

2019

# Improving resolution of mixtures by DNA sequencing using the Illumina MiSeq FGx system

---

<https://hdl.handle.net/2144/38673>

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

BOSTON UNIVERSITY  
SCHOOL OF MEDICINE

Thesis

**IMPROVING RESOLUTION OF MIXTURES BY DNA SEQUENCING USING  
THE ILLUMINA MISEQ FGX SYSTEM**

by

**MICHAEL DENNIS MORETTO**

B.S., Michigan State University, 2016

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

2019

© 2019 by  
MICHAEL DENNIS MORETTO  
All rights reserved

Approved by

First Reader

---

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

Second Reader

---

Susanne Hoffmann-Benning, Ph.D.  
Assistant Professor, Biochemistry and Molecular Biology  
Michigan State University

## ACKNOWLEDGMENTS

First and foremost, I would like to thank my mom and dad for supporting me in everything I do. Regardless of how ridiculous the hobby or how large the decision, you were always on board. Even when it meant moving halfway across the country and only getting to see me for holidays, you supported me 100%. You continued to push me when I was feeling stressed and help me up when I'm feeling down. Without you both, none of this would have been possible.

A special acknowledgment goes to Dr. Susanne Hoffmann-Benning. You were the beginning of biological sciences career and helped me get to where I am now. From allowing me to volunteer in your lab to working under you for over 2 years, you've taught me much and more, from professional advice to life advice. From the United States to Germany, you have pushed me and helped me become the scientist I am today. Thank you for all your help.

I would also like to thank Dr. Robin Cotton for her support through these last two years at Boston University. You put up with my sarcasm and helped me to get through the stress that comes with a thesis. It has been a learning experience for both of us and you always seem to have a positive attitude through it all.

Lastly, I would like to acknowledge my fellow sequencing members, Tyler McDermott and Andre Porto. We decided to work on a project without knowing anything about the system and it has been quite the ride for all of us. None of this would have been possible without the help we have provided each other to keep ourselves sane.

**IMPROVING RESOLUTION OF MIXTURES BY DNA SEQUENCING USING  
THE ILLUMINA MISEQ FGX SYSTEM**

**MICHAEL DENNIS MORETTO**

**ABSTRACT**

The use of short tandem repeats (STRs) for genotyping forensic case samples has long been an effective tool for human identification. However, interpretation of forensic STR mixture samples can be difficult and any additional information to aid in this process can be invaluable. Allele overlap and stutter during PCR can cause drop out of the minor contributor's alleles and result in incorrect allele calling. The Scientific Working Group on DNA Analysis Methods (SWGDM) provides a list of guidelines on how to interpret DNA typing results from forensic STRs and mixtures, but there is still a significant variation in the interpretation of mixture samples between analysts in the same laboratory and between laboratories. The Illumina MiSeq Forensic Genomics™ system (Illumina Inc., San Diego, CA) is a massively parallel sequencing instrument that was developed specifically for the use in forensic DNA typing and which could provide sequence variations among on mixture samples. The ForenSeq™ DNA Signature Prep Kit is a kit that can be used with the MiSeq FGx™ platform. The DNA Primer Mix A (DPMA) included in the ForenSeq™ kit targets 27 autosomal STRs, 24 Y-STRs, 7 X-STRs and 94 identity single nucleotide polymorphisms (SNPs) on up to 32 or 96 samples, depending on the flow cell used. This study compares the STR performance on DNA mixtures of the MiSeq FGx™ and CE and evaluates its reliability and robustness.

The MiSeq FGx™ provides data in read count and the CE in relative fluorescence units (RFU), so the two output data cannot be directly compared to one another. Instead, the ratio of two contributors was calculated at three mixture ratios (1:1, 1:4, and 1:9) to use as a mean of comparison. The mean contributor ratios calculated on the MiSeq FGx™ were 1.799, 7.595, and 13.524 for the 1:1, 1:4, and 1:9 mixtures, respectively. This was not significantly different from the CE mean contributor ratios of 1.818, 7.722, and 14.827, respectively. More allele dropouts occurred on the MiSeq FGx™ than the CE at both 1:4 and 1:9 mixture ratios, but sequencing provided the detection of six isoalleles based on sequence variants that could not be discerned by CE. Other studies have shown full profile generation at these ratios, indicating there could have been some issues during library preparation. Further studies should be performed to thoroughly validate the ForenSeq™ process and evaluate the sensitivity of the instrument. Until then, it is recommended that the ForenSeq™ kit and MiSeq FGx™ system be used at close to equal mixture ratios or in tandem with the CE to prevent genotypes miscalling.

## TABLE OF CONTENTS

	Page
Title Page	i
Reader's Approval Page	iii
Acknowledgments	iv
Abstract	v
Table of Contents	vii
List of Tables	x
List of Figures	xi
List of Abbreviations	xiv
1. Introduction	1
1.1 Forensic DNA Analysis	1
1.2 What is DNA?	1
1.3 DNA Profiling Methods	2
1.3.1 Short Tandem Repeats	3
1.3.2 Sanger Sequencing	4
1.3.2.1 Mitochondrial DNA Testing	5
1.3.3 Next Generation Sequencing	6
1.4 Difficulties in Forensic Analysis of Mixture Samples	7
1.4.1 Stutter	8
1.4.2 Allele Overlap	8
1.4.3 Preferential Amplification	9

1.5 The MiSeq Forensic Genomics System	10
1.5.1 Improving mixture resolution with the MiSeq FGx™	11
1.6 Aim of this Study	11
2. Materials and Methods	13
2.1 Sample Preparation	13
2.2 Amplification and Fragment Separation	13
2.3 Sample Selection	14
2.4 Mixture Preparation	15
2.5 ForenSeq™ DNA Library Preparation	15
2.5.1 Amplification and Target Tagging	16
2.5.2 Target Enrichment	16
2.5.3 Library Purification	17
2.5.4 Library Normalization and Pooling	17
2.5.5 Pooling and Denaturing the Libraries	18
2.6 MiSeq FGx™ Sequencing	19
2.7 Contributor Ratio Determination	19
3. Results	23
3.1 Sample Selection	23
3.2 Capillary Electrophoresis	23
3.2.1 1:1 Mixtures	23
3.2.2 1:4 Mixtures	25
3.2.3 1:9 Mixtures	27

3.3 MiSeq FGx™	29
3.3.1 1:1 Mixtures	31
3.3.2 1:4 Mixtures	33
3.3.3 1:9 Mixtures	36
3.4 CE Versus MiSeq™	39
4. Discussion	44
5. Conclusions	47
Appendix A	48
Bibliography	50
Curriculum Vitae	54

## LIST OF TABLES

	Page
Table 1. Equations used for contributor ratio calculations of various autosomal STR combinations at an individual locus.	21
Table 2. Mean contributor ratios and their standard deviations calculated compared on the CE and MiSeq FGxTM	43
Table 3. Table of GlobalFiler <sup>TM</sup> alleles of mixture contributors	48
Table 4. Table of ForenSeq <sup>TM</sup> alleles of mixture contributors	49

## LIST OF FIGURES

	Page
Figure 1. Visual depiction of peak height imbalance in a STR profile	10
Figure 2. Flowchart illustrating the experimental set-up of both runs for each mixture.	15
Figure 3. Mean contributor ratio of 1:1 mixture samples by locus run on the CE	24
Figure 4. Mean contributor ratio across replicates of 1:1 mixture samples run on the CE	25
Figure 5. Mean contributor ratio of 1:4 mixture samples by locus run on the CE	26
Figure 6. Mean contributor ratio across replicates of 1:4 mixture samples run on the CE	27
Figure 7. Mean contributor ratio across all replicates of 1:4 mixture samples minus D16S539 locus in replicate 1 run on the CE	27
Figure 8. Mean contributor ratio of 1:9 mixture samples by locus run on the CE	28
Figure 9. Mean contributor ratio across replicates of 1:9 mixture samples run on the CE	29
Figure 10. Genotype of individual 434's D3S1358 locus illustrating the sequence difference of the isoalleles	30

Figure 11. Individual 434's 9 allele (top) and individual 438's 9 allele (bottom) at the D13S3171 locus, illustrating their sequence differences	30
Figure 12. Individual 438's 30 allele (top) and individual 434's 30 allele (bottom) at the D21S11 locus, illustrating their sequence differences	30
Figure 13. Mean contributor ratio of 1:1 mixture samples by STR locus run on the MiSeq FGx™	31
Figure 14. Mean contributor ratio of 1:1 mixture samples by SNP locus run on the MiSeq FGx™	32
Figure 15. Mean contributor ratio of STRs across replicates of 1:1 mixture samples run on the MiSeq FGx™	32
Figure 16. Mean contributor ratio of SNPs across replicates of 1:1 mixture samples run on the MiSeq FGx™	33
Figure 17. Mean contributor ratio of 1:4 mixture samples by STR locus run on the MiSeq FGx™	34
Figure 18. Mean contributor ratio of STRs across replicates of 1:4 mixture samples run on the MiSeq FGx™	35
Figure 19. Mean contributor ratio of SNPs across replicates of 1:4 mixture samples run on the MiSeq FGx™	35
Figure 20. Mean contributor ratio of 1:4 mixture samples by SNP locus run on the MiSeq FGx™	36
Figure 21. Mean contributor ratio of 1:9 mixture samples by STR locus run on the MiSeq FGx™	37

Figure 22. Mean contributor ratio of STRs across replicates of 1:9 mixture samples run on the MiSeq FGx™	38
Figure 23. Mean contributor ratio of SNPs across replicates of 1:9 mixture samples run on the MiSeq FGx™	38
Figure 24. Mean contributor ratio of 1:9 mixture samples by SNP locus run on the MiSeq FGx™	39
Figure 25. Comparison of mean contributor ratio of STR loci of 1:1 mixture samples on the CE versus MiSeq FGx™	40
Figure 26. Total mean contributor ratio of all 1:1 mixture samples on the CE versus MiSeq FGx™	41
Figure 27. Comparison of mean contributor ratio of STR loci of 1:4 mixture samples on the CE versus MiSeq FGx™	41
Figure 28. Total mean contributor ratio of all 1:4 mixture samples on the CE versus MiSeq FGx™	42
Figure 29. Total mean contributor ratio of all 1:9 mixture samples on the CE versus MiSeq FGx™	42
Figure 30. Comparison of mean contributor ratio of STR loci of 1:9 mixture samples on the CE versus MiSeq FGx™	43

## LIST OF ABBREVIATIONS

DNA	Deoxyribonucleic acid
DNL	Diluted Normalized Libraries
ddNTP	Dideoxyribose nucleotide triphosphate
dNTP	Deoxyribose nucleotide triphosphate
DPMA	DNA Primer Mix A
DPMB	DNA Primer Mix B
EPG	Electropherogram
FSP	ForenSeq Sample Plate
HP3	2N NaOH
HSC	Human Sequencing Control
kV	kilovolt
LNA1	Library Normalization Additives 1
LNB1	Library Normalization Beads 1
LNS2	Library Normalization Storage Buffer 2
LNW1	Library Normalization Wash 1
mtDNA	Mitochondrial DNA
NGS	Next generation sequencing
NLP	Normalized Library Plate
NWP	Normalized Working Plate
PBP	Purification Bead Plate
PLP	Purified Library Plate

PNL	Pooled Normalized Libraries
RFU	Relative Fluorescent Unit
RSB	Resuspension Buffer
SBS	Sequencing-by-synthesis
SGS	Second-generation sequencing
SPB	Sample Purification Beads
SWGDM	Scientific Working Group on DNA Analysis Methods

## **1. INTRODUCTION**

### **1.1 Forensic DNA Analysis**

The use of human deoxyribonucleic acid (DNA) as an identification tool has produced a profound change in how criminal investigations are carried out. Its ability to identify one individual out of the entire population quickly earned DNA profiling the status of “gold standard” [1] in the criminal justice system. Since its discovery by Sir Alec Jeffreys in 1985 [2], significant advancements in the forensic DNA analysis methods have been made to improve its sensitivity, reproducibility, and discriminatory power [3]. Early analysis methods used restriction fragment length polymorphisms (RFLPs) or variable number tandem repeats (VNTRs), which were subsequently phased out in the 1990s by the current mainstream method of genotyping short tandem repeats (STRs) [4]. Since this transition, the pace of change has slowed in the advancement of incorporating new technologies. However, in recent years, the use of what is known as massively parallel sequencing (MPS), or next generation sequencing (NGS), has emerged as an alternative technology in forensic DNA profiling [4].

### **1.2 What is DNA?**

DNA is a long chain, highly negatively charged molecule that is located in nucleus of most cells of the human body in the form of a double helix [5]. Each strand of the DNA molecule is made up of a chain of units known as nucleotides, which consist of a phosphate-deoxyribose sugar backbone and one of four bases: adenine, guanine, cytosine, and thymine. The bases of the two strands join together in pairs, where a base from one strand bonds via hydrogen-bonding to its complimentary base from the other strand. The only

pairs possible are cytosine to guanine and thymine to adenine, making the two strands complimentary to each other [5]. This means that if the order of bases for one strand were known, the order for the other strand could be deduced because of the presence base pairing. It is the order of these bases that makes up the genetic code responsible for creating and maintaining an organism.

In human somatic cells, DNA is subjected to an organized folding consisting of several levels of compaction using proteins known as histones. One single strand of highly compacted double-stranded DNA is known as a chromatid, and there are 2 chromatids (or 1 pair of chromatids) in a cell which make up one chromosome [6]. An individual has 23 pairs of chromosomes, inheriting one chromosome from their mother and one chromosome from their father. In eukaryotic DNA, much of the DNA is not translated to genes, and some of these non-translated regions consist of highly repetitive sequences of DNA [6]. It is the combination of these highly repetitive regions and the principles of Mendelian genetics that makes STR profiling possible.

### **1.3 DNA Profiling Methods**

All current DNA-typing methods rely on the same revolutionary technique, the polymerase chain reaction (PCR). With its ability to create millions of copies of multiple, specific DNA regions at once in just a few hours, PCR has revolutionized the forensic DNA world since its discovery in 1985 [7]. In tandem with capillary electrophoresis (CE), forensic DNA profiling took a massive leap forward.

### 1.3.1 Short Tandem Repeats

STR sequences are regions of DNA that are shorter in length (~100-400 bp) than their VNTR predecessor (~400-1000 base pairs) and consist of a variable number of tandemly repeated sequences. In forensic DNA-typing, the typical STR loci contain tetranucleotide repeats, or a four base pair (bp) sequence that is repeated a certain number of times [7]. Due to their high level of polymorphism and because of their short length, they are easily amplified with the PCR process [8]. The shorter length is advantageous for forensic samples because they are often degraded. Where longer stretches of DNA would be broken up into smaller pieces, the shorter STR regions are more likely to be intact. In 1994, the use of STRs in conjunction with mitochondrial DNA sequencing on degraded DNA was demonstrated by analyzing 70 year-old bones to identify the remains of the Romanov family [9].

PCR plays an important role in forensic DNA-typing because the primers used to locate and amplify the target STR regions can be modified with different fluorescent tags. This allows for many STR markers to be multiplexed together and still be resolved from each other by electrophoretic separation. One of the first STR multiplexes in the early 1990s comprised of only four loci with a matching probability of 1 in 10,000, and by 1997, 13 core STR loci were identified with a matching probability of more than one in a trillion unrelated individuals [7]. The scientific community has continued to improve these STR multiplex kits, with some of the latest kits containing up to 24 loci.

Individuals are identified based on the number of repeats at a specific locus, which can be determined by comparing the length of the DNA segment to an allelic ladder. For

example, an individual with a 9 allele at the TPOX locus has a four-base unit repeated nine times and will be 4 bp longer than someone with an 8 allele, or 8 repeats. These repeats may be different in DNA sequence but identical in repeat length. CE analysis does not allow the determination of the sequence of the repeats while MPS can. This sequencing adds information beyond the length of the allele.

### 1.3.2 Sanger Sequencing

Sanger sequencing was first described in 1977 with a methodology that uses some of the same mechanics as PCR. It relies on DNA polymerase in the presence of a DNA template and the DNA building blocks, deoxyribose nucleotide triphosphates (dNTPs), to build a new strand of DNA. However, it also utilizes dideoxyribose nucleotide triphosphates (ddNTPs) to terminate the DNA extension wherever they are incorporated because these do not contain a 3'-hydroxyl group [10] and, hence cannot form the phosphodiester bond necessary for chain extension. Four reaction mixtures are set up containing all the dNTPs and each reaction mix receiving a different ddNTP: ddATP, ddGTP, ddTTP, or ddCTP. When the template DNA is incubated in the presence of a polymerase and a mixture of, for instance, a ddATP and dNTPs, a mixture of different fragment lengths will be obtained based on whether a dNTP or the ddNTP is added at the location. This occurs in each reaction, and when the 4 reactions are separated by gel electrophoresis in parallel, the bands will indicate where in the DNA sequence each ddNTP is added.

The Sanger method was widely used for short DNA regions of 500 to 700 bp [11], but larger regions of DNA posed a problem. The main limitation with Sanger sequencing

was the time required to create the fragments for sequencing by cloning into the bacteriophage lambda. This is highlighted by the Human Genome Project, where it took multicenter collaborations more than 10 years to sequence 5% of the human genome using Sanger sequencing, with the last 95% of the genome sequenced in a year by incorporation of a new shotgun sequencing approach [11,12].

#### 1.3.2.1 Mitochondrial DNA Testing

Sometimes in forensic science, conventional STR typing does not work if the sample is very old or degraded. In these cases, mitochondrial DNA (mtDNA) testing can be used. While nuclear DNA testing is more valuable and discriminatory, there are hundreds of copies of mtDNA in each cell compared to two copies of nuclear DNA, making it more likely to survive degradation. However, the mtDNA genome is only 16,569 bp long and has only one region that has many bases not coding for a gene, known as the control region, or D-loop, that is 1,122 bp [7]. Because it is such a short region and not highly variable, typical STR-type analysis is not viable. Testing of mtDNA first introduced forensic DNA analysis to DNA sequencing.

The first human mtDNA was sequenced in 1981 [13] with Sanger sequencing and became the reference sequence to which new sequences were compared. This early identification of the mtDNA sequence allowed for PCR primers to be developed for the D-loop, removing the long process of fragment cloning from the procedure. Sequencing is carried out using four different fluorescent dyes attached to the four different ddNTPs and separation occurs using capillary electrophoresis (CE). As the DNA passes through the capillary, the different lengths of DNA fragments are separated by size, with each fragment

fluorescing the color of the ddNTP that was incorporated as the terminating base. The order in which the fluorescing dyes are read by the CE indicates the DNA sequence.

Mitochondrial DNA is inherited only from the maternal lineage, with all siblings and maternal relatives having identical mtDNA sequences [7]. Mass disaster or missing person cases can benefit from this type of testing for familial linkages but matches in forensic cases will be less significant. While having another method for individualization when traditional methods are unavailable would be very helpful, mtDNA cannot be used alone for identification. While effective for forensic mtDNA testing, CE-based Sanger sequencing does not have the resolution, speed, or throughput that is required to sequence the STRs currently used by forensic labs. New technologies have been introduced to allow for the sequencing of multiple human genomes in a single run, known as second-generation sequencing (SGS) [12].

### 1.3.3 Next Generation Sequencing

SGS, also known as next generation sequencing (NGS), has effectively overcome all the limitations that faced Sanger sequencing. Higher throughput, increased resolution, and speed has made NGS a cheaper and more effective alternative to Sanger sequencing. One method in particular, sequencing-by-synthesis (SBS), has become widely adopted in areas of research ranging from microbial communities [14] to human virus and cancer research [15,16]. What distinguishes SBS from Sanger sequencing is the use of fluorescently labeled, reversible terminator dNTPs [17]. The bases are identified based on their fluorescence after addition, followed by the removal of the terminating portion which allows the addition of another reversible terminator dNTP. This is more advantageous than

Sanger sequencing because it allows all the terminator dNTPs to be added simultaneously, instead of in separate tubes, improving accuracy and time and reducing the cost per sample. With the SBS method, it has been possible to sequence a whole human genome in a matter of 8 weeks [17] compared to the 10 years it took using Sanger sequencing.

#### **1.4 Difficulties in Forensic Analysis of Mixture Samples**

Using currently available STR tests, single source samples are relatively straightforward to interpret, but with each additional contributor, the difficulty of interpreting a sample increases exponentially [18]. Forensic case samples can come in any number of contributors, and with the high regard that DNA is held at in the courtroom, it is important that these samples are interpreted accurately. One of the most prevalent forensic case types are sexual assault cases, where the samples often contain DNA from two or more individuals. The Scientific Working Group on DNA Analysis Methods (SWGDM) provides a list of guidelines on how to interpret DNA typing results from forensic STRs and mixtures [19], but there is still significant difference in the interpretation of mixture samples between laboratories and between individual analysts in the same laboratory [20,21].

The large variation in mixture interpretation between labs and individuals has sparked the development of computational models for mixture interpretation, known as probabilistic genotyping. Different software, such as STRmix, EuroForMix, and Lab Retriever, have been tested and validated for use in forensic cases [22–24].

With the advancement of DNA-typing methods, new forensic kits can detect smaller and smaller amounts of DNA, increasing the likelihood of individuals being

detected. Current STR kits are so sensitive that they have been able to obtain complete profiles from as little as 100 picograms (pg) of template DNA, and some studies have shown to be able to obtain profiles with as little as a single diploid cell [25]. As template level decreases, an inherent issue with PCR known as stutter becomes more of a problem, even with a single-source profile. Other issues, such as preferential amplification of alleles and allele overlap (sharing) add to the difficulty of mixture interpretation.

#### 1.4.1 Stutter

STR analysis by PCR has revolutionized the forensic field, but it has its limits. A common artifact of PCR occurs when the DNA polymerase skips, or sometimes repeats, one of the repeat motifs of the template DNA [26]. As the reaction continues, these strands of DNA then continue to be copied like the template strand and can be in high enough concentrations to be read by the CE. They typically show up in a DNA profile as a peak one repeat (4 bp) shorter or longer than the parent allele and typically at a much lower peak height [27]. While it is usually easy to detect stutter in a single-source sample because of the peak height difference, it can be difficult to discern a stutter peak from a peak legitimately from a minor contributor, especially at low concentrations of the minor contributor.

#### 1.4.2 Allele Overlap

While forensic STR typing is a very powerful identification tool, mixture samples can be difficult to interpret if neither individuals profile can be assumed to be present in the mixture. STRs can only be separated by length, and when two individuals share the same allele at a locus, it is impossible to know if they share that allele or not using CE

analysis. If a 2-person mixture has 3 alleles at a locus, the total number of possible allele combinations is 6. One individual could be homozygous for one of those alleles and the other individual heterozygous with the other 2 alleles. Or, they could both be heterozygous sharing one of the 3 alleles. With the addition of a third person, the number of combinations for a 3 allele locus jumps from 6 to 29 [18]. When concentrations of one of the contributors falls to a low enough level, stutter peaks and minor contributor's alleles can have similar peak heights. Stutter has been known to range from 5% to 30% of the height of the parent allele [27], which can make determining if the minor contributor has an allele in the stutter position difficult.

#### 1.4.3 Preferential Amplification

Sometimes the ratio between allele peak heights can be used 2 contributor mixture analysis to deduce contributor genotypes. SWGDAM recommends a peak height ratio >60% as a measure for two alleles (sister alleles) coming from the same individual [19]. This is because of another inherent issue with PCR which is greater amplification of a locus relative to another. This can result in large peak height differences between two alleles of one individual in a single source sample that may drop below the 60% cutoff recommendation (Figure 1). If this occurs in a mixture sample and one relies solely on the peak height ratio, an incorrect interpretation of the profile can be made.

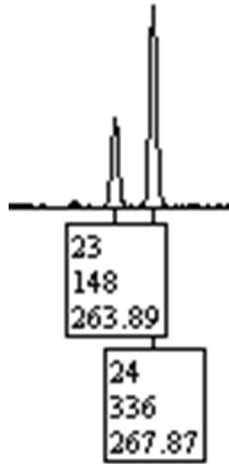


Figure 1: Visual depiction of peak height imbalance in a STR profile

### 1.5 The MiSeq Forensic Genomics System

To be able to produce a forensic STR profile and to know the sequence of those STRs could prove incredibly useful in identifying one contributor in a mixture. One system looking to break out in the forensic field is the MiSeq Forensic Genomics System (FGx™).

The MiSeq FGx™ is a massively parallel sequencing (MPS) platform that sequences all of the currently mandated forensic STR loci as well as single nucleotide polymorphism (SNP) loci simultaneously using SBS chemistry. The system uses the Verogen ForenSeq™ DNA Signature Prep Kit (Verogen, Inc., San Diego, CA), which comes with 2 different primer mixes: DNA Primer Mix A (DPMA) and DNA Primer Mix B (DPMB). DPMA identifies 27 autosomal STRs, 24 Y-STRs, 7 X-STRs and 94 identity SNPs. DPMB can identify everything from DPMA in addition to 56 ancestry informative SNPs and 24 phenotypic SNPs [28]. Additionally, the MiSeq FGx™ system utilizes one of two flow cells that can sequence up to 32 or 96 samples.

### 1.5.1 Improving mixture resolution with the MiSeq FGx™

The clearest advantage that the MiSeq FGx™ has over conventional STR-based CE mixture analysis is that it can determine the specific DNA sequence at the forensically relevant STR loci. This makes it possible for alleles of the same length at a locus to be identified as two different alleles based on differences their sequence, or isoalleles. This can not only separate 2 people from an apparent single allele but adding sequence variation to the already determined allele frequencies can also increase the power of discrimination. On top of being able to identify isoalleles, the sequencing data may also aid in the detection of the minor contributor's alleles when they coincide with the stutter position of the major contributor's alleles.

### 1.6 Aim of this Study

Deconvolution of forensic STR mixture samples can be difficult and being able to obtain additional information to aid in this process will be important. Compared to traditional CE-based STR analysis, the MiSeq FGx™ not only targets an additional 6 autosomal STRs, 24 Y-STRs, 7 X-STRs and 94 SNPs on up to 96 samples at one time, it also provides the DNA sequence of those targets. The additional loci as well as separation of alleles by sequence should provide much more information for resolving mixture samples. The contributor ratio accuracy and MiSeq FGx™ performance is analyzed here and compared to current CE-based methods. The DNA sequencing process used here requires 3 PCR amplification steps overall, which could increase the likelihood of preferential amplification. Additionally, there are many wash steps and transfer steps involved in the purification and normalization of the libraries prior to sequencing, which

may increase profile variability. A side by side assessment of the ForenSeq™ Signature Prep Kit with the MiSeq FGx™ system and the GlobalFiler™ PCR Amplification Kit (Applied Biosystems, Foster City, CA) using equivalent samples containing two person DNA mixtures at three different mixture ratios is presented.

## **2. Materials and Methods**

DNA Quantitation was performed in a 7500 Real-Time PCR (RT-PCR) Instrument (Applied Biosystems, Foster City, CA) using Quantifiler Duo Kits (Applied Biosystems, Foster City, CA) according to the manufacturer's specifications using 2 microliters ( $\mu\text{L}$ ) of sample. DNA concentrations were then calculated using a previously calibrated standard curve [29].

### **2.1 Sample Preparation**

DNA samples were obtained as saliva samples collected anonymously from 7 individuals. The samples were washed prior to DNA extraction by mixing 300 microliters ( $\mu\text{L}$ ) of laboratory prepared TE Buffer and 300  $\mu\text{L}$  of neat saliva. Cells were pelleted by centrifugation at 3000 rotations per minute (rpm) for five minutes. The supernatant was removed, the pellet was resuspended in 400  $\mu\text{L}$  TE Buffer, and centrifuged again at 3000 rpm for five minutes. This step was repeated, and the pellet was resuspended in 50  $\mu\text{L}$  of TE Buffer.

Nuclear DNA was extracted from each of the samples using the QIAmp DNA Investigator Kit (QIAGEN, Hilden, Germany) according to the QIAmp DNA Investigator Handbook Protocol "Isolation of Total DNA from Small Volumes of Blood or Saliva" [30] using 100  $\mu\text{L}$  of saliva and an incubation at 56°C for 1 hour instead of the recommended 10 minutes. The sample DNA was then stored at -20°C until further use.

### **2.2 Amplification and Fragment Separation**

DNA amplifications were performed using GlobalFiler™ PCR Amplification Kit in a GeneAmp® PCR System 9700 thermal cycler (Applied Biosystems, Foster City, CA)

with an initial amount of 0.5 ng of input DNA. Each run contained a positive control using 5  $\mu$ L DNA Control 007 and a negative control that should not contain DNA. Amplification was carried out with the following cycling parameters: 95°C for 1 minute, then a cycle of 94°C for 10 seconds and 59°C for 90 seconds for 30 cycles, followed by a hold at 60°C for 10 minutes and then 4°C indefinitely.

After amplification, 1  $\mu$ L of the amplified samples were added to the appropriate wells of a 96-well reaction plate. Each well that received a sample contained 9.5  $\mu$ L of HiDi deionized formamide and 0.5  $\mu$ L of 600 LIZ<sup>®</sup> Size Standard. The samples were then denatured by heating them at 95°C for 3 minutes and then chilled for 3 minutes. Fragment separation was carried out using an ABI 3130 Genetic Analyzer Capillary Electrophoresis instrument (Applied Biosystems, Foster City, CA) using POP-4<sup>™</sup> Polymer (Applied Biosystems, Foster City, CA) for separation with a 1.2 kilovolt (kV) injection for 2 seconds. Electropherograms (EPGs) were analyzed using GeneMapper<sup>®</sup> ID-X version 1.4 with the stutter filter on and an analytical threshold set at 30 relative fluorescence units (RFU).

### **2.3 Sample Selection**

To determine which samples would be used for mixture analysis, each sample was diluted down to 0.2 ng/ $\mu$ L, amplified with a target of 0.5 ng, and separated via CE. Each individual's EPG was analyzed to confirm there was no contamination and then compared against each other EPG to identify the number of overlapping alleles. The samples chosen were to have high heterozygosity with the highest amount of allele overlap to get the best chance for DNA sequence variation of overlapping alleles.

## 2.4 Mixture Preparation

The two samples selected with the most allele overlap were quantified in triplicate and then made into 3 mixture ratios: 1:1, 1:4, and 1:9. Each mixture was made to a final volume of 100  $\mu\text{L}$  at a concentration of 0.2  $\text{ng}/\mu\text{L}$ , using TE Buffer if necessary. Each mixture was then amplified in quadruplicate over 2 runs following the chart in Figure 2 using either GlobalFiler™ (as described above) or the ForenSeq™ DNA Signature Prep kit (as described in section 2.3).

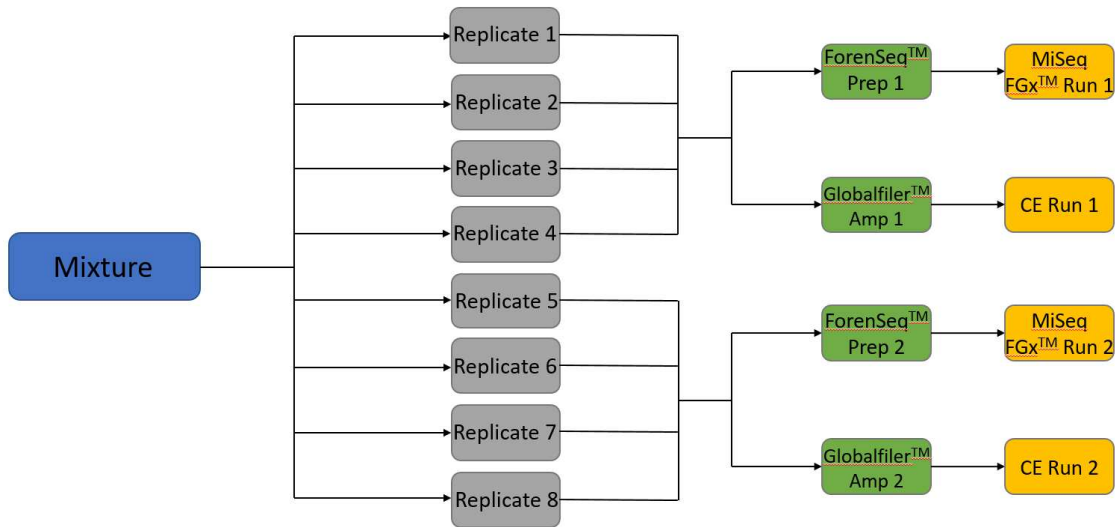


Figure 2: Flowchart illustrating the experimental set-up of both runs for each mixture.

## 2.5 ForenSeq™ DNA Library Preparation

Before the mixture samples were processed for sequencing on the MiSeq FGx™ instrument (Illumina Inc., San Diego, CA), the samples must go through a 2-step PCR amplification process to amplify target loci and to add the necessary indexes (i5 and i7). Each sample is then purified through a series of wash steps, normalized to ensure samples of varying concentrations are represented equally, and eventually pooled and denatured for

sequencing. Library preparation was conducted using the ForenSeq™ DNA Signature Prep Kit.

#### 2.5.1 Amplification and Target Tagging

Using a 96-well reaction plate labeled ForenSeq Sample Plate (FSP), 10  $\mu\text{L}$  of a master mix (containing 4.7  $\mu\text{L}$  PCR1, 0.3  $\mu\text{L}$  enzyme mix, and 5.0  $\mu\text{L}$  of DNA Primer Mix A [DPMA] per sample) was added to each well that would contain a sample. 5.0  $\mu\text{L}$  of each sample, diluted down to 0.2 ng/ $\mu\text{L}$ , was added to the appropriate wells so that 1 ng of total DNA was in each well. 5.0  $\mu\text{L}$  of 2800M control DNA (single-source human male genomic DNA) was used as a positive control and 5.0  $\mu\text{L}$  of nuclease-free water as a negative control. The plate was sealed and centrifuged at 1000 x g for 30 seconds and placed in the thermal cycler for amplification of target loci as follows: initial incubation at 98°C for 3 minutes, 8 cycles of [45 seconds at 96°C, 30 seconds at 80°C, 2 minutes at 54°C, 2 minutes at 68°C], 10 cycles of [30 seconds at 96°C, 3 minutes at 68°C], a final extension at 68°C for 10 minutes, and an infinite hold at 10°C.

#### 2.5.2 Target Enrichment

DNA is further amplified in this step along with the addition of Index 1 (i7) and Index 2 (i5) adapters that are required for cluster amplification and sample separation downstream. There are 12 i7 adapters (701-712) which correspond to one column of a 96-well plate (1-12) and 8 i5 adapters (501-508) which correspond to one row of a 96-well plate (A-H). Using a multichannel pipette, 4  $\mu\text{L}$  of each i7 adapter was added to the corresponding column and 4  $\mu\text{L}$  of each i5 adapter was added to the corresponding row. A total of 27  $\mu\text{L}$  of PCR2 solution was then added to each well, followed by centrifugation

at 1000 x g for 30 seconds. PCR was then performed as follows: initial incubation for 30 seconds at 98°C, 15 cycles of [98°C for 20 seconds, 66°C for 30 seconds, 68°C for 90 seconds], a final extension at 68°C for 10 minutes, and an infinite hold at 10°C.

### 2.5.3 Library Purification

Samples are purified using Sample Purification Beads (SPB) to bind the DNA while the other reaction components are washed away. 45 µL of SPB are added to each well of a midi plate labeled Purification Bead Plate (PBP). 45 µL of each sample of the FSP was transferred to the corresponding well of the PBP, the plate was sealed, and left to shake at 1800 rpm for 2 minutes. The PBP plate was placed on a magnetic stand, the supernatant removed and discarded from each well, and then washed two times with freshly prepared 80% ethanol, removing the supernatant each time. To ensure all the ethanol was removed, the PBP plate was centrifuged at 1000 x g for 30 seconds, placed back on the magnetic stand, and then any residual ethanol was removed. The plate was removed from the magnetic stand and the beads were resuspended in 52.5 µL of Resuspension Buffer (RSB) by shaking at 1800 rpm for 2 minutes. The plate was then placed back on the magnetic stand and the solution was left until it cleared before 50 µL of each well was transferred to the corresponding well of a new 96-well PCR plate labeled Purified Library Plate (PLP).

### 2.5.4 Library Normalization and Pooling

To a new 96-well midi plate labeled Normalization Working Plate (NWP), 45 µL of a master mix, made up of 46.8 µL Library Normalization Additives 1 (LNA1) per sample and 8.5 µL Library Normalization Beads 1 (LNB1) per sample, was transferred to each well that will contain a library. 20 µL from each well of the PLP were transferred to the

corresponding well of the NWP, the NWP was sealed and shaken for 30 minutes at 1800 rpm. After shaking, the NWP was placed on a magnetic stand and left until the liquid cleared before removing the supernatant of each well. The NWP was then removed from the magnetic stand and washed twice as follows: 45  $\mu$ L Library Normalization Wash 1 (LNW1) was added to each well, the plate was sealed and shaken for 5 minutes at 1800 rpm, placed back on the magnetic stand, and all of the supernatant from each well was removed. The plate was then centrifuged for 30 seconds at 1000 x g, placed back on the magnetic stand, and any residual supernatant was removed from each well. A solution of 2N sodium hydroxide (HP3) was diluted down to 0.1N and 32  $\mu$ L was added to each well before shaking for 5 minutes at 1800 rpm to resuspend the beads. The plate was placed back on the magnetic stand until the solution cleared before 30  $\mu$ L of each well was transferred to the corresponding well of a new 96-well reaction plate labeled Normalization Library Plate (NLP).

#### 2.5.5 Pooling and Denaturing the Libraries

To pool the libraries, 5  $\mu$ L of each well of the NLP was transferred to a 1.5 mL microcentrifuge tube. 7  $\mu$ L of this library pool was then transferred to a new tube containing 591  $\mu$ L Hybridization Buffer (HT1) and 2  $\mu$ L Human Sequencing Control (HSC) mixture (2  $\mu$ L HSC, 2  $\mu$ L HP3, and 36  $\mu$ L water). The tube was then heated at 96°C for 2 minutes for denaturation, inverted several times, and immediately placed in an ice-

water bath for 5 minutes before loading the full volume into the MiSeq FGx™ Reagent Cartridge for sequencing.

## **2.6 MiSeq FGx™ Sequencing**

DNA sequencing was carried out on the MiSeq FGx™ instrument using the forensic genomics run type and micro flow cell, which allows for a maximum of 32 single-source samples using DPMA, or 12 mixture/case samples. Each run consisted of 12 pooled mixture libraries (four 1:1, four 1:4, and four 1:9) as well as the positive and negative controls for a total of 14 libraries per run. There are 398 total sequencing cycles (each sequencing cycle reads one nucleotide base) consisting of 4 reads: Read 1, Index 1, Index 2, and Read 2. Read 1 is 351 cycles that sequences the first 351 bases in each of the forensic STR and SNP amplicons. Index 1 and Index 2 are both 8 cycle reads that sequences the i7 index and i5 index, respectively, for sample determination. Lastly, Read 2 sequences the last 31 nucleotides of the amplicons in the reverse direction of Read 1 to aid in sequence alignment and to sequence any amplicons longer than 351 bp.

Samples were analyzed using Verogen's ForenSeq™ Universal Analysis Software (UAS) version 1.3.6767 (Verogen, Inc., San Diego, CA) using the manufacturer set stutter filter percentages. The analytical threshold was set at 1.5% of the total reads of a locus and the interpretation threshold was set at 4.5% of the total reads of a locus as recommended by the manufacturer.

## **2.7 Contributor Ratio Determination**

For both the CE and sequencer, contributor ratios of autosomal STRs were calculated using the RFU (for the CE) and the read count (for sequencing) of each allele at

loci chosen for analysis. Loci were chosen for contributor ratio calculations only if they contained 3 or 4 alleles at a given locus as shown in Table 1, otherwise it would not have been possible to distinguish the major and minor alleles. When 4 alleles were present at a locus, all 4 were used in calculating contributor ratios. However, in cases where a minor allele dropped out, the allele still present was doubled to replace the missing allele. When only 3 alleles were present, it could mean both contributors are heterozygous and share one allele, or one contributor is homozygous and the other is heterozygous with no shared alleles. In the first case, it was not possible to determine how much of the shared allele each individual contributed, so only the unshared alleles were used in the calculations. In cases where both minor alleles dropped out of a 4-allele locus or the unshared minor allele dropped out of a 3-allele locus, that locus was excluded in the use of contributor ratio calculations.

**Table 1. Equations used for contributor ratio calculations of various autosomal STR combinations at an individual locus.** The RFU or read count of alleles A, B, C, and D are represented as  $\varphi_A$ ,  $\varphi_B$ ,  $\varphi_C$ ,  $\varphi_D$ , respectively.

Major Contributor Genotype	Minor Contributor Genotype	Contributor Ratio Calculation
AB	CD	$\frac{\varphi_A + \varphi_B}{\varphi_C + \varphi_D}$
AB	C + Dropout	$\frac{\varphi_A + \varphi_B}{2\varphi_C}$
AB	BC	$\frac{\varphi_A}{\varphi_C}$
AA	BC	$\frac{\varphi_A}{\varphi_B + \varphi_C}$
AB	CC	$\frac{\varphi_A + \varphi_B}{\varphi_C}$

For each replicate, the mean contributor ratio over all the loci that fit the criteria described above was calculated. For each run, the mean contributor ratio of each locus was calculated using the 4 replicates in that run and the mean contributor ratio over all replicated was calculated. Then the mean contributor ratio was calculated for each locus and all replicates over both runs. A standard deviation of each mean was also calculated and represented as error bars in the figures or reported with the mean.

The ForenSeq™ kit also includes SNPs in its primer set and the SNPs were looked at to determine if they could give a more accurate indication of mixture ratio than the autosomal STR loci. SNPs do not have the same variability as STRs as they only look at a single base pair difference, and only one of two bases are seen at a specific locus. For example, the rs735155 SNP can only have an A or G. This makes any overlap between the two individuals impossible to distinguish the major and minor contributor.

Therefore, SNP loci were chosen only if both individuals were homozygous for the two different bases.

### **3. Results**

#### **3.1 Sample Selection**

DNA was isolated from 7 anonymous donors. The two individuals that were selected for two person mixtures were individual 434 and individual 438 because they were found to have 18 and 20 heterozygous loci out of the possible 21, respectively (Table 2). This high level of heterozygosity also resulted in the overlap of 14 alleles, which might result in allele differences based on sequence.

#### **3.2 Capillary Electrophoresis**

Each mixture was amplified in quadruplicate twice, over two different days for between run and within run variation. In the first set of amplifications, however, one of the 1:4 mixtures did not amplify and resulted in only 3 replicates for this run. It was found that 18 of the 21 autosomal STRs fit the criteria for contributor ratio calculations (as described in Section 2.7) and these loci (D1S1656, TPOX, D2S441, D2S1338, FGA, D5S818, CSF1PO, D7S820, D8S1179, vWA, D12S391, D13S317, D16S539, D18S51, D19S433, D21S11, D22S1045, and SE33) were used in the results that follow.

##### **3.2.1 1:1 Mixtures**

The 1:1 mixture was prepared to be equal amounts of the 434 and 438 contributors' DNA based on the quantification results; therefore, the contributor ratio was expected to be approximately 1. However, it was apparent in each run that 434 was

present at a higher concentration and was treated as the major contributor for all calculations of the 1:1 mixture.

The mean contributor ratio for amplifications 1-4 (n=72 loci) and amplifications 5-8 (n=72 loci) were calculated to be  $1.862 \pm 0.744$  and  $1.774 \pm 0.595$ , respectively. Overall, the mean contributor ratio for all 1:1 samples (n=144) came out to  $1.818 \pm 0.675$ . Of the 18 loci the overall contributor ratio was consistent across all loci (Figure 3). A two-tailed t-test with an  $\alpha$ -level of 0.05 was performed and it was found that there were no loci that were significantly different from the overall mean contributor ratio. On top of consistency among loci, Figure 4 shows that there was high consistency of contributor ratios between replicates. A t-test also confirmed that there was no significant difference among replicates.

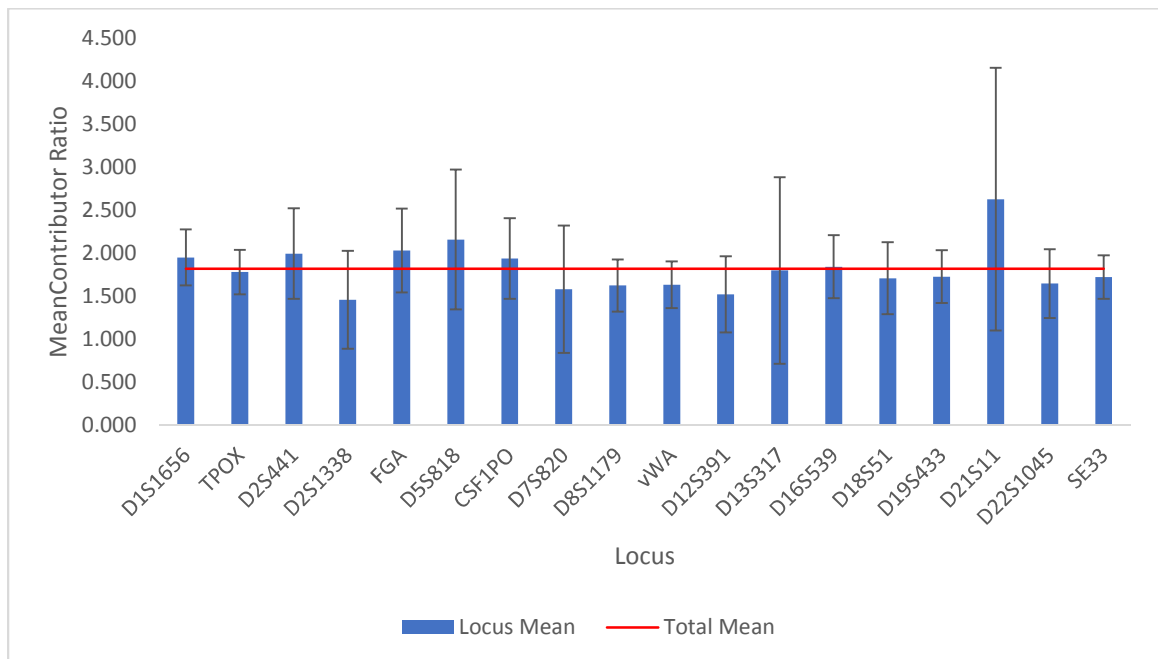
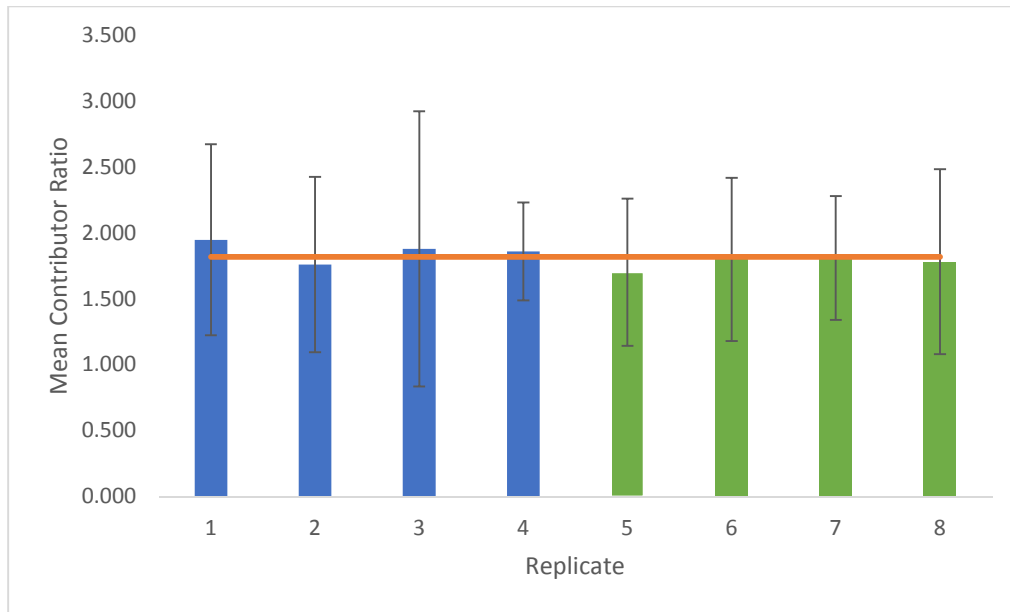


Figure 3: Mean contributor ratio of 1:1 mixture samples by locus run on the CE.

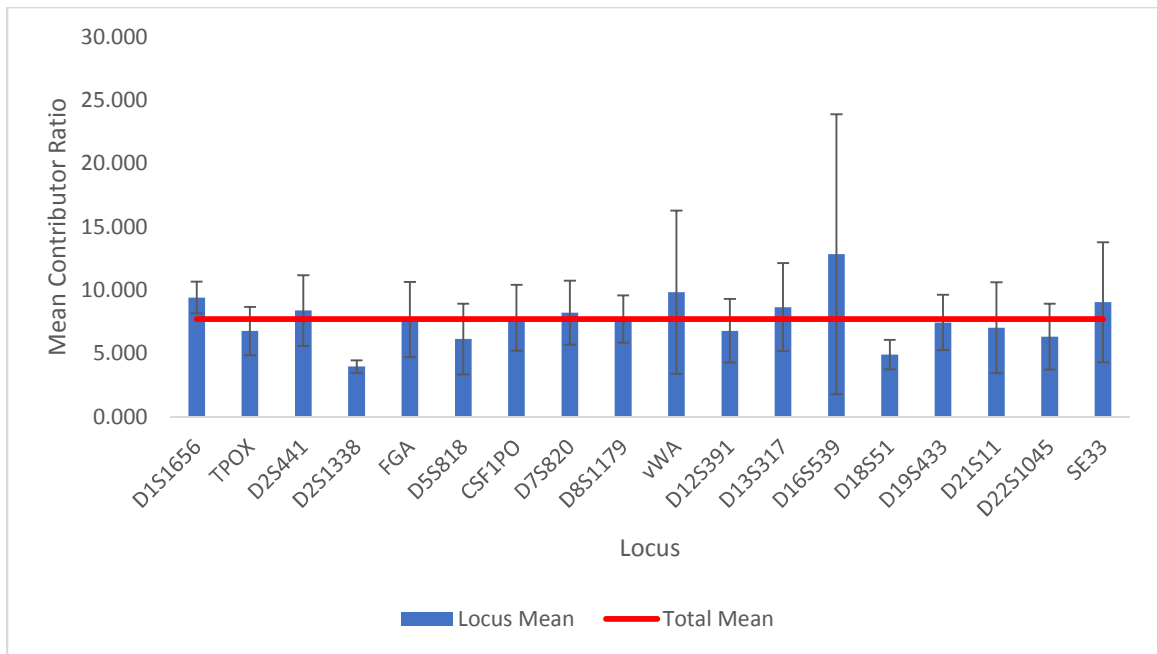


**Figure 4: Mean contributor ratio across replicates of 1:1 mixture samples run on the CE.** The blue bars represent those in amplification 1, the green bars represent those in amplification 2, and the orange line represents the mean contributor ratio of all replicates at 1:1.

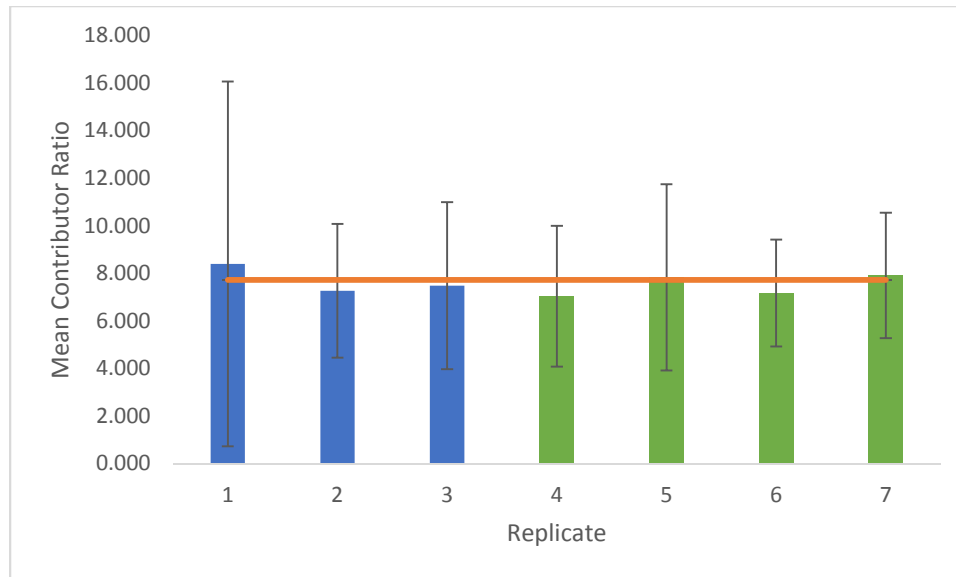
### 3.2.2 1:4 Mixtures

The mean contributor ratio for amplifications 1-3 (n=54) and amplifications 5-8 (n=72) were found to be  $8.040 \pm 5.649$  and  $7.483 \pm 3.030$ , respectively. Overall, the mean contributor ratio for all samples (n = 126) was  $7.722 \pm 4.359$ , while the expected ratio was 4. Variability across loci appeared to be slightly increased in the 1:4 samples (Figure 5) compared to the 1:1 and D16S539 had an unusually high mean contributor ratio and an equally high variation when compared to other loci ( $12.833 \pm 11.054$ ). In the other direction, both D2S1338 ( $3.966 \pm 0.491$ ) and D18S51 ( $4.918 \pm 1.158$ ) performed very well when compared to the expected value of 4 for this mixture. However, the contributor ratio of these two loci and D1S1656 was found to be significantly lower and higher than the mean contributor ratio, respectively. Similar to the 1:1 samples, the 1:4 replicates showed good reproducibility across all amplifications (Figure 6), but the

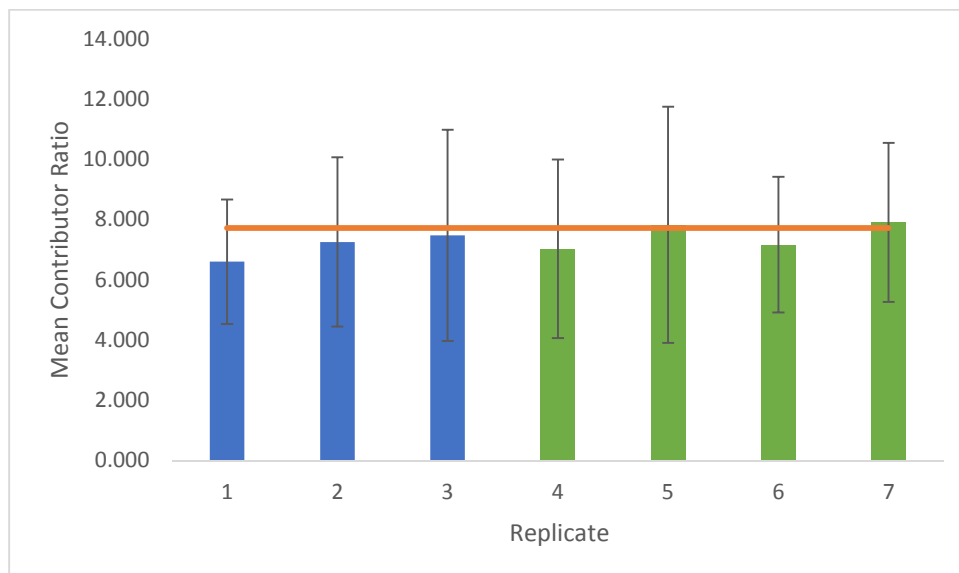
variation of replicate 1 is very apparent and likely caused by a very high ratio of 38.897 at D16S539. A box and whisker plot showed that this value was an outlier, and when removed, accounted for much of the variance (Figure 6). Before removing the outlier, none of the replicates showed a difference from the mean but removing D16S539 caused replicate 1 to be significantly different. A value of 25.103 at the vWA locus of replicate 2 was also determined to be an outlier and updated in Figure 7. Across all samples, 2 allele dropouts occurred: one at D7S820 and one at SE33. Both dropouts occurred during the second amplification and were at a four-allele locus which meant the contributor ratio was still able to be calculated.



**Figure 5: Mean contributor ratio of 1:4 mixture samples by locus run on the CE.**



**Figure 6: Mean contributor ratio across replicates of 1:4 mixture samples run on the CE.** The blue bars represent those in amplification 1, the green bars represent those in amplification 2, and the orange line represents the mean contributor ratio of all replicates at 1:4.

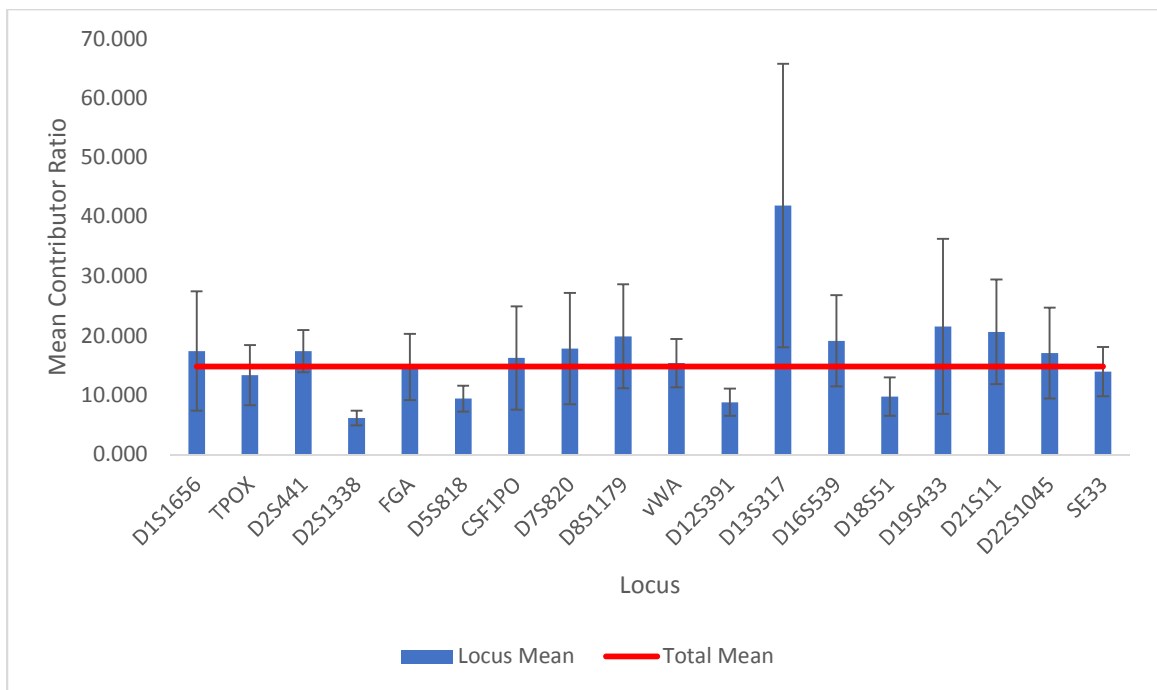


**Figure 7: Mean contributor ratio across all replicates of 1:4 mixture samples minus D16S539 locus in replicate 1 run on the CE.** The blue bars represent those in amplification 1, the green bars represent those in amplification 2, and the orange line represents the mean contributor ratio of all replicates at 1:4.

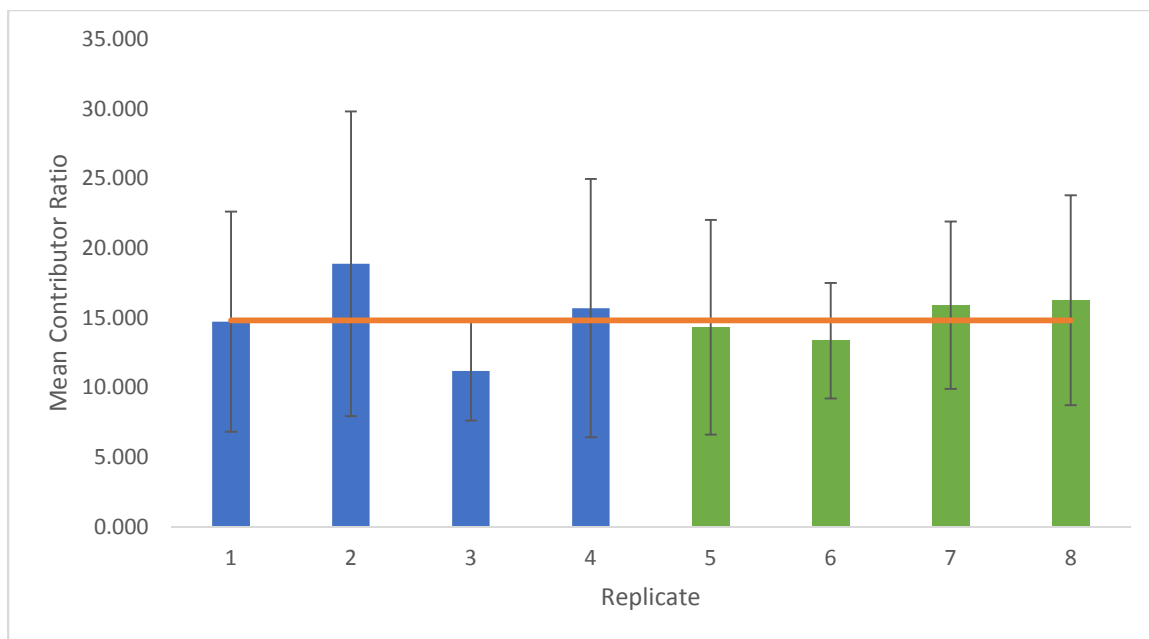
### 3.2.3 1:9 Mixtures

Where the 1:4 samples only had two allele dropouts, the 1:9 samples had 18 allele dropouts which resulted in the loss of 7 loci for contributor ratio calculations. The mean

values for all amplifications were very consistent, but the standard deviation was extremely high in all cases. The mean for amplifications 1-4 (n=67) was  $14.689 \pm 8.575$ ,  $14.955 \pm 6.548$  for amplifications 5-8 (n=70) and was  $14.827 \pm 7.592$  for the overall mean contributor ratio (n=137), almost double the expected ratio of 9. This can be seen in Figure 8 by the large variance across all loci, most notably D13S317. With the large amount of dispersion, 5 loci were found to be significantly different from the mean: D2S1338, D5S818, D12S391, D13S317, and D18S51. When it came to reproducibility, the mean contributor ratios between runs were not significantly different from each other but replicate 3 was significantly different from the overall mean (Figure 9).



**Figure 8: Mean contributor ratio of 1:9 mixture samples by locus run on the CE.**



**Figure 9: Mean contributor ratio across replicates of 1:9 mixture samples run on the CE.** The blue bars represent those in amplification 1, the green bars represent those in amplification 2, and the orange line represents the mean contributor ratio of all replicates at 1:9.

### 3.3 MiSeq FGx™

Using the ForenSeq™ Signature Prep Kit, each mixture was amplified in quadruplicate on two different days and sequenced on the Illumina MiSeq™. Of the 27 autosomal loci amplified in the ForenSeq™ kit, 23 were determined to be suitable for contributor ratio calculations: D1S1656, TPOX, D2S441, D2S1338, D3S1358, D4S2408, FGA, D5S818, CSF1PO, D6S1043, D7S820, D8S1179, vWA, D12S391, D13S317, PentaE, D16S539, D18S51, D19S433, D20S482, D21S11, PentaD, D22S1045 (Table 3). The D22S1045 locus was consistently significantly higher in all mixture ratios and was not used for mean contributor ratio calculations. These 23 loci included all the 18 used for calculations with the CE samples, except for SE33. An issue during the first round of



### 3.3.1 1:1 Mixtures

The overall autosomal STR mean contributor ratio (n=154) was calculated to be  $1.799 \pm 0.475$  and the SNP mean contributor ratio (n=70) was  $1.787 \pm 0.462$ . There was little variation between amplifications, with mean value of run 1 (n= 66) equaling  $1.780 \pm 0.463$  and run 2 (n=88)  $1.814 \pm 0.484$  for the STR loci and  $1.760 \pm 0.422$  and  $1.807 \pm 0.489$  for the SNP loci for run 1 (n=30) and run 2 (n=40), respectively. While the overall variation was low, the mean contributor ratio by locus was surprisingly variable (Figure 13). Of the 22 autosomal loci, 9 of the loci were calculated to be significantly different from the mean and 3 of the 10 SNP loci were significantly different from their mean (Figure 14). However, between amplifications, none of the autosomal STRs (Figure 15) or the SNPs (Figure 16) showed a significant difference from the mean.

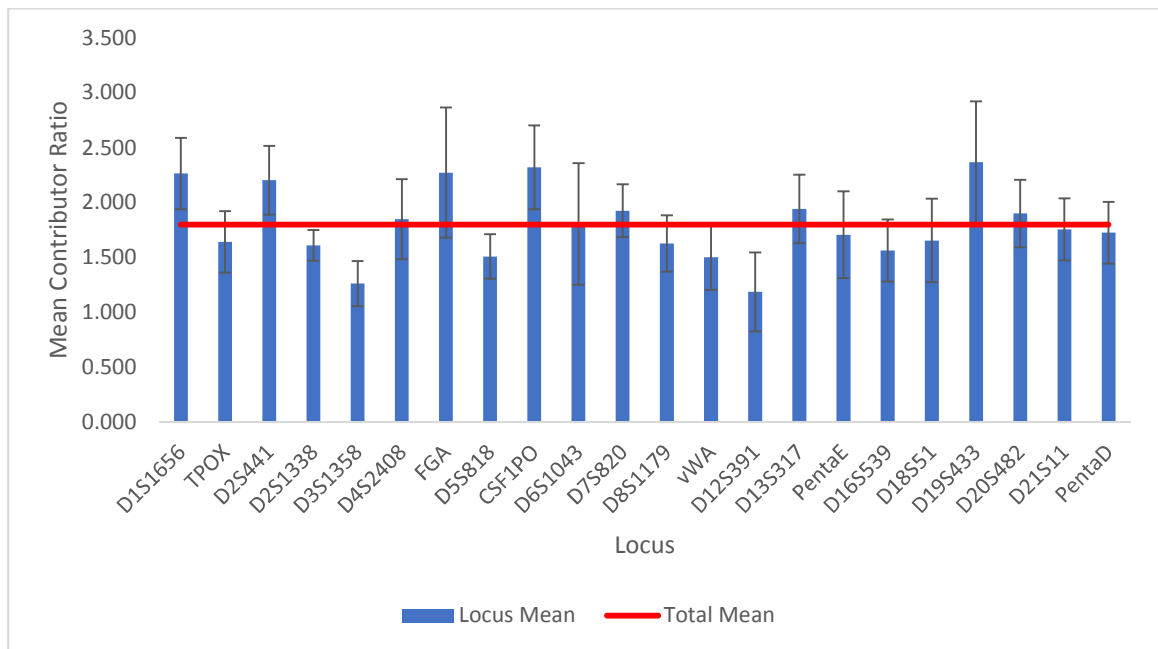
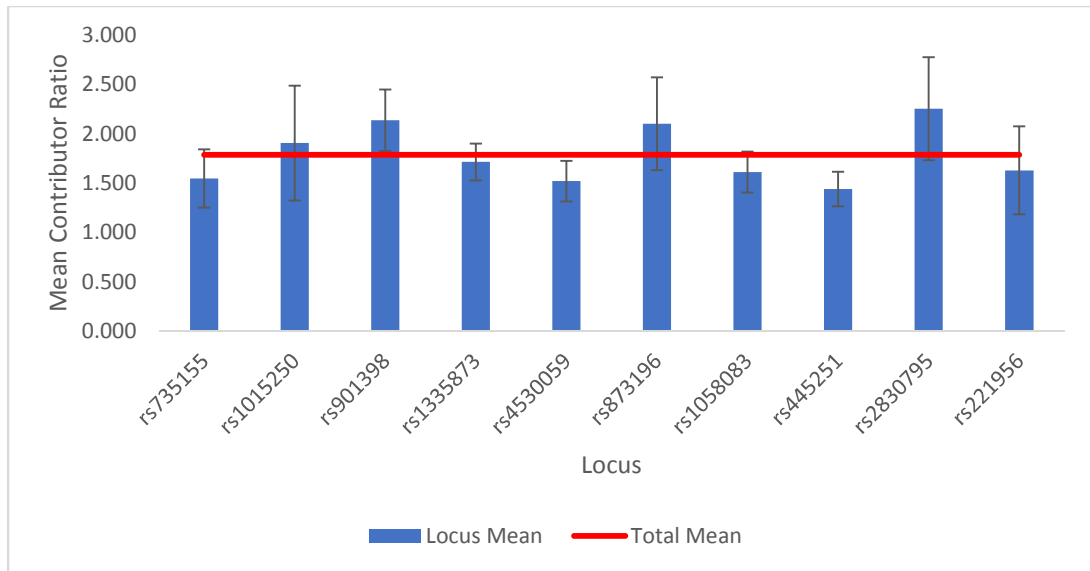
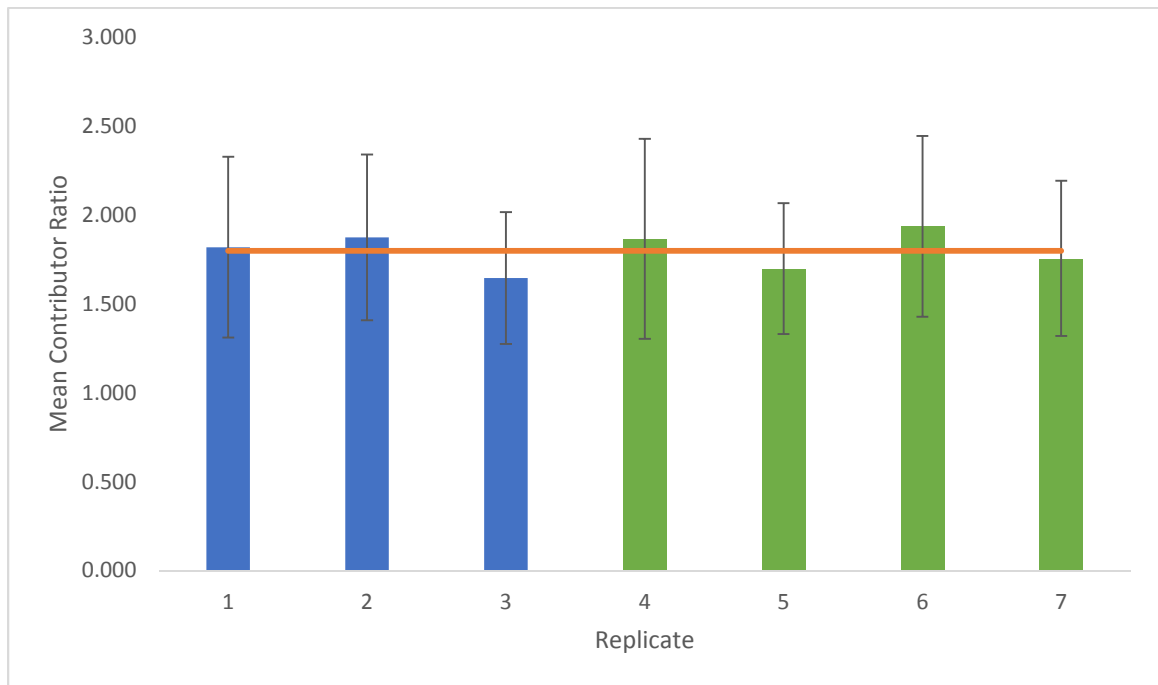


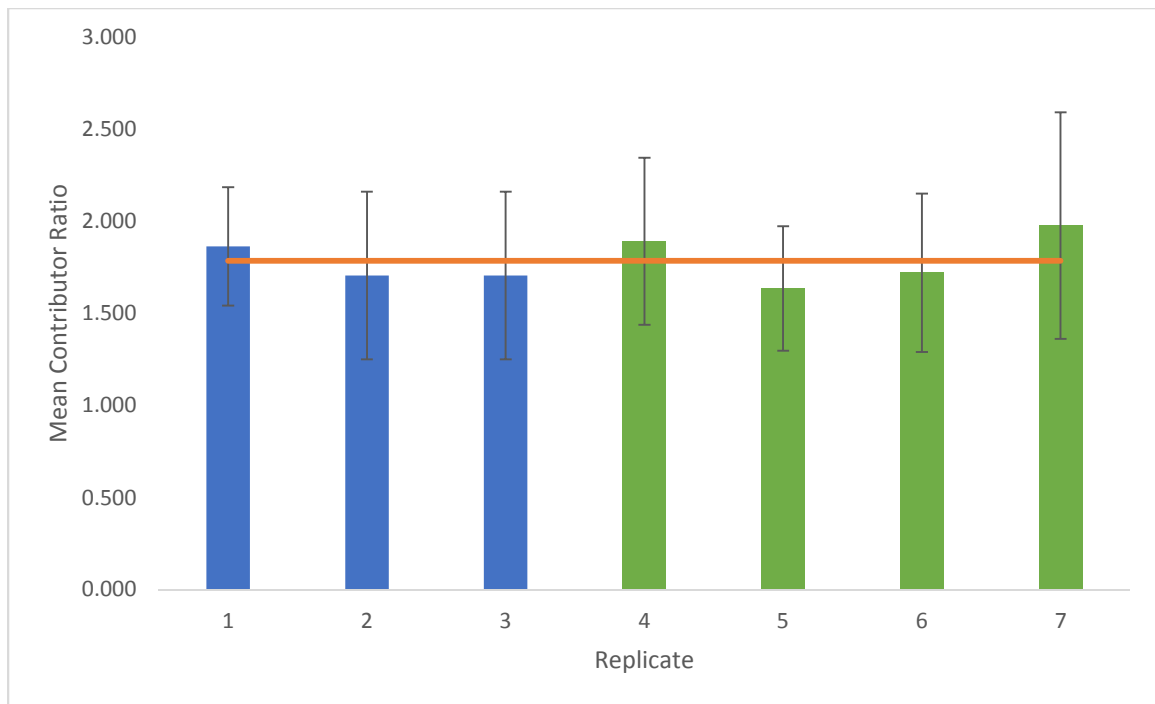
Figure 13: Mean contributor ratio of 1:1 mixture samples by STR locus run on the MiSeq FGx™.



**Figure 14: Mean contributor ratio of 1:1 mixture samples by SNP locus run on the MiSeq FGx™.**



**Figure 15: Mean contributor ratio of STRs across replicates of 1:1 mixture samples run on the MiSeq FGx™.** The blue bars represent those in amplification 1, the green bars represent those in amplification 2, and the orange line represents the mean contributor ratio of all STR replicates at 1:1.



**Figure 16: Mean contributor ratio of SNPs across replicates of 1:1 mixture samples run on the MiSeq FGx™.** The blue bars represent those in amplification 1, the green bars represent those in amplification 2, and the orange line represents the mean contributor ratio of all SNP replicates at 1:1.

### 3.3.2 1:4 Mixtures

As with the CE, dropouts occurred at 1:4 ratio. A total of eight STR alleles and one of the SNP alleles dropped out, but only one STR locus was unable to be recovered for contributor ratio calculations. The variance increased greatly from the 1:1 samples, but it was lower than the CE variance at 1:4. The overall mean contributor ratio (n=153) of the autosomal STRs was  $7.452 \pm 3.515$ . The mean contributor ratio for the SNPs (n=69) was  $8.726 \pm 3.473$ , which was significantly higher than the STRs. Even though the variation increased, there was no difference in contributor ratio of the STR loci ( $7.783 \pm 3.612$  and  $7.452 \pm 3.515$ ) or the SNP loci ( $9.235 \pm 4.202$  and  $8.334 \pm 2.721$ ) between runs. D3S1358, D6S1043, D12S391, and D18S51 were all around the expected ratio of 4 (Figure 17) but were also significantly different from the mean along with 4 additional

loci. This could be explained by the minor contributor having an allele in stutter position at all these loci, but since the sequences of the stutter and minor alleles were the same, they could not be separated. Figure 18 shows the consistency by replicate between each other and compared to the mean. While the SNPs were significantly higher than the STRs, they were very consistent among themselves (Figure 19) and had only one locus that was significantly different (Figure 20).

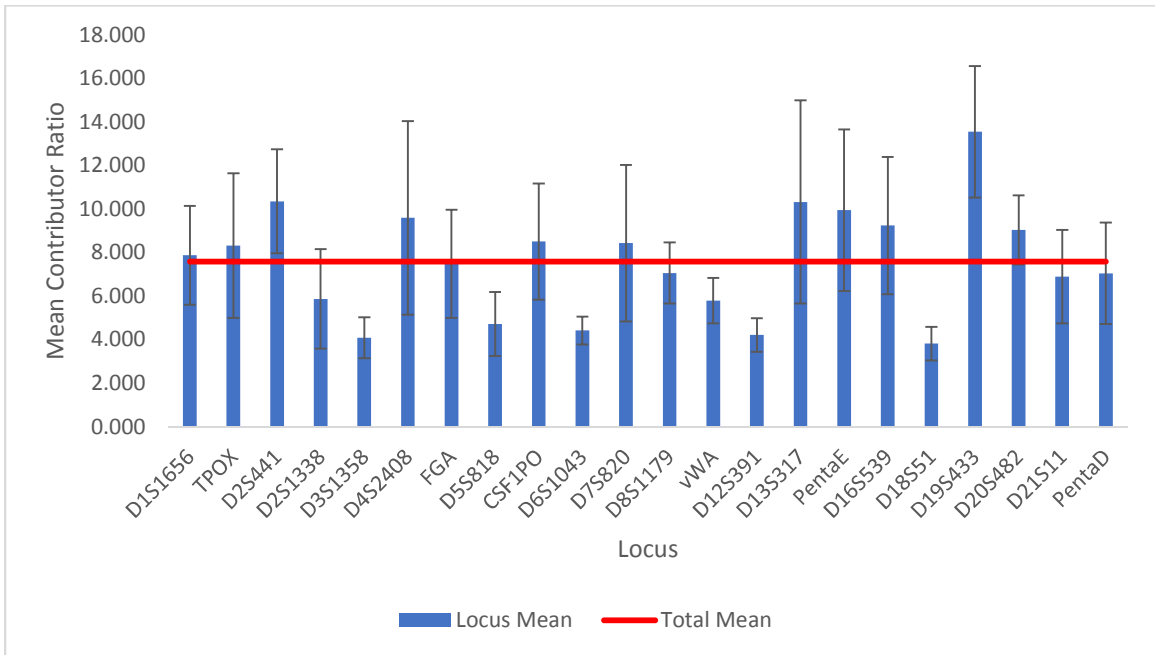
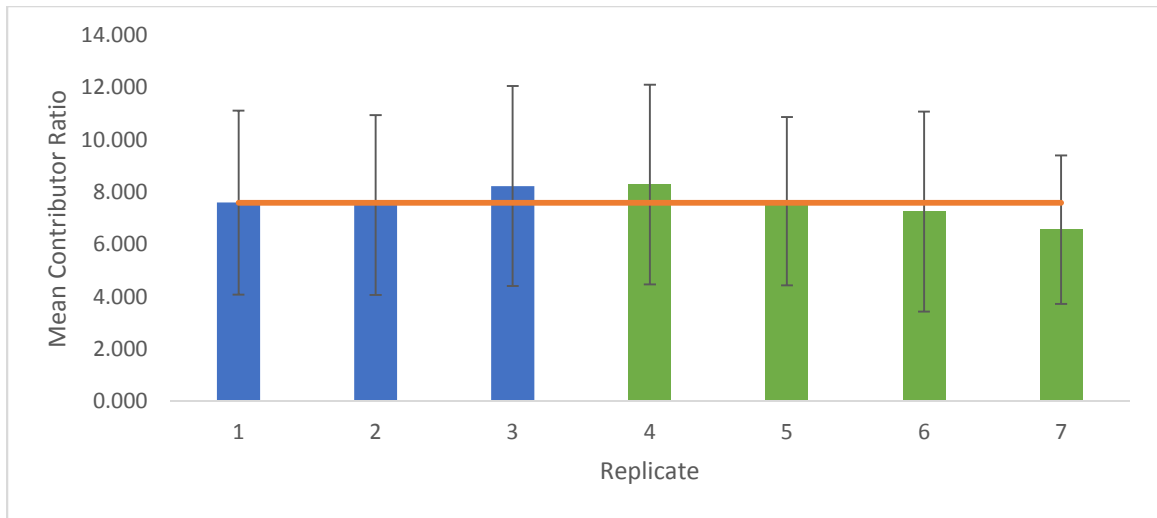
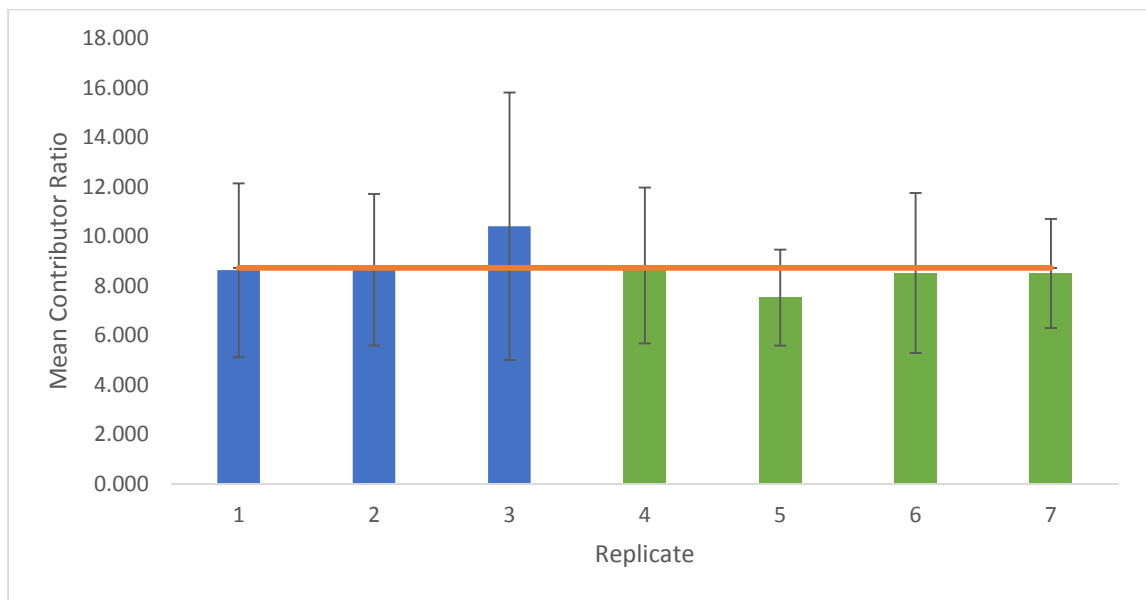


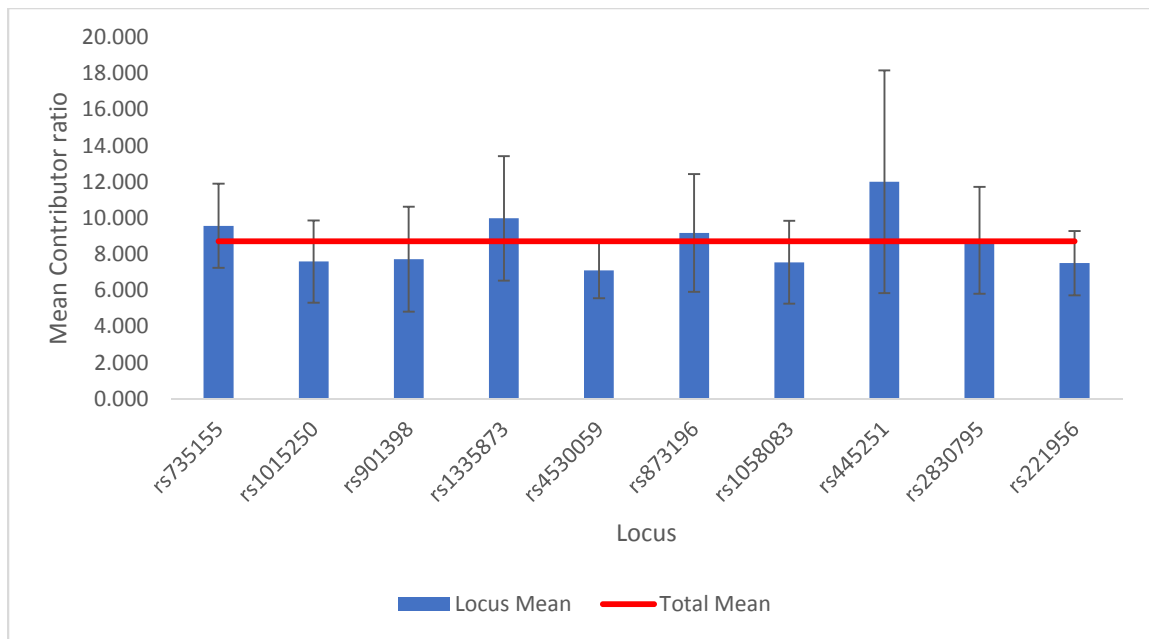
Figure 17: Mean contributor ratio of 1:4 mixture samples by STR locus run on the MiSeq FGx™.



**Figure 18: Mean contributor ratio of STRs across replicates of 1:4 mixture samples run on the MiSeq FGx™.** The blue bars represent those in amplification 1, the green bars represent those in amplification 2, and the orange line represents the mean contributor ratio of all STR replicates at 1:4.



**Figure 19: Mean contributor ratio of SNPs across replicates of 1:4 mixture samples run on the MiSeq FGx™.** The blue bars represent those in amplification 1, the green bars represent those in amplification 2, and the orange line represents the mean contributor ratio of all SNP replicates at 1:4.

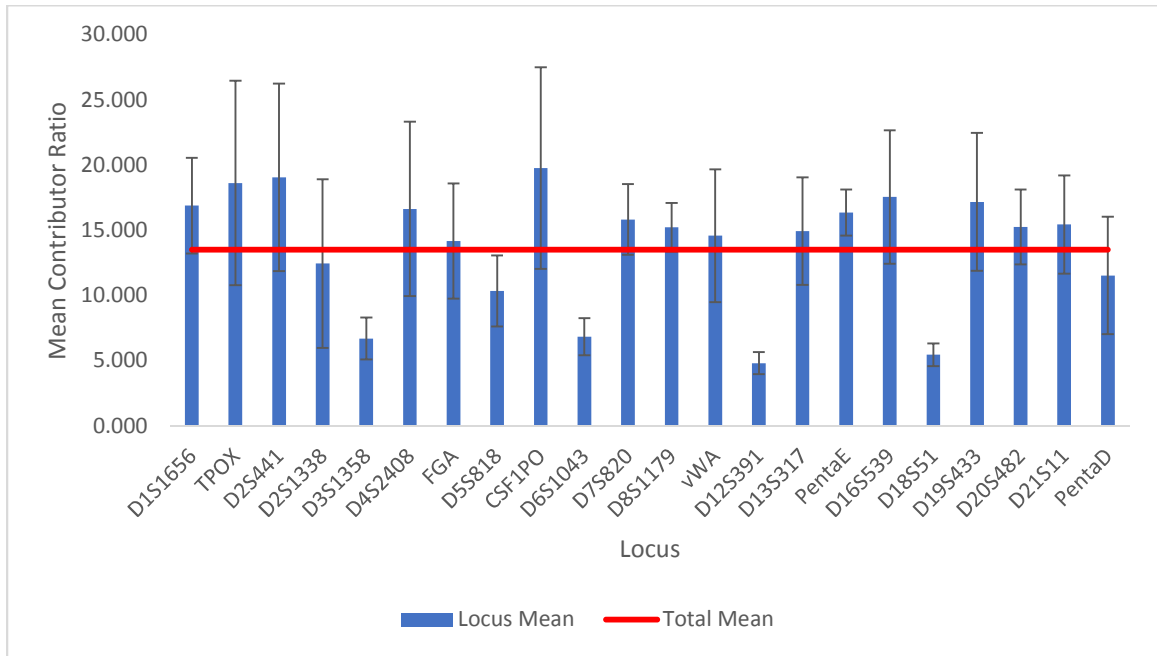


**Figure 20: Mean contributor ratio of 1:4 mixture samples by SNP locus run on the MiSeq FGx™.**

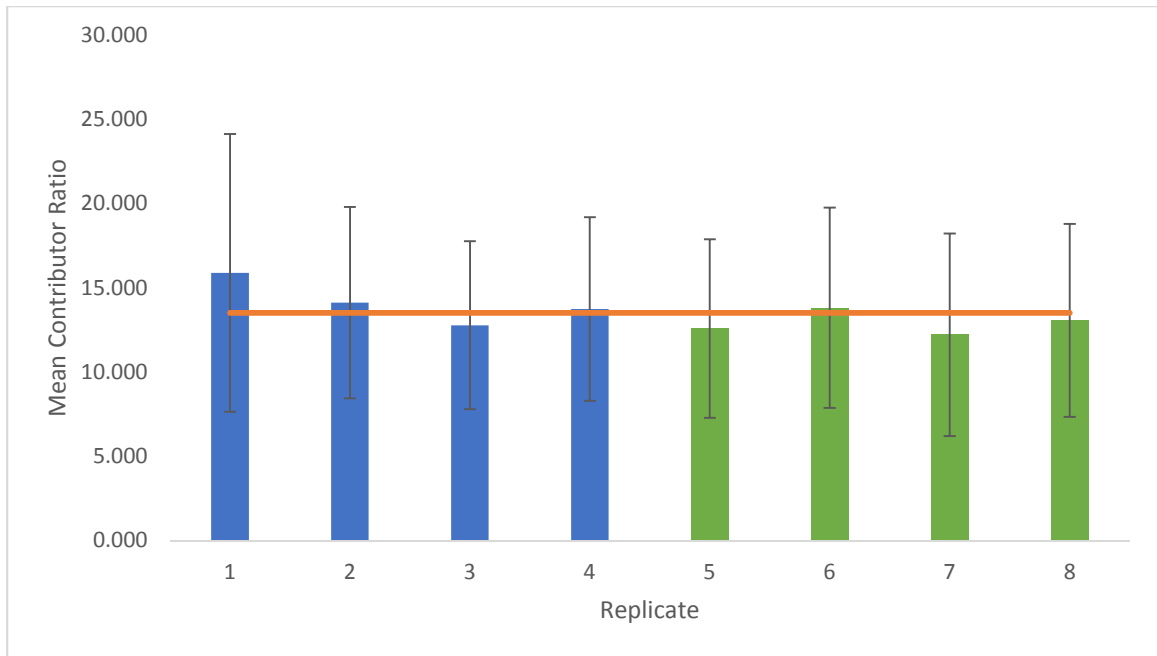
### 3.3.3 1:9 Mixtures

The 1:9 samples had many allele dropouts. 61 STR alleles and 2 SNP alleles dropped out resulting in the loss of 22 loci for contributor ratio calculations. The overall (n=162) mean contributor ratio for the STRs was  $13.524 \pm 6.063$  and the SNPs (n=78) were significantly higher at  $21.077 \pm 9.265$ . Like the 1:4 mixtures, D3S1358, D6S1043, D12S391, and D18S51 were significantly lower than all other loci (Figure 21) and there were 4 other loci significantly different from the mean contributor ratio. However, none of the replicates were significantly different (Figure 22). This is also shown in the SNP replicates, even though they are significantly higher than the STRs (Figure 23). Where

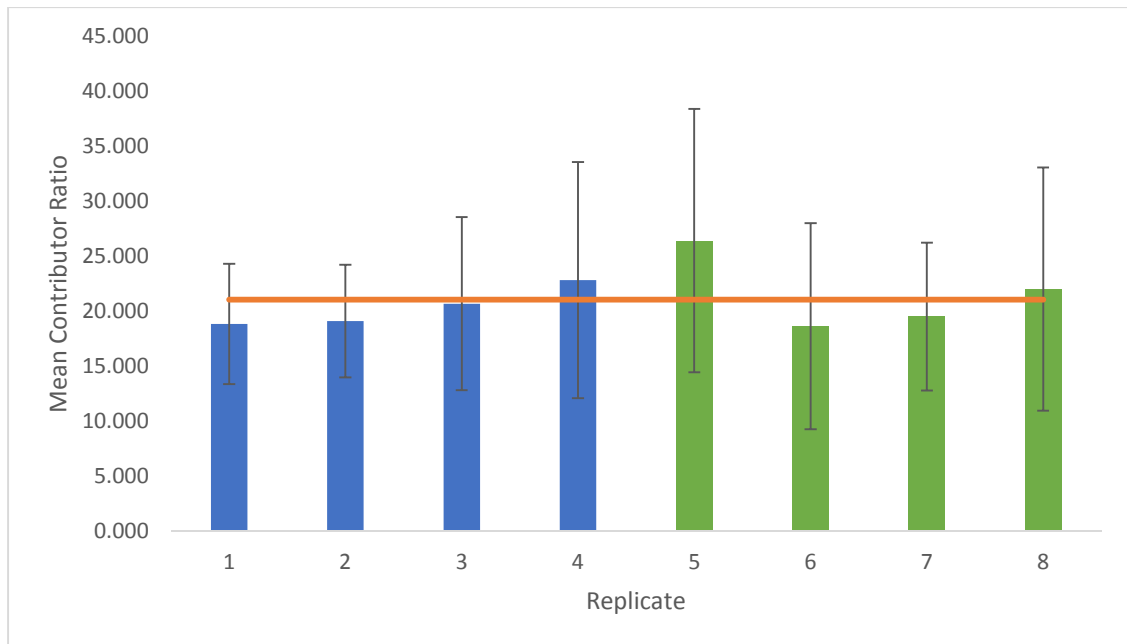
the STRs had high variability by locus, the SNPs had no significant difference between loci for 1:9 mixtures (Figure 24).



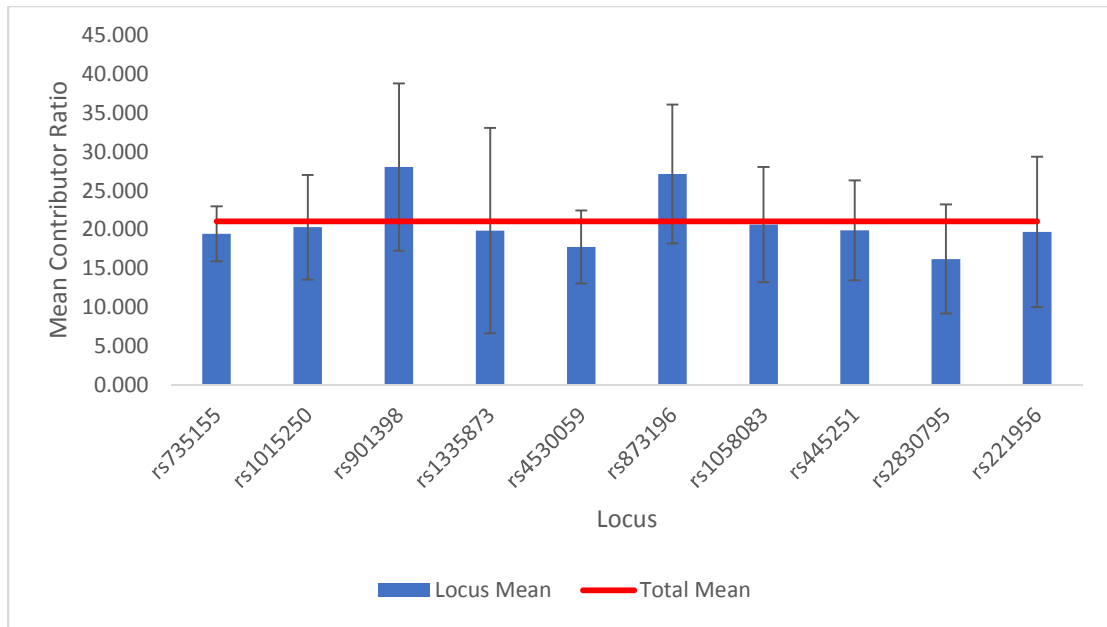
**Figure 21: Mean contributor ratio of 1:9 mixture samples by STR locus run on the MiSeq FGx™**



**Figure 22: Mean contributor ratio of STRs across replicates of 1:9 mixture samples run on the MiSeq FGx™.** The blue bars represent those in amplification 1 and the green bars represent those in amplification 2, and the orange line represents the mean contributor ratio of all STR replicates at 1:9.



**Figure 23: Mean contributor ratio of SNPs across replicates of 1:9 mixture samples run on the MiSeq FGx™.** The blue bars represent those in amplification 1, the green bars represent those in amplification 2, and the orange line represents the mean contributor ratio of all SNP replicates at 1:9.

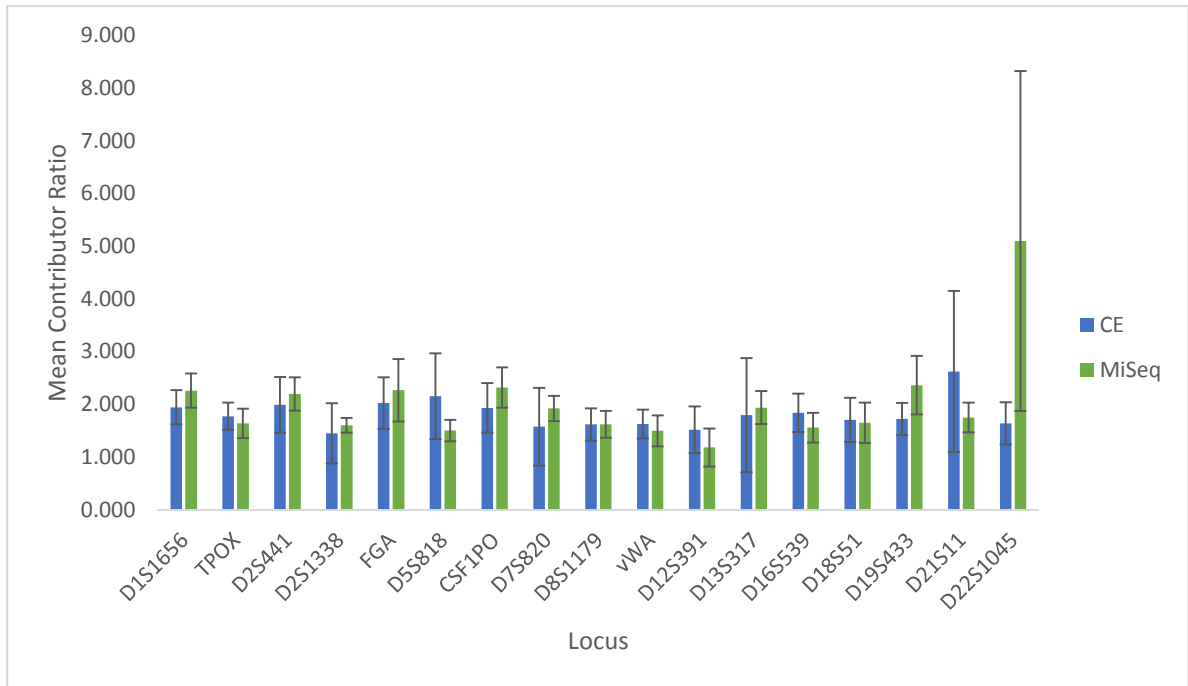


**Figure 24: Mean contributor ratio of 1:9 mixture samples by SNP locus run on the MiSeq FGx™**

### 3.4 CE Versus MiSeq™

Figures 25 and 26 show a comparison of the 1:1 samples on the CE and MiSeq™ by locus and the overall mean contributor ratios, respectively. The overall means were almost identical and the MiSeq FGx™ showed slightly less variance compared to the CE, although not significant (Table 2). Only two loci showed a significant difference between the two analysis methods, D19S433 and D22S1045, both of which were higher on the MiSeq™. The two loci were also significantly higher on the MiSeq FGx™ at 1:4 ratio as was D2S1338 (Figure 27); however, there was one locus, D12S391, that was significantly higher on the CE. When it came to overall mean contributor ratio, neither the CE nor MiSeq FGx™ was different from the other at the 1:4 ratio (Figure 28). Even at 1:9 ratio, the mean contributor ratio was consistent between the two methods, although the MiSeq™ showed a slightly smaller ratio and variance (Figure 29). Between loci the

number of differences increased to 5 at 1:9 (Figure 30). D22S1045 and D2S1338 were higher on the MiSeq™ than CE, but D18S51, D13317, and D12S391 were all higher on the CE.



**Figure 25: Comparison of mean contributor ratio of STR loci of 1:1 mixture samples on the CE versus MiSeq FGx™**

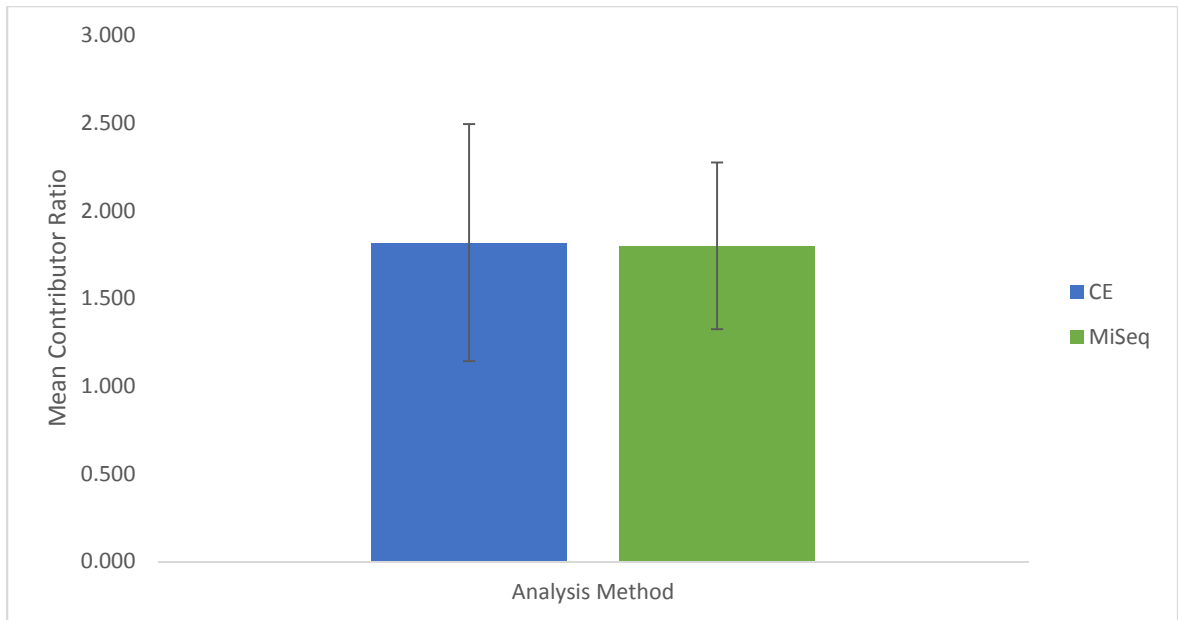


Figure 26: Total mean contributor ratio of all 1:1 mixture samples on the CE versus MiSeq FGx™

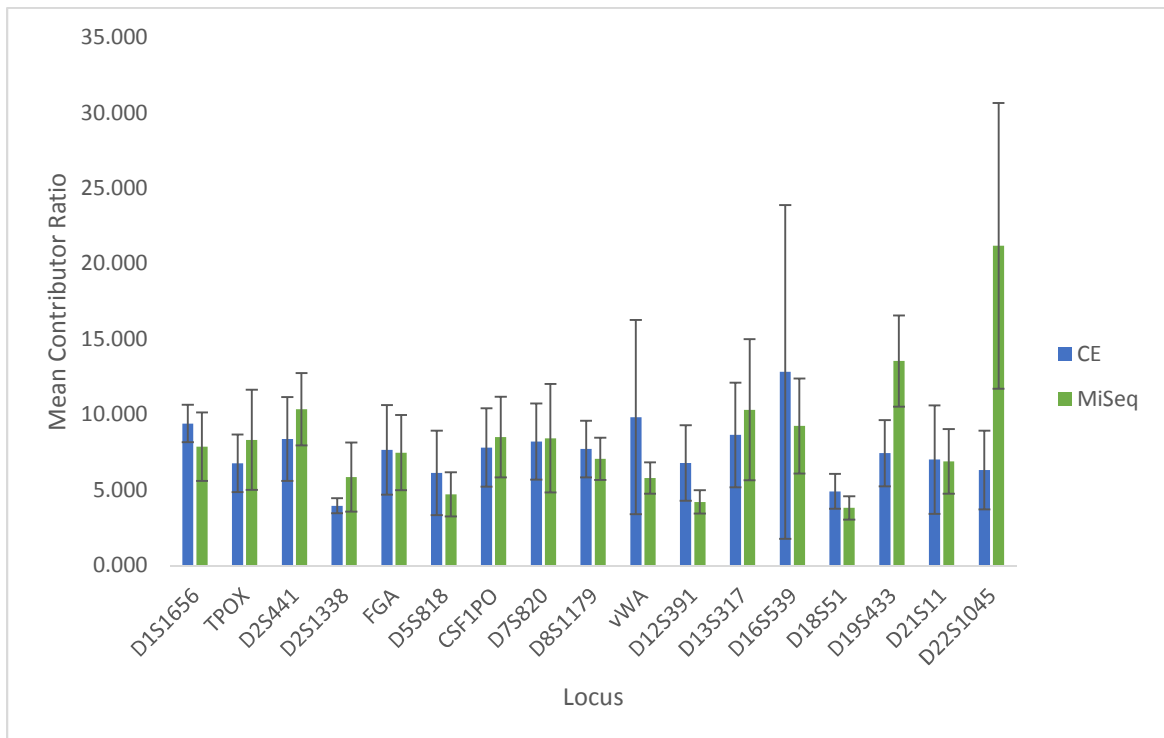
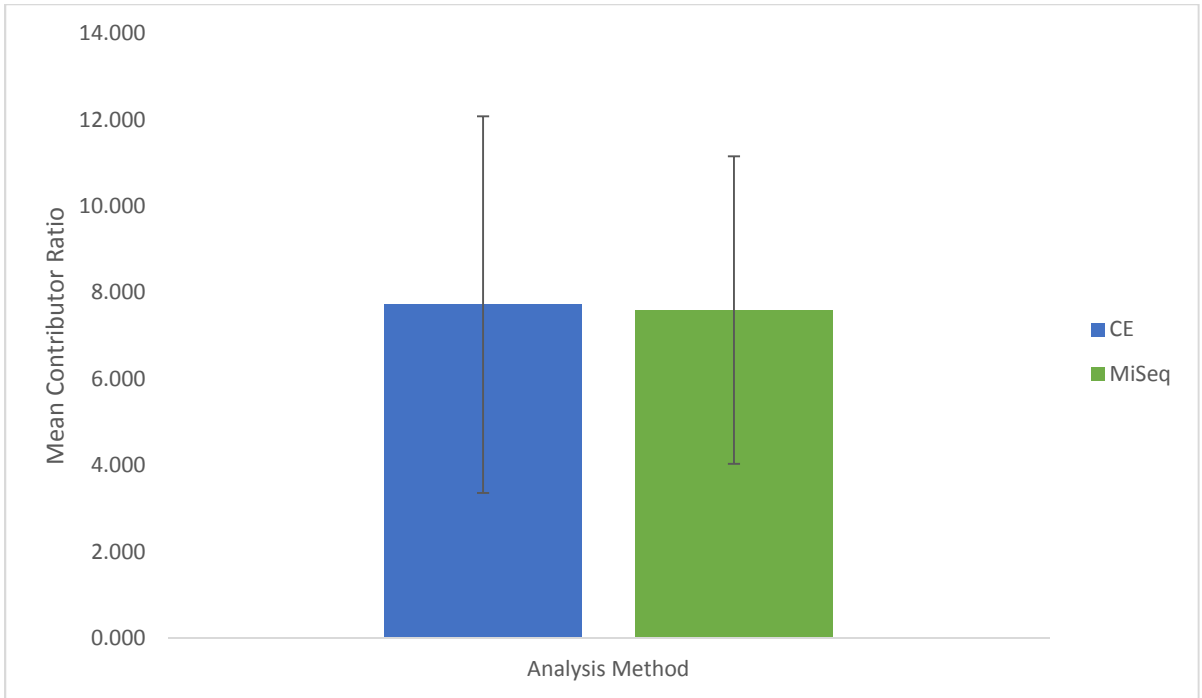
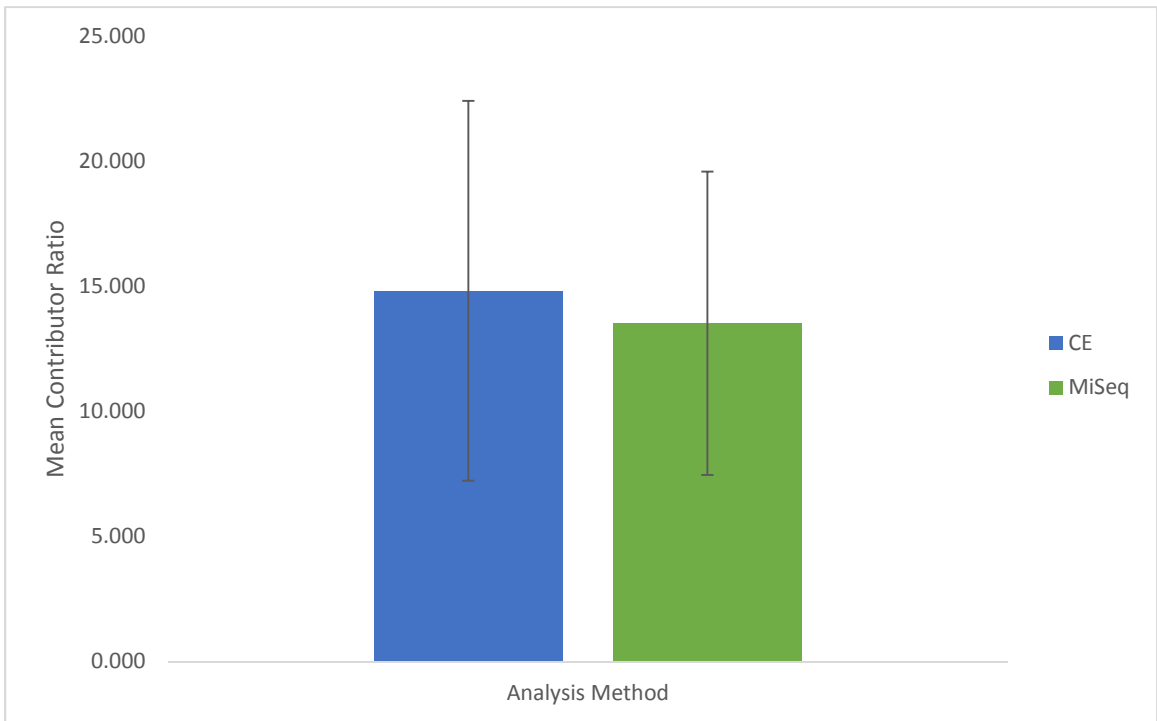


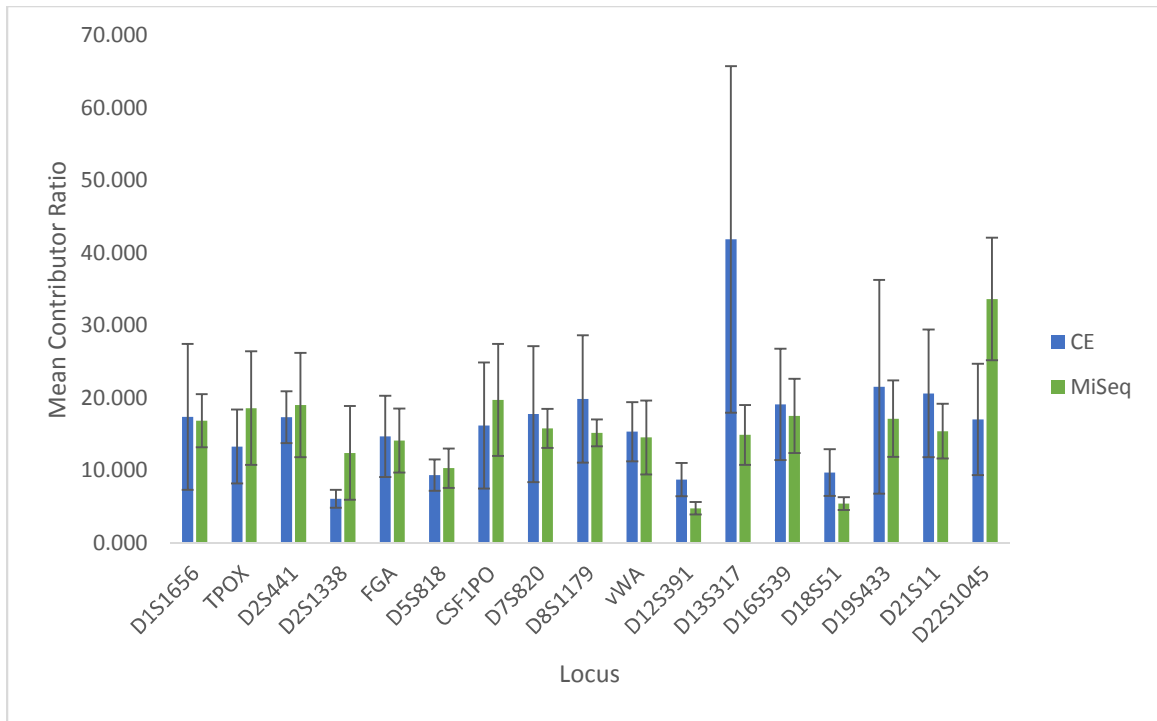
Figure 27: Comparison of mean contributor ratio of STR loci of 1:4 mixture samples on the CE versus MiSeq FGx™



**Figure 28: Total mean contributor ratio of all 1:4 mixture samples on the CE versus MiSeq FGx™**



**Figure 29: Total mean contributor ratio of all 1:9 mixture samples on the CE versus MiSeq FGx™**



**Figure 30: Comparison of mean contributor ratio of STR loci of 1:9 mixture samples on the CE versus MiSeq FGx™**

**Table 2: Mean contributor ratios and their standard deviations compared on the CE and MiSeq FGx™**

Analysis Method	CE		MiSeq FGx™	
	Mean	Standard Deviation	Mean	Standard Deviation
1:1	1.818	0.675	1.799	0.475
1:4	7.722	4.359	7.452	3.515
1:9	14.827	7.592	13.524	6.063

#### 4. DISCUSSION

NGS enables the sequencing of a large battery of forensic biomarkers including STR and SNP loci and could be the future of forensic DNA profiling. The MiSeq FGx<sup>TM</sup> system is at the forefront of development and implementation of DNA sequencing in forensics and has been shown to be concordant with fragment-length CE analysis while overcoming some of the limitations of CE. It has been demonstrated to meet the strict forensic validation guidelines set by SWGDAM to ensure the reproducibility and reliability of the system [4,31,32]. The data presented here demonstrates the reproducibility of the MiSeq FGx<sup>TM</sup> system and the benefits it can provide to mixture deconvolution. The MiSeq FGx<sup>TM</sup> system produces an enormous amount of data and this study was only focused on the autosomal STRs and SNPs. Additional X and Y-STRs could potentially be used to increase mixture resolution as more individualizing loci.

The mixtures were prepared based on quantitation results using RT-PCR to be 1:1, 1:4, and 1:9 at 0.2 ng/ $\mu$ L. However, the fact that both the CE and MiSeq<sup>TM</sup> results were consistent across all replicates and concordant with each other indicate that the major contributor, 434, was double the expected quantitative value, thus resulting in higher ratios. The concentrations determined by contributor ratios were found to be closer to 1:2, 1:8 and 1:14.

A total of 1 ng input DNA was used for sequenced-based STR profiling of all samples, with the minor contributor at 0.5 ng, 0.2 ng, and 0.1 ng for 1:1, 1:4, and 1:9 mixture ratios, respectively. There have been numerous studies that showed full profiles can be generated at all of these ratios or amounts [28,32,33] using the MiSeq FGx<sup>TM</sup>

system; however, in this study full STR profiles could not be generated at 1:9 ratio and only two out of seven replicates at 1:4 ratio produced a full profile. The higher than expected concentration of the major contributor is also a likely factor in the poor allele coverage of the minor contributor. Using the contributor ratio instead of the mixture ratio, it is reasonable to observe allele dropout at 1:4 (1:8 contributor ratio) and with no full profiles being generated at 1:9 (1:14 contributor ratio). In comparison, full profiles for five of the seven 1:4 mixtures (1:8 contributor ratio) and two of the eight 1:9 mixtures (1:14 contributor ratio) were generated on the CE platform. This is likely due to the less hands-on nature of the GlobalFiler™ amplification process.

The MiSeq FGx™ performance was found to be concordant with that of the GlobalFiler™-CE method in terms of contributor ratio. However, there was a total of 68 minor contributor alleles (9.09% of the total possible alleles) that were lost on the MiSeq FGx™ and only 20 on the CE (3.38% of the total possible alleles). Most of those alleles lost on the MiSeq FGx™ (61 or 8.15%) and CE (18 or 3.03%) were from the 1:9 (1:14 contributor ratio) mixture. This difference may seem large, but there were an additional 5 minor contributor alleles, per replicate, that could be identified by sequence. A total of 22 samples were analyzed for an increase in 110 potential alleles that could be used, not including six additional loci present in the ForenSeq™ kit but not in the GlobalFiler™ kit.

MiSeq FGx™ appeared to perform worse in terms of variation in the mean contributor ratio on a per locus-basis. For instance, high preferential amplification was observed at locus D22S1045 (see Section 3.4). So much so that even the major

contributor lost an allele in one of the 1:9 mixtures. The manufacturer mentioned this was possible and there have been other studies that report the same levels of preferential amplification at this locus [4]. Run by run and between replicates, however, the MiSeq FGx™ showed consistent contributor ratios and less variation in comparison to the CE.

The SNP loci are located on much shorter regions of DNA compared to STRs. At lower concentrations or in degraded DNA samples, shorter DNA strands are more likely to still be intact and be successfully amplified while longer DNA strands have a higher likelihood of being degraded and unable to be amplified. The SNPs were analyzed here to determine if they present a more accurate representation of mixture ratio at lower concentrations. Only 3 SNP alleles dropped out, one at 1:4 (1:8 contributor ratio) and two at 1:9 (1:14 contributor ratio); however, the contributor ratio values were significantly higher than STR loci and were therefore less reliable than the STRs.

## 5. CONCLUSIONS

The MiSeq FGx™ system was developed specifically for the use in forensic DNA typing. Since its release, there have been many studies evaluating its robustness, accuracy, and reproducibility. The largest advantage next generation sequencing has over traditional CE methods is the ability to determine sequence variants in alleles when looking at mixture samples. Between the two individuals sequenced in these mixtures, two minor contributor's alleles could to be separated by sequence variation and three of them could be distinguished from the stutter peak of the major contributor. In addition to the X- and Y-STRs and SNPs, the MiSeq FGx™ can provide a larger amount of information of mixture samples that cannot be obtained by the CE.

As the MiSeq FGx™ provides data in read counts and the CE in relative fluorescence units, the two cannot be directly compared to one another. Instead, the ratio of the two contributors on each instrument was calculated for data comparison. The data showed that the MiSeq FGx™ was concordant with the CE which should facilitate the introduction of sequence data as additional tool for use in forensic DNA testing.

## APPENDIX A:

**Table 3: Table of GlobalFiler™ alleles of mixture contributors.** Loci that were used in contributor ratio calculations are marked with a <sup>1</sup>.

Locus	434 Alleles	438 Alleles
D3S1358	17,17	16,17
vWA <sup>1</sup>	16,19	14,17
D16S539 <sup>1</sup>	9,12	10,12
CSF1PO <sup>1</sup>	10,11	11,12
TPOX <sup>1</sup>	8,9	9,11
Y-Indel	1	-
Amelogenin	X,Y	X,X
D8S1179 <sup>1</sup>	10,14	15,16
D21S11 <sup>1</sup>	29,30	30,30.3
D18S51 <sup>1</sup>	15,16	14,16
DYS391	10	-
D2S441 <sup>1</sup>	10,11.3	11,14
D19S433 <sup>1</sup>	13,14	14,15.2
TH01	9,9	8,9
FGA <sup>1</sup>	22,25	23,25
D22S1045 <sup>1</sup>	11,17	15,17
D5S818 <sup>1</sup>	7,12	11,12
D13S317 <sup>1</sup>	8,9	9,11
D7S820 <sup>1</sup>	11,11	12,13
SE33 <sup>1</sup>	16,27.2	17,25.2
D10S1248	13,15	13,13
D1S1656 <sup>1</sup>	12,13	15,16
D12S391 <sup>1</sup>	20,22	15,15
D2S1338 <sup>1</sup>	19,24	19,23

**Table 4: Table of ForenSeq™ alleles of mixture contributors.** Where the individual is homozygous in allele length and sequence are listed as one number. Where the individual has an allele of the same length but different sequence, the number is listed twice. Loci that were used in contributor ratio calculations are marked with a <sup>1</sup> and alleles shared by the two individuals but with different sequences are marked with an asterisk.

Locus	434 Allele	438 Allele
Amelogenin	X,Y	X,X
D1S1656 <sup>1</sup>	12,13	15,16
TPOX <sup>1</sup>	8,9	9,11
D2S441 <sup>1</sup>	10,11.3	11,14
D2S1338 <sup>1</sup>	19,24	19,23
D3S1358 <sup>1</sup>	17,17	16,17
D4S2408 <sup>1</sup>	8,10	9,10
FGA <sup>1</sup>	22,25	23,25
D5S818 <sup>1</sup>	7,12	11,12
CSF1PO <sup>1</sup>	10,11	11,12
D6S1043 <sup>1</sup>	13,14	12,14
D7S820 <sup>1</sup>	11	12,13
D8S1179 <sup>1</sup>	10,14	15,16
D9S1122	12	12,13
D10S1248	13,15	13
TH01	9	8,9
vWA <sup>1</sup>	16,19	14,17
D12S391 <sup>1</sup>	20,22	15,19
D13S317 <sup>1*</sup>	8,9	9,11
PentaE <sup>1</sup>	5,20	10,14
D16S539 <sup>1</sup>	9,12	10,12
D17S1301	12	11,12
D18S51 <sup>1</sup>	15,16	14,16
D19S433 <sup>1</sup>	13,14	14,15.2
D20S482 <sup>1</sup>	13,14	10,16
D21S11 <sup>1*</sup>	29,30	30,30.3
PentaD <sup>1</sup>	11	10,12
D22S1045 <sup>1</sup>	11,17	15,17

## BIBLIOGRAPHY

1. Lynch M. God's signature: DNA profiling, the new gold standard in forensic science. *Endeavour* 2003;27(2):93–7. Available from: <https://www.sciencedirect.com/science/article/pii/S0160932703000681#BIB6>
2. Jeffreys AJ, Wilson V, Thein SL. Individual-specific 'fingerprints' of human DNA. *Nature* 1985;316(6023):76–9. Available from: <http://www.nature.com/articles/316076a0>
3. Roewer L. DNA fingerprinting in forensics: past, present, future. *Investigative Genetics* 2013;4(1):22. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/24245688>
4. 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 Science International: Genetics* 2017;28:52–70.
5. Watson JD, Crick FHC. Genetical implications of the structure of deoxyribonucleic acid. *Nature* 1953;171(4361):964–7.
6. Nelson D, Cox M. *Principles of Biochemistry*. Sixth. New York: Susan Winslow, 2013;
7. Butler J. *Advanced Topics in Forensic DNA Typing: Methodology*. Waltham, MA: Elsevier, 2012;
8. Edwards A, Civitello A, Hammond HA, It CTC. DNA Typing and Genetic Mapping with Trimeric and Tetrameric Tandem Repeats. *American Journal of Human Genetics* 1991;(49):746–56.
9. Gill P, Ivanov PL, Kimpton C, Piercy R, Benson N, Tully G, et al. Identification of the remains of the Romanov family by DNA analysis. *Nature Genetics* 1994;6(2):130–5. Available from: <http://www.nature.com/articles/ng0294-130>
10. Sanger F, Nicklen S, Coulson AR. DNA sequencing with chain-terminating inhibitors. 1977; Available from: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC431765/pdf/pnas00043-0271.pdf>
11. Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, et al. The Sequence of the Human Genome. *Science* 2001;291(5507):1304–51. Available from: <https://science.sciencemag.org/content/291/5507/1304>
12. Berglund EC, Kiialainen A, Syvänen A-C. Next-generation sequencing

technologies and applications for human genetic history and forensics. *Investigative Genetics* 2011;2(1):23. Available from: <http://investigativegenetics.biomedcentral.com/articles/10.1186/2041-2223-2-23>

13. Anderson S, Bankier AT, Barrell BG, de Bruijn MHL, Coulson AR, Drouin J, et al. Sequence and organization of the human mitochondrial genome. *Nature* 1981;290(5806):457–65. Available from: <http://www.nature.com/articles/290457a0>
14. Qin J, Li R, Raes J, Arumugam M, Burgdorf KS, Manichanh C, et al. A human gut microbial gene catalogue established by metagenomic sequencing. *Nature* 2010;464(7285):59–65. Available from: <http://www.nature.com/articles/nature08821>
15. Meiring TL, Salimo AT, Coetzee B, Maree HJ, Moodley J, Hitzeroth II, et al. Next-generation sequencing of cervical DNA detects human papillomavirus types not detected by commercial kits. *Virology Journal* 2012;9:164. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/22897914>
16. Kothari N, Schell MJ, Teer JK, Yeatman T, Shibata D, Kim R. Comparison of KRAS mutation analysis of colorectal cancer samples by standard testing and next-generation sequencing. *Journal of Clinical Pathology* 2014;67(9):764–7. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/25004944>
17. Bentley DR, Balasubramanian S, Swerdlow HP, Smith GP, Milton J, Brown CG, et al. Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* 2008;456(7218):53–9. Available from: <http://www.nature.com/articles/nature07517>
18. Lynch PC, Cotton RW. Determination of the possible number of genotypes which can contribute to DNA mixtures: Non-computer assisted deconvolution should not be attempted for greater than two person mixtures. *Forensic Science International: Genetics* 2018;37:235–40. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/30261423>
19. SWGDAM. SWGDAM Interpretation Guidelines for Autosomal STR Typing by Forensic DNA Testing Laboratories. 2017; Available from: [https://docs.wixstatic.com/ugd/4344b0\\_50e2749756a242528e6285a5bb478f4c.pdf](https://docs.wixstatic.com/ugd/4344b0_50e2749756a242528e6285a5bb478f4c.pdf)
20. David L. Duewer, Margaret C. Kline, Janette W. Redman A, Butler JM. NIST Mixed Stain Study 3: Signal Intensity Balance in Commercial Short Tandem Repeat Multiplexes. *Analytical Chemistry* 2004;76(23):6928–34. Available from: <https://pubs.acs.org/doi/10.1021/ac049178k>



<http://doi.wiley.com/10.1002/elps.201600511>

29. Grgicak CM, Urban ZM, Cotton RW. Investigation of reproducibility and error associated with qPCR methods using Quantifiler® Duo DNA quantification kit. *Journal of Forensic Sciences* 2010;55(5):1331–9.
30. QIAGEN. QIAamp® DNA Investigator Handbook. 2012;(June):1–60.
31. Sharma V, Chow HY, Siegel D, Wurmbach E. Qualitative and quantitative assessment of Illumina's forensic STR and SNP kits on MiSeq FGx™. *PLoS ONE* 2017;12(11):1–21.
32. 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 Science International: Genetics* 2017;31:135–48. Available from:  
<https://www.sciencedirect.com/science/article/pii/S1872497317301898>
33. Churchill JD, Schmedes SE, King JL, Budowle B. Evaluation of the Illumina® Beta Version ForenSeq™ DNA Signature Prep Kit for use in genetic profiling. *Forensic Science International: Genetics* 2016;20:20–9. Available from:  
<https://www.sciencedirect.com/science/article/pii/S1872497315300715?via%3Dihub>

**CURRICULUM VITAE**

