, which is discussed below.

**Table 1. Y-STR Genotype Results from Sequencing.** The genotypic data of each pedigree sample was determined through amplification with ForenSeq™ DNA Signature Prep Kit and the MiSeq FGx™ system. The bolded samples with an asterisk represent a father-son pair that is in the pedigree, but not in the same paternal line as the other samples. The discordant locus, DYS385a-b, is shown in the last panel.

Y-STR Genotypes, MiSeq FGx™

Sample	Relationship	DYS505	DYS570	DYS576	DYS522	DYS481	DYS19	DYS391	DYS635
441	Grandfather	12	15	19	10	22	14	11	23
448	Father	12	15	19	10	22	14	11	23
455	Son	12	15	19	10	22	14	11	23
451	Father	12	15	19	10	22	14	11	23
458	Son	12	15	19	10	22	14	11	23
449	Father	12	15	19	10	22	14	11	23
457	Son	12	15	19	10	22	14	11	23
<b>454*</b>	<b>Father</b>	11	17	17	10	22	15	10	23
<b>460*</b>	<b>Son</b>	11	17	17	10	22	15	10	23

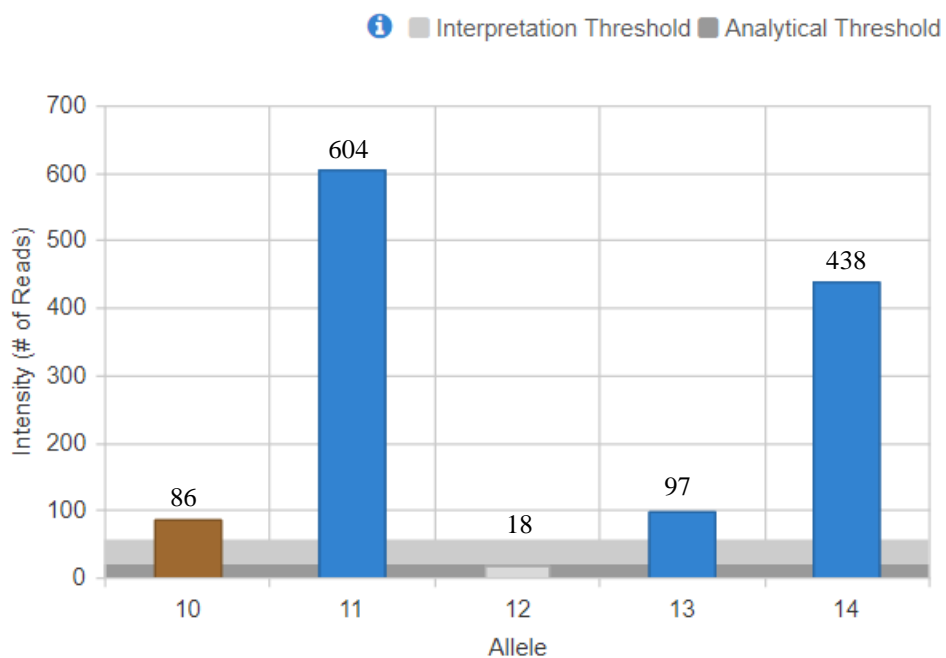
Y-STR Genotypes, MiSeq FGx™

Sample	Relationship	DYS437	DYS439	DYS389I	DYS389II	DYS438	DYS612	DYS390	DYS643
441	Grandfather	14	12	13	29	12	30	24	10
448	Father	14	12	13	29	12	30	24	10
455	Son	14	12	13	29	12	30	24	10
451	Father	14	12	13	29	12	30	24	10
458	Son	14	12	13	29	12	30	24	10
449	Father	14	12	13	29	12	30	24	10
457	Son	14	12	13	29	12	30	24	10
<b>454*</b>	<b>Father</b>	15	12	12	28	12	32	24	10
<b>460*</b>	<b>Son</b>	15	12	12	28	12	32	24	10

Y-STR Genotypes, MiSeq FGx™

Sample	Relationship	DYS533	Y-GATA-H4	DYS385a-b	DYS460	DYS549	DYS392	DYS448	DYF387S1
441	Grandfather	12	11	11, 14	11	11	13	18	35, 36
448	Father	12	11	11, 14	11	11	13	18	35, 36
455	Son	12	11	11, 13, 14	11	11	13	18	35, 36
451	Father	12	11	11, 14	11	11	13	18	35, 36
458	Son	12	11	11, 14	11	11	13	18	35, 36
449	Father	12	11	11, 14	11	11	13	18	35, 36
457	Son	12	11	11, 14	11	11	13	18	35, 36
<b>454*</b>	<b>Father</b>	12	12	11, 14	11	13	13	17	34, 37
<b>460*</b>	<b>Son</b>	12	12	11, 14	11	13	13	17	34, 37

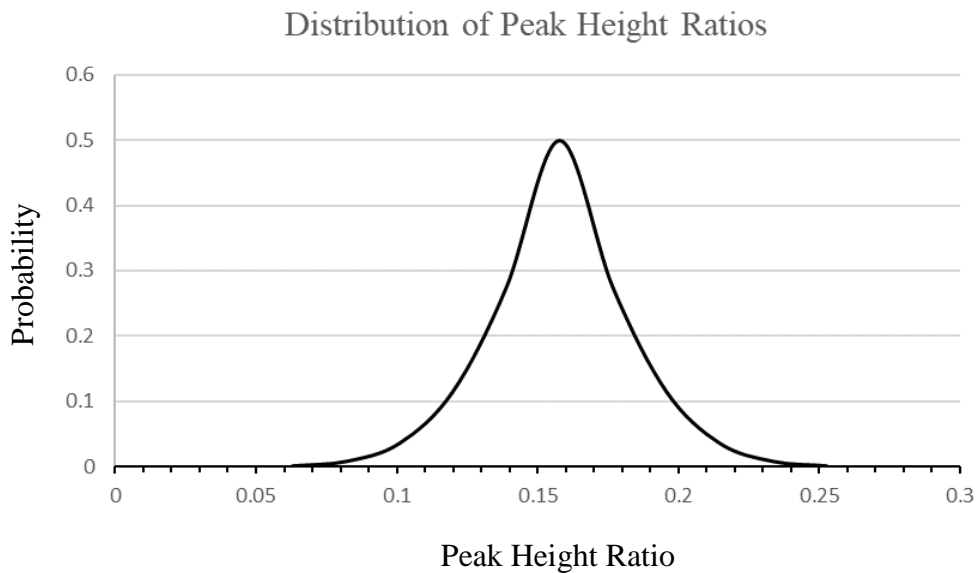
While the 11 allele had an ARC over 600 and the 14 allele had an ARC over 400, the 13 allele had an ARC of only 97. The resulting PHR of the 13 and 14 alleles was 0.22. A 10 allele was observed, but determined to be stutter as the PHR was 0.14, below the stutter filter of 20% for the DYS 385a/b locus.



**Figure 5. Alleles Detected at DYS385a/b in Sample 455.** Four different alleles with corresponding ARC values at DYS385a/b. This locus typically has two alleles present: the extra 10 allele was determined to be stutter, but the extra 13 allele may be a potential mutational allele insertion.

**Table 2. Comparison of Allele Read Counts of DYS385a/b.** The ARC values of the 13 allele and the 14 allele were used to determine the PHR of each sample analyzed in this study. The PHRs can be compared to the stutter filter of DYS385a/b to determine if the 13 allele could be a potential mutation. The bolded samples with an asterisk represent a father-son pair that is in the pedigree, but not in the same paternal line as the other samples.

Sample	Relationship	13 Allele	14 Allele	PHR	Stutter Filter
441	Grandfather	73	499	0.146	20%
448	Father	63	400	0.157	20%
455	Son	97	438	0.221	20%
451	Father	76	598	0.127	20%
458	Son	95	514	0.185	20%
449	Father	74	641	0.115	20%
457	Son	144	779	0.185	20%
<b>454*</b>	<b>Father</b>	81	602	0.135	20%
<b>460*</b>	<b>Son</b>	88	591	0.149	20%



**Figure 6. Distribution of Peak Height Ratios for DYS385a/b.** An average value and standard deviation were calculated from the peak height ratios of the 13 and 14 alleles. The PHR ratio of Sample 455, 0.221, is two standard deviations away from the mean.

### 3.2 Capillary Electrophoresis Results

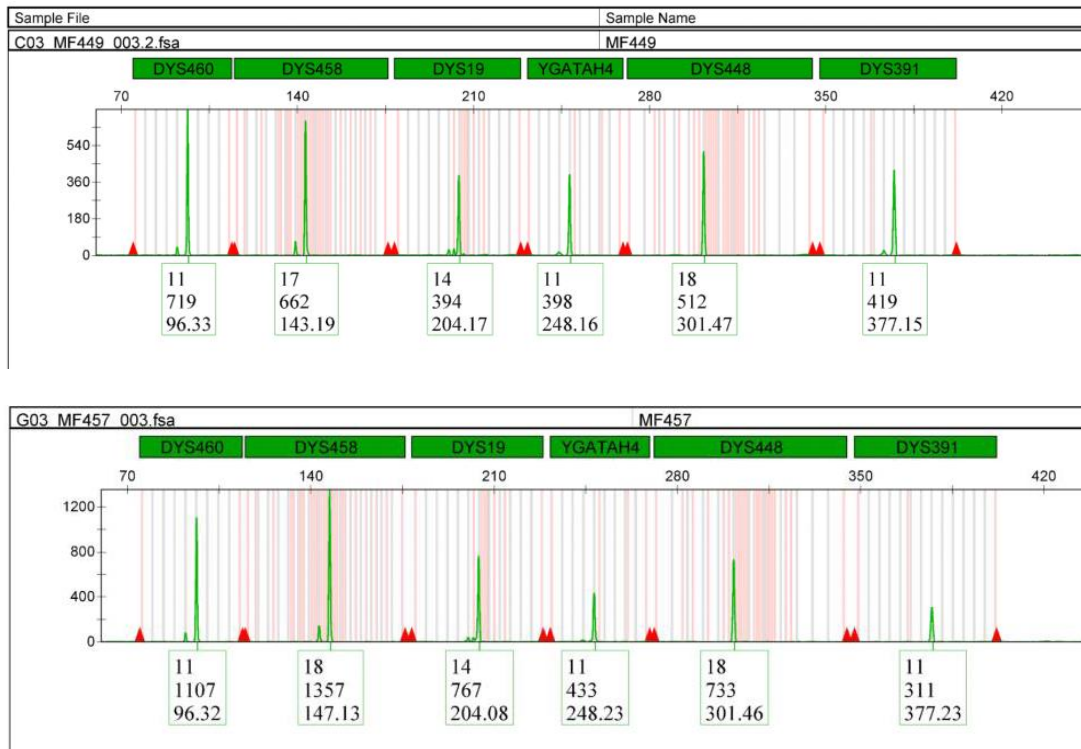
Analysis of CE data of the multi-generational paternal lines were conducted with GeneMapper ID-X V1.4 Software to determine genotypes of the samples, and to identify any artifacts or discordance present. In the initial injection, the grandfather, sample 441 from the multi-generational pedigree, was discordant with his son, sample 451, at locus DYS627 and DYS389II. At locus DYS627, the grandfather had a 22 allele, however the son had an additional 21 allele with 65 RFU. If this 21 allele was stutter, the stutter percentage would be 0.163, which is higher than the normal stutter range of 0-15%. At locus DYS389II, the grandfather had a 29 allele, but the son had an additional 28 allele with 72 RFU. If this 28 allele was stutter, the stutter percentage would be 0.190, which is higher than the normal stutter range of 0-15%.

Sample 451 was reinjected to ensure optimal resolution, resulting in the disappearance of the additional 21 allele being called at DYS627. The 21 allele was most likely stutter, and upon reinjection the RFU fell below the stutter filter. However, the additional 28 allele at DYS389II was still present, with an RFU of 103. The stutter percentage at this locus was 0.189, which is close to the previous value of 0.190. However, this additional allele was only present in sample 451; neither the grandfather, sample 441, or the son, sample 457, carried this allele.

There was a discordance between sample 448, the father, and sample 455, the son. At locus DYS518 the father had a 37 allele, but there was no allele called in the son's sample. Sample 455 was reinjected to ensure optimal resolution, but the reinjected sample still displayed the absence of an allele at DYS518.

A discordance between sample 454, the father, and sample 460, the son, occurred at multiple loci. At DYS576 and DYS460, multiple OL alleles were observed in the son's sample but were not present in the father's sample. At DYS458, the father had a 16 allele, but the son had an additional 18 allele. The resulting PHR for this locus was 0.095. Upon reinjection of sample 460 to ensure optimal resolution, all of these additional alleles were resolved, but another OL allele was discovered at DYS481. The son's sample had an OL allele with 2 bp less than the 22 allele that both he and his father shared.

The last discordance was observed between sample 449, the father, and sample 457, the son. The discordance occurred at DYS458: the father had allele 17, whereas the son had allele 18 (Figure 6).



**Figure 7. Y-STR Allele Comparison.** The father's EPG is above and the son's EPG is below. The allele calls, RFU, and length in bp are framed in boxes below the graph, while the loci are above each peak in green. The discordance between samples is shown at DYS458.

**Table 3. Y-STR Genotype Results from Capillary Electrophoresis..** The genotypic data of each pedigree sample was determined through amplification with Yfiler™ Plus PCR Amplification Kit and a Genetic Analyzer. The bolded samples with an asterisk represent a father-son pair that is in the pedigree, but not in the same paternal line as the other samples. Discordant loci DYS389II and DYS458 are shown in the first panel. The potential null allele at DYS518 and the OL allele at DYS481 are shown in the second panel.

Sample	Relationship	DYS576	DYS389I	DYS635	DYS389II	DYS627	DYS460	DYS458	DYS19	YGATAH4	DYS448	DYS391	DYS456	DYS390
441	Grandfather	19	13	23	29	22	11	17	14	11	18	11	15	24
448	Father	19	13	23	29	22	11	17	14	11	18	11	15	24
455	Son	19	13	23	29	22	11	17	14	11	18	11	15	24
451	Father	19	13	23	28, 29	22	11	17	14	11	18	11	15	24
458	Son	19	13	23	29	22	11	17	14	11	18	11	15	24
449	Father	19	13	23	29	22	11	17	14	11	18	11	15	24
457	Son	19	13	23	29	22	11	18	14	11	18	11	15	24
<b>454*</b>	Father	17	12	23	28	22	11	16	15	12	17	10	16	24
<b>460*</b>	Son	17	12	23	28	22	11	16	15	12	17	10	16	24

Sample	Relationship	DYS438	DYS392	DYS518	DYS570	DYS437	DYS385a/b	DYS449	DYS393	DYS439	DYS481	DYF387S1	DYS533
441	Grandfather	12	13	37	15	14	11, 14	30	13	12	22	35, 36	12
448	Father	12	13	37	15	14	11, 14	30	13	12	22	35, 36	12
455	Son	12	13		15	14	11, 14	30	13	12	22	35, 36	12
451	Father	12	13	37	15	14	11, 14	30	13	12	22	35, 36	12
458	Son	12	13	37	15	14	11, 14	30	13	12	22	35, 36	12
449	Father	12	13	37	15	14	11, 14	30	13	12	22	35, 36	12
457	Son	12	13	37	15	14	11, 14	30	13	12	22	35, 36	12
<b>454*</b>	Father	12	13	40	17	15	11, 14	30	13	12	22	34, 37	12
<b>460*</b>	Son	12	13	40	17	15	11, 14	30	13	12	21.1, 22	34, 37	12

#### 4. DISCUSSION

The data presents two definitive discordances between father-son pairs in a multi-generational pedigree, along with the possibility of two other artifacts being related to a mutational event. The discordance observed at DYS458 displayed an allele mutation involving the gain of a repeat (GAAA), rather than the deletion or insertion of an additional allele. The father's allele was 17, whereas the son's resulting allele at DYS458 was 18. Though the mechanism of allele mutation is not well understood, Y-STR mutability has been determined to be affected by total repeat number, where longer alleles tend to lose repeats and shorter alleles tend to gain repeats [37]. At 19 repeats and above, it is more probable that a loss of repeats would occur than a gain of repeats [38]. Another factor of Y-STR mutability is the complexity of the repetitive structure. Most Y-STRs have simple repeats, such as the (GAAA) sequence of DYS458. Other loci have compound repeats where two sequences alternate (GATA)(GACA), and a few others have complex repeats that have multiple repeat sequences of different sizes (GATA)(GACA)(CA)(CATA) [37]. Other variables that can factor into Y-STR mutability include the length of the motif and the father's age at the time of birth of the son [38]. Regardless of these factors, DYS458 is documented to be one of the most mutable Y-STR loci. According to Yang et al., 60% of mutations at DYS458 involved gains in repeats, and the calculated mutation rate for the locus was  $8.7 \times 10^{-3}$  [39]. This rate is substantially higher than the study's average Y-STR mutation rate, calculated as being  $3.4 \times 10^{-3}$  [39]. Though DYS458 is known to be more prone to mutating than most Y-STRs, a few Y-STRs have been categorized as highly-mutating. Yang et al.

determined DYS449 to be a highly-mutating Y-STR, calculating its mutation rate to be twice that of DYS458 at  $15.6 \times 10^{-3}$  [39].

The discordance observed at DYS389II and the discordance observed during sequencing at DYS385a/b indicated an allele insertion. For DYS389II, an allele insertion was detected in the stutter position, though the stutter percentage was calculated as 0.190 in both initial injection and reinjection. It is possible that an allele insertion occurred and there is significant allelic imbalance present. Allelic imbalance can occur due to differences in gene expression, specifically epigenetic inactivation or variation in regulatory regions [40]. Additionally, allelic imbalance can occur due to preferential amplification of one allele over another. When analyzing the CE data by itself, due to the stutter percentage staying consistent across injections, it appears to be allelic imbalance due to an aforementioned gene expression issue. However, upon comparison with the sequencing data, the addition of an allele can be precluded, as no additional allele was visualized in the sequencing data. This peak is therefore concluded to be stutter, though its consistent RFU is not understood. Had only CE been performed, analysis could have deemed a stutter peak an actual allele, leading to issues down the line with identification of the person producing the sample. The stutter percentage of the additional allele observed at DYS385a/b in the sequencing data was determined as being 0.22, over the locus-specific stutter threshold of 20%. Upon statistical analysis of the DYS385a/b stutter values across samples, 0.22 was determined to be outside two standard deviations away from the mean. Only five percent of values fall outside of two standard deviations from the mean, so this value could be significant. Though it is still possible it could be stutter,

the fact that three alleles are present at a single locus indicates the possibility that DYS385a/b is tri-allelic in this sample. In a Type 1 tri-allelic pattern, the sum of the two smaller peaks RFUs will equal the RFU of the bigger peak [41]. This type of pattern is generally associated with mutation during development, where some cells contain the normally-inherited allele and others contain the mutant allele [41]. In the instance of this sample, the two smaller peaks do not exactly equal the intensity of the bigger peak, but the values are relatively close. However, upon comparison with the CE data, the addition of an allele can be precluded as no additional peak was visualized on the EPG of the sample. Had only sequencing been performed, analysis could potentially have deemed this a tri-allelic locus, which could cause a problem in identifying the contributor of the profile.

The two remaining artifacts, a potential null allele and an off-ladder allele, have the potential to be results of mutation events other than allele insertions or gain of repeats. At DYS518, no allele was called, though there was a small peak around 350bp. It is possible that a mutational event could have resulted in a nucleotide change in the primer binding site of the sample, leading to the allele failing to amplify. The other possibility is that a mutational event, specifically the rearrangement of Yq, caused the deletion of the entire locus. At DYS481, an OL allele was identified at 2bp less than the called allele. Since stutter occurs at one repeat less than the allele, the OL allele cannot be attributed to stutter since the DYS481 repeat is trimeric (CTT). Additionally, the OL allele cannot be attributed to minus A, as it occurs at one bp less than the allele. Through comparison to other peaks present in the EPG, the OL allele cannot be attributed to bleed-

through from other dyes present. Furthermore, in the NGS data, a high stutter peak was observed at minus one repeat, but nothing was observed at minus 2 bp. A 2 bp deletion in the flanking region could be the source of this artifact, which, although it was only identified in one sample, was present and uncalled in nearly all samples.

## 5. CONCLUSION

With the growing prevalence of Y-STR testing, especially in the context of potential mixture samples where sensitivity is of the utmost importance, ensuring that the best methodology is implemented may be the defining factor for identification of a contributor. Between CE methodology and that of NGS, they each have their merits and their faults. Through separation by length as well as sequence, an allele insertion was identified with NGS, though corresponding CE data disproved it. CE-based methods were able to identify four artifacts, of which one was a definitive mutational event involving a gain of a repeat and another was an allele insertion, disproved using NGS data. The other two artifacts, a potential flanking region deletion and a potential null allele, were unable to be confirmed or denied. Through these results, CE methods were determined to be more useful in identifying Y-STR mutations, mainly due to the contradiction of the NGS mutation. This may be due to the set of Y-STR loci analyzed in the YFiler™ Plus PCR Amplification Kit being more mutable than those in the ForenSeq™ DNA Signature Prep Kit.

Though these occurrences may be rare, especially in the case of Y-STRs, the potential impact on forensic casework, genealogy, paternity testing, population genetics, and medicine cannot be ignored. Visualization of allele deletions could cause an analyst to assume homozygosity of a heterozygous locus, or the presence of a null allele that will not amplify. Visualization of allele insertions could cause an analyst to automatically assume that a single-source sample is, in fact, a mixture. Visualization of an allele mutation could cause an analyst to assume that two samples are not paternally linked.

These mutations and artifacts can be misleading, but extensive knowledge and thorough analysis can ensure the correct conclusion is reached.

## 6. FUTURE DIRECTIONS

In this study, Y-STRs were analyzed using both CE and NGS methodology, but the comparison between the two procedures could prove useful in detection of mutations in autosomal STRs and X-STRs. Through further comparison of these two methods in the analysis of different STR types, the best methodology in terms of detecting mutations can be determined. In addition to detecting mutations of single-source paternally-linked profiles, mixture samples containing multiple male profiles can be analyzed to help establish parameters for deconvolution, specifically in 2 or more male contributor mixtures with highly similar Y-STR profiles. With the analysis of mixtures, the study of different mixture ratios of contributors and mixtures with a higher number of contributors would also be useful.

With the MiSeq FGx™ system, not only are autosomal and sex chromosome STRs evaluated, but SNPs as well. Perhaps analysis of SNPs should be in consideration for future forensic DNA analysis, preferably in tandem with STRs for higher discriminating power, as they are more useful in samples containing degraded DNA. The analysis of degraded DNA versus non-degraded DNA from the same contributor can be useful in showing how the MiSeq FGx™ profiles differ, allowing degraded samples to be more readily identified and by attempting to calculate the effects on allelic drop-out. The MiSeq FGx™ can also be useful in identifying mutations of flanking regions and sequence specific stutter, which could result in the identification of alleles presumed to be null alleles, or the identification of stutter presumed to be an actual allele present at the locus.

## REFERENCES

1. Dahm, R. (2007). Discovering DNA: Friedrich Miescher and the early years of nucleic acid research. *Human Genetics*, 122(6), 565–581. doi:10.1007/s00439-007-0433-0
2. WATSON, J. D., & CRICK, F. H. C. (1953). Molecular Structure of Nucleic Acids: A Structure for Deoxyribose Nucleic Acid. *Nature*, 171(4356), 737–738. doi:10.1038/171737a0
3. Watson, J D, and F H C Crick. 2003. “A Structure for Deoxyribose Nucleic Acid.” *Nature* 421(6921): 397–8; discussion 396.
4. Strachan, T., & Read, A. (2010). *Human Molecular Genetics, Fourth Edition* (4th ed.). Garland Science.
5. Bichile, D. et al. (2014). Y chromosome: Structure and Biological Functions. *Indian Journal of Basic and Applied Medical Research*, 3(3), 152-160.
6. Hanson, E. K., & Ballantyne, J. (2006). Comprehensive annotated STR physical map of the human Y chromosome: Forensic implications. *Legal Medicine*, 8(2), 110–120. doi:10.1016/j.legalmed.2005.10.001
7. Ballantyne, K. N., Ralf, A., Aboukhalid, R., Achakzai, N. M., Anjos, M. J., Ayub, Q., Balažic, J. ž., Ballantyne, J., Ballard, D. J., Berger, B., Bobillo, C., Bouabdellah, M., Burri, H., Capal, T., Caratti, S., Cárdenas, J., Cartault, F., Carvalho, E. F., Carvalho, M., ... Kayser, M. (2014). Toward Male Individualization with Rapidly Mutating Y-Chromosomal Short Tandem Repeats. *Human Mutation*, 35(8), 1021–1032. <https://doi.org/10.1002/humu.22599>
8. Quintana-Murci, L., & Fellous, M. (2001). The Human Y Chromosome: The Biological Role of a “Functional Wasteland.” *Journal of Biomedicine and Biotechnology*, 1(1), 18–24. <https://doi.org/10.1155/s1110724301000080>
9. Sinclair, A. H., Berta, P., Palmer, M. S., Hawkins, J. R., Griffiths, B. L., Smith, M. J., Foster, J. W., Frischauf, A.-M., Lovell-Badge, R., & Goodfellow, P. N. (1990). A gene from the human sex-determining region encodes a protein with homology to a conserved DNA-binding motif. *Nature*, 346(6281), 240–244. <https://doi.org/10.1038/346240a0>
10. Heather, J. M., & Chain, B. (2016). The sequence of sequencers: The history of sequencing DNA. *Genomics*, 107(1), 1–8. <https://doi.org/10.1016/j.ygeno.2015.11.003>

11. Maxam, A. M., & Gilbert, W. (1977). A new method for sequencing DNA. *Proceedings of the National Academy of Sciences*, 74(2), 560–564. <https://doi.org/10.1073/pnas.74.2.560>
12. Sanger, F., Nicklen, S., & Coulson, A. R. (1977). DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences*, 74(12), 5463–5467. <https://doi.org/10.1073/pnas.74.12.5463>
13. Jeffreys, A. J., Wilson, V., & Thein, S. L. (1985). Individual-specific ‘fingerprints’ of human DNA. *Nature*, 316(6023), 76–79. <https://doi.org/10.1038/316076a0>
14. Saad, R. (2005). Discovery, Development, and Current Applications of DNA Identity Testing. *Baylor University Medical Center Proceedings*, 18(2), 130–133. <https://doi.org/10.1080/08998280.2005.11928051>
15. Gill, P., Jeffreys, A. J., & Werrett, D. J. (1985). Forensic application of DNA ‘fingerprints.’ *Nature*, 318(6046), 577–579. <https://doi.org/10.1038/318577a0>
16. Saiki, R., Scharf, S., Faloona, F., Mullis, K., Horn, G., Erlich, H., & Arnheim, N. (1985). Enzymatic amplification of beta-globin genomic sequences and restriction site analysis for diagnosis of sickle cell anemia. *Science*, 230(4732), 1350–1354. <https://doi.org/10.1126/science.2999980>
17. Garibyan, L., & Avashia, N. (2013). Polymerase Chain Reaction. *Journal of Investigative Dermatology*, 133(3), 1–4. <https://doi.org/10.1038/jid.2013.1>
18. Mullis, K. B. (1990). The Unusual Origin of the Polymerase Chain Reaction. *Scientific American*, 262(4), 56–65. doi:10.1038/scientificamerican0490-56
19. Butler, J. M. (2007). Short tandem repeat typing technologies used in human identity testing. *BioTechniques*, 43(4), Sii–Sv. doi:10.2144/000112582
20. Barron, Annelise E., and Harvey W. Blanch. (1995). DNA Separations by Slab Gel and Capillary Electrophoresis: Theory and Practice. *Separation & Purification Reviews* 24(1), 1–118.
21. Gupta, V., Dorsey, G., Hubbard, A. E., Rosenthal, P. J., & Greenhouse, B. (2010). Gel versus capillary electrophoresis genotyping for categorizing treatment outcomes in two anti-malarial trials in Uganda. *Malaria Journal*, 9(1), 19. <https://doi.org/10.1186/1475-2875-9-19>

22. Butler, J. M. (2015). The future of forensic DNA analysis. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 370(1674), 20140252. <https://doi.org/10.1098/rstb.2014.0252>
23. Gill, P., Urquhart, A., Millican, E., Oldroyd, N., Watson, S., Sparkes, R., & Kimpton, C. P. (1996). A new method of STR interpretation using inferential logic -development of a criminal intelligence database. *International Journal of Legal Medicine*, 109(1), 14–22. <https://doi.org/10.1007/bf01369596>
24. Holt, A., Wootton, S. C., Mulero, J. J., Brzoska, P. M., Langit, E., & Green, R. L. (2016). Developmental validation of the Quantifiler® HP and Trio Kits for human DNA quantification in forensic samples. *Forensic Science International: Genetics*, 21, 145–157. <https://doi.org/10.1016/j.fsigen.2015.12.007>
25. Butler, J. M., Buel, E., Crivellente, F., & McCord, B. R. (2004). Forensic DNA typing by capillary electrophoresis using the ABI Prism 310 and 3100 genetic analyzers for STR analysis. *ELECTROPHORESIS*, 25(1011), 1397–1412. <https://doi.org/10.1002/elps.200305822>
26. Butler, JM. Introduction to interpretation issues. NIST DNA Mixture Interpretation Webcast; 2013 Apr 12 [updated 2013; cited 2018 Nov 19]. Available from: <https://www.nist.gov/sites/default/files/documents/2017/04/28/DNA-MIXTURE-INTERPRETATION-WEBCAST-PrintableSlides-6-per-page.pdf>
27. Brookes, C., Bright, J.-A., Harbison, S., & Buckleton, J. (2012). Characterising stutter in forensic STR multiplexes. *Forensic Science International: Genetics*, 6(1), 58–63. <https://doi.org/10.1016/j.fsigen.2011.02.001>
28. Bentley, D. R., Balasubramanian, S., Swerdlow, H. P., Smith, G. P., Milton, J., Brown, C. G., Hall, K. P., Evers, D. J., Barnes, C. L., Bignell, H. R., Boutell, J. M., Bryant, J., Carter, R. J., Keira Cheetham, R., Cox, A. J., Ellis, D. J., Flatbush, M. R., Gormley, N. A., Humphray, S. J., ... Smith, A. J. (2008). Accurate whole human genome sequencing using reversible terminator chemistry. *Nature*, 456(7218), 53–59. <https://doi.org/10.1038/nature07517>
29. Liu, L., Li, Y., Li, S., Hu, N., He, Y., Pong, R., Lin, D., Lu, L., & Law, M. (2012). Comparison of Next-Generation Sequencing Systems. *Journal of Biomedicine and Biotechnology*, 2012, 1-11. <https://doi.org/10.1155/2012/251364>
30. Gettings, K. B., Borsuk, L. A., Steffen, C. R., Kiesler, K. M., & Vallone, P. M. (2018). Sequence-based U.S. population data for 27 autosomal STR loci. *Forensic Science International: Genetics*, 27, 106-115. <https://doi.org/10.1016/j.fsigen.2018.07.013>

31. Ballard, D., Winkler-Galicki, J., & Wesoły, J. (2020). Massive parallel sequencing in forensics: advantages, issues, technicalities, and prospects. *International Journal of Legal Medicine*, *134*(4), 1291–1303. <https://doi.org/10.1007/s00414-020-02294-0>
32. Woerner, A., King, J., & Budowle, B. (2017). Flanking Variation Influences Rates of Stutter in Simple Repeats. *Genes*, *8*(11), 329. <https://doi.org/10.3390/genes8110329>
33. Oldoni, F., & Podini, D. (2019). Forensic molecular biomarkers for mixture analysis. *Forensic Science International: Genetics*, *41*, 107-119. <https://doi.org/10.1016/j.fsigen.2019.04.003>
34. Schirmer, M., Ijaz, U. Z., D'Amore, R., Hall, N., Sloan, W. T., & Quince, C. (2015). Insight into biases and sequencing errors for amplicon sequencing with the Illumina MiSeq platform. *Nucleic Acids Research*, *43*(6), e37. <https://doi.org/10.1093/nar/gku1341>
35. England, R., & Harbison, S. (2019). A review of the method and validation of the MiSeq FGx™ Forensic Genomics Solution. *WIREs Forensic Science*, *2*(1). <https://doi.org/10.1002/wfs2.1351>
36. McEvoy, D. P. (2020). *A comparison of the Illumina MiSeq FGx™ System against capillary electrophoresis in the analysis of two-person mixtures*. [Unpublished master's thesis]. Boston University.
37. Butler, J. M., Kline, M. C., & Decker A. E. (2008). Addressing Y-Chromosome Short Tandem Repeat Allele Nomenclature. *Journal of Genetic Genealogy*, *4*(2), 125-148. [https://strbase.nist.gov/pub\\_pres/Butler2008-JoGG-YSTR-nomenclature.pdf](https://strbase.nist.gov/pub_pres/Butler2008-JoGG-YSTR-nomenclature.pdf)
38. Ballantyne, K. N., Goedbloed, M., Fang, R., Schaap, O., Lao, O., Wollstein, A., Choi, Y., van Duijn, K., Vermeulen, M., Brauer, S., Decorte, R., Poetsch, M., von Wurmb-Schwark, N., de Knijff, P., Labuda, D., Vézina, H., Knoblauch, H., Lessig, R., Roewer, L., ... Kayser, M. (2010). Mutability of Y-Chromosomal Microsatellites: Rates, Characteristics, Molecular Bases, and Forensic Implications. *The American Journal of Human Genetics*, *87*(3), 341–353. <https://doi.org/10.1016/j.ajhg.2010.08.006>

39. Yang, Y., Wang, W., Cheng, F., Chen, M., Chen, T., Zhao, J., Chen, C., Shi, Y., Li, C., Chen, C., Liu, Y., & Yan, J. (2018). Haplotypic polymorphisms and mutation rate estimates of 22 Y-chromosome STRs in the Northern Chinese Han father–son pairs. *Scientific Reports*, 8(1). <https://doi.org/10.1038/s41598-018-25362-3>
40. Wagner, J. R., Ge, B., Pokholok, D., Gunderson, K. L., Pastinen, T., & Blanchette, M. (2010). Computational Analysis of Whole-Genome Differential Allelic Expression Data in Human. *PLoS Computational Biology*, 6(7), e1000849. <https://doi.org/10.1371/journal.pcbi.1000849>
41. Huel, R. L., Basić, L., Madacki-Todorović, K., Smajlović, L., Eminović, I., Berbić, I., Milos, A., & Parsons, T. J. (2007). Variant alleles, triallelic patterns, and point mutations observed in nuclear short tandem repeat typing of populations in Bosnia and Serbia. *Croatian medical journal*, 48(4), 494–502.

**CURRICULUM VITAE**

