

Warshauer, David H. Development of a Comprehensive Massively Parallel Sequencing Panel of Single Nucleotide Polymorphism and Short Tandem Repeat Markers for Human Identification. Doctor of Philosophy (Biomedical Sciences), July 2015, 233 pp., 96 tables, 32 figures, 115 references.

Massively parallel sequencing (MPS) technologies allow for the detection of an unparalleled amount of genetic information with unprecedented speed and relative ease. These qualities make the technology desirable for generating DNA profiles that may be uploaded into forensic offender, arrestee, and family reference database files. This doctoral dissertation research was conducted under the hypothesis that MPS, with its exquisitely high throughput, can provide a system whereby reference samples can be typed for a large battery of markers, providing more discrimination power for forensic DNA typing and offering increased opportunities to develop investigative leads. The design and implementation of large marker panels for the typing of reference samples will reduce debates on the best core markers for forensic utility, generate innovation because focus will not be solely on a core set of autosomal STRs, promote the development of better systems that can analyze more challenging samples, and enable sharing of data across laboratories worldwide.

The primary goal of this project was to develop the capability of typing reference samples for a large battery of markers: 84 autosomal, Y-chromosome, and X-chromosome short tandem repeats (STRs), Amelogenin, and 275 human identity single nucleotide polymorphisms (SNPs), in a single multiplex analysis. To that end, a bioinformatic software package, STRait Razor, was developed to detect STR alleles in raw MPS data. A proof-of-concept study was performed to evaluate the efficacy of using MPS to type forensically relevant markers, using a PCR multiplex-based SNP assay. The proposed comprehensive capture-based MPS panel then was designed and extensively tested. Finally, the benefits of the additional genetic data afforded by MPS, as opposed to traditional methods, were illustrated through the characterization of intra-repeat nucleotide variation within Y-chromosome STR alleles.

The results of this dissertation research indicate that MPS is capable of providing robust genetic data from a wide variety of forensically-relevant STR and SNP loci in a single analysis. To date, the comprehensive MPS panel developed during the course of these studies is the most potentially informative assay for reference sample testing for human identification.

KEYWORDS: Comprehensive panel · Massively parallel sequencing · STR · SNP · Bioinformatics · Nextera™ Rapid Capture · MiSeq™ · STRait Razor · Sequence polymorphism

**DEVELOPMENT OF A COMPREHENSIVE
MASSIVELY PARALLEL SEQUENCING PANEL
OF SINGLE NUCLEOTIDE POLYMORPHISM AND
SHORT TANDEM REPEAT MARKERS
FOR HUMAN IDENTIFICATION**

DISSERTATION

Presented to the Graduate Council of the
Graduate School of Biomedical Sciences
University of North Texas
Health Science Center at Fort Worth
In Partial Fulfillment of the Requirements

For the Degree of

DOCTOR OF PHILOSOPHY

By

David H. Warshauer, B.S., M.S.
Fort Worth, Texas
July, 2015

ACKNOWLEDGMENTS

As I look back on my academic career, recalling the challenges and successes that I have experienced, I am overwhelmed by a deep sense of appreciation for the many amazing personalities I have encountered. Throughout the course of my doctoral research, I have had the opportunity to work with a variety of extremely knowledgeable, creative, and supportive individuals, without whom I would never have reached this point.

I would like to begin by thanking Dr. Bruce Budowle, my Major Professor, mentor, and friend. Dr. Budowle possesses the unique gift of being able to recognize and accept the personalities and individual styles of his students so that he can interact with them on a personal level. He not only shared his extensive scientific knowledge with me, but he also fostered a sense of curiosity, independence, and creativity in me that has helped me become a better researcher. He diligently ensured that I stayed focused on my goals, and I could always count on him for reassurance and a light-hearted joke during times of stress. He is the type of scientist that I would one day like to become, and I will be eternally grateful for all he has done for me.

I would also like to thank the other members of my PhD advisory committee: Dr. Ranajit Chakraborty, Dr. Bobby LaRue, Dr. Harlan Jones, and Dr. Ren-Qi Huang. Their guidance and support have been invaluable to me throughout my doctoral research.

None of my research would have been possible without the assistance and involvement of my colleagues and co-workers. I would like to express my appreciation to Jonathan King,

Xiangpei Zeng, Carey Davis, Jennifer Churchill, Nicole Novroski, Sarah Schmedes, and Pamela Marshall for their selfless help, which was instrumental in the accomplishment of my research goals. I am also grateful to Cydne Holt, Yonmee Han, Paulina Walichiewicz, Tom Richardson, Kathryn Stephens, and Anne Jager of Illumina, Inc., as well as David Lin, Kumar Hari, and Ravi Jain of cBio, Inc., for their technical guidance.

I would like to express my most heartfelt gratitude to my incredible family, who have been my rock from the very beginning. I would not be the man I am today without the love and support of my parents, Victor and Mary, and my late grandmother, Jo. They nurtured my inquisitive nature from a young age and taught me that, with enough determination and effort, I could accomplish absolutely anything I desired. My amazing wife, Cortney, has been equally important in my life. She has given me the strength to persevere through even the most challenging situations, and I am grateful for her encouragement and belief in me.

Finally, and most importantly, I would like to thank God. All of my talents and achievements are from Him alone, and I have always been amazed at the way all the events of my life have fallen so perfectly into place with Him as my guide. My scientific career thus far has truly served to strengthen my faith. I am grateful to have the opportunity and ability to observe the biological intricacy and efficiency of His design.

This project was supported by Award No. 2012-DN-BX-K033, awarded by the National Institute of Justice, Office of Justice Programs, U.S. Department of Justice. The opinions,

findings, and conclusions or recommendations expressed in this publication are those of the author(s) and do not necessarily reflect those of the Department of Justice.

TABLE OF CONTENTS

ACKNOWLEDGEMENTS	ii
LIST OF FIGURES	vi
LIST OF TABLES	viii
INTRODUCTION	1
<i>Development of a Novel Bioinformatic Tool for the Detection of STR Alleles in Raw Massively Parallel Sequencing Data</i>	
Chapter 1: STRait Razor: A Length-Based Forensic STR Allele-Calling Tool for Use with Second Generation Sequencing Data	18
Chapter 2: STRait Razor v2.0: the Improved STR Allele Identification Tool – Razor	43
<i>Initial Evaluation of Massively Parallel Sequencing as a Tool for Comprehensive Human Identification Marker Testing</i>	
Chapter 3: Massively Parallel Sequencing of Forensically Relevant Single Nucleotide Polymorphisms Using TruSeq™ Forensic Amplicon	60
<i>Design and Testing of an Extensive Capture-Based Assay for the Detection of Human Identification Markers</i>	
Chapter 4: Development of a Comprehensive Massively Parallel Sequencing Panel of Single Nucleotide Polymorphism and Short Tandem Repeat Markers for Human Identification	78
<i>Detection of Intra-Repeat Nucleotide Variation in Short Tandem Repeat Alleles Through the Use of the Comprehensive Massively Parallel Sequencing Panel</i>	
Chapter 5: Novel Y-Chromosome Short Tandem Repeat Variants Detected Through the Use of Massively Parallel Sequencing	118
SUMMARY	137
CONCLUSIONS	142
SUPPLEMENTAL MATERIAL	146
REFERENCES	224

LIST OF FIGURES

INTRODUCTION

- Figure 1. Sanger sequencing
- Figure 2. dNTP and ddNTP structure
- Figure 3. Bridge amplification
- Figure 4. MiSeq™ sequencing-by-synthesis
- Figure 5. Index barcoding
- Figure 6. Capture-based library preparation

CHAPTER 1: *STRait Razor: A Length-Based Forensic STR Allele-Calling Tool for Use with Second Generation Sequencing Data*

- Fig. 1. STRait Razor algorithm
- Fig. 2. Read length-related issues

CHAPTER 2: *STRait Razor v2.0: the Improved STR Allele Identification Tool – Razor*

- Fig. 1. Sequence data sorting
- Fig. 2. Histogram generated for locus D6S474 in dataset 3 (Read 1)
- Fig. 3. Example of intra-repeat variation detected at locus D21S11 in dataset 4

CHAPTER 3: *Massively Parallel Sequencing of Forensically Relevant Single Nucleotide Polymorphisms Using TruSeq™ Forensic Amplicon*

- Fig. 1. Heterozygous allele balance for sample 2
- Fig. 2. Average sequence coverage for apSNP loci

CHAPTER 4: *Development of a Comprehensive Massively Parallel Sequencing Panel of Single Nucleotide Polymorphism and Short Tandem Repeat Markers for Human Identification*

- Figure 1. Relative locus performance – autosomal STRs and Amelogenin
- Figure 2. Relative locus performance – X-chromosome STRs (females only)
- Figure 3. Relative locus performance – X-chromosome STRs (males only)

LIST OF FIGURES (CONTINUED)

CHAPTER 4: *Development of a Comprehensive Massively Parallel Sequencing Panel of Single Nucleotide Polymorphism and Short Tandem Repeat Markers for Human Identification*

Figure 4. Relative locus performance – Y-chromosome STRs

Figure 5. Heterozygosity balance – autosomal STRs and Amelogenin

Figure 6. Heterozygosity balance – X-chromosome STRs (females only)

Figure 7. Relative locus performance – autosomal SNPs

Figure 8. Relative locus performance – Y-chromosome SNPs

Figure 9. Heterozygosity balance – Autosomal SNPs

SUPPLEMENTAL MATERIAL

Chapter 4, Supplemental Figure 1. Relative locus performance – autosomal SNPs (part 1)

Chapter 4, Supplemental Figure 2. Relative locus performance – autosomal SNPs (part 2)

Chapter 4, Supplemental Figure 3. Relative locus performance – autosomal SNPs (part 3)

Chapter 4, Supplemental Figure 4. Relative locus performance – autosomal SNPs (part 4)

Chapter 4, Supplemental Figure 5. Relative locus performance – autosomal SNPs (part 5)

Chapter 4, Supplemental Figure 6. Heterozygosity balance – Autosomal SNPs (part 1)

Chapter 4, Supplemental Figure 7. Heterozygosity balance – Autosomal SNPs (part 2)

Chapter 4, Supplemental Figure 8. Heterozygosity balance – Autosomal SNPs (part 3)

Chapter 4, Supplemental Figure 9. Heterozygosity balance – Autosomal SNPs (part 4)

Chapter 4, Supplemental Figure 10. Heterozygosity balance – Autosomal SNPs (part 5)

LIST OF TABLES

INTRODUCTION

Table 1. Bioinformatic software used in dissertation research

CHAPTER 1: *STRait Razor: A Length-Based Forensic STR Allele-Calling Tool for Use with Second Generation Sequencing Data*

Table 1. Autosomal STR loci detected by STRait Razor

Table 2. Y-chromosome STR loci detected by STRait Razor

Table 3. Comparison of CE allele calls and STRait Razor results – Autosomal STRs

Table 4. Comparison of CE allele calls and STRait Razor results – Y-Chromosome STRs

CHAPTER 2: *STRait Razor v2.0: the Improved STR Allele Identification Tool – Razor*

Table 1. Datasets used for concordance testing

Table 2. Additional alleles and loci detected in the original 7 datasets

Table 3. Allele and locus detection comparison for the additional 12 datasets

CHAPTER 3: *Massively Parallel Sequencing of Forensically Relevant Single Nucleotide Polymorphisms Using TruSeq™ Forensic Amplicon*

Table 1. SNP discordance (in-house vs. WGS)

CHAPTER 4: *Development of a Comprehensive Massively Parallel Sequencing Panel of Single Nucleotide Polymorphism and Short Tandem Repeat Markers for Human Identification*

Table 1. Panel markers – STRs and Amelogenin

Table 2. Panel markers – SNPs

Table 3. Discordant STR loci

Table 4. Significant LD results – Y-chromosome STRs

Table 5. Significant LD results – Autosomal SNPs

Table 6. Significant LD results – Y-chromosome SNPs

Table 7. Significant LD results – Autosomal SNP and STR pairs

LIST OF TABLES (CONTINUED)

CHAPTER 4: *Development of a Comprehensive Massively Parallel Sequencing Panel of Single Nucleotide Polymorphism and Short Tandem Repeat Markers for Human Identification*

Table 8. Y-chromosome haplotypes and haplogroup predictions for each sample in each population group

CHAPTER 5: *Novel Y-Chromosome Short Tandem Repeat Variants Detected Through the Use of Massively Parallel Sequencing*

Table 1. Nominal allele variant sequences

Table 2. Novel allele variant sequences

SUPPLEMENTAL MATERIAL

Chapter 3, Supplemental Table 1. iSNPs and apSNPs included in the TruSeq™ Forensic Amplicon multiplexes

Chapter 3, Supplemental Table 2. SNP discordance (in-house vs. Illumina)

Chapter 3, Supplemental Table 3. Heterozygous SNP loci with allelic balance values below 50%

Chapter 4, Supplemental Table 1. Autosomal STR allele counts – African Americans

Chapter 4, Supplemental Table 2. Autosomal STR allele counts – Asians

Chapter 4, Supplemental Table 3. Autosomal STR allele counts – Caucasians

Chapter 4, Supplemental Table 4. Autosomal STR allele counts – Hispanics

Chapter 4, Supplemental Table 5. X-chromosome STR allele counts – African American Females

Chapter 4, Supplemental Table 6. X-chromosome STR allele counts – Asian Females

Chapter 4, Supplemental Table 7. X-chromosome STR allele counts – Caucasian Females

Chapter 4, Supplemental Table 8. X-chromosome STR allele counts – Hispanic Females

Chapter 4, Supplemental Table 9. X-chromosome STR allele counts – African American males

LIST OF TABLES (CONTINUED)

SUPPLEMENTAL MATERIAL

Chapter 4, Supplemental Table 10. X-chromosome STR allele counts – Asian males

Chapter 4, Supplemental Table 11. X-chromosome STR allele counts – Caucasian males

Chapter 4, Supplemental Table 12. X-chromosome STR allele counts – Hispanic males

Chapter 4, Supplemental Table 13. Y-chromosome STR allele counts – African Americans

Chapter 4, Supplemental Table 14. Y-chromosome STR allele counts – Asians

Chapter 4, Supplemental Table 15. Y-chromosome STR allele counts – Caucasians

Chapter 4, Supplemental Table 16. Y-chromosome STR allele counts – Hispanics

Chapter 4, Supplemental Table 17. Autosomal SNP allele counts – African Americans

Chapter 4, Supplemental Table 18. Autosomal SNP allele counts – Asians

Chapter 4, Supplemental Table 19. Autosomal SNP allele counts – Caucasians

Chapter 4, Supplemental Table 20. Autosomal SNP allele counts – Hispanics

Chapter 4, Supplemental Table 21. Y-chromosome SNP allele counts – African Americans

Chapter 4, Supplemental Table 22. Y-chromosome SNP allele counts – Asians

Chapter 4, Supplemental Table 23. Y-chromosome SNP allele counts – Caucasians

Chapter 4, Supplemental Table 24. Y-chromosome SNP allele counts – Hispanics

Chapter 4, Supplemental Table 25. Autosomal STR allele frequencies – African Americans

Chapter 4, Supplemental Table 26. Autosomal STR allele frequencies – Asians

Chapter 4, Supplemental Table 27. Autosomal STR allele frequencies – Caucasians

Chapter 4, Supplemental Table 28. Autosomal STR allele frequencies – Hispanics

LIST OF TABLES (CONTINUED)

SUPPLEMENTAL MATERIAL

Chapter 4, Supplemental Table 29. X-chromosome STR allele frequencies – African American females

Chapter 4, Supplemental Table 30. X-chromosome STR allele frequencies – Asian females

Chapter 4, Supplemental Table 31. X-chromosome STR allele frequencies – Caucasian females

Chapter 4, Supplemental Table 32. X-chromosome STR allele frequencies – Hispanic females

Chapter 4, Supplemental Table 33. X-chromosome STR allele frequencies – African American males

Chapter 4, Supplemental Table 34. X-chromosome STR allele frequencies – Asian males

Chapter 4, Supplemental Table 35. X-chromosome STR allele frequencies – Caucasian males

Chapter 4, Supplemental Table 36. X-chromosome STR allele frequencies – Hispanic American males

Chapter 4, Supplemental Table 37. Y-chromosome STR allele frequencies – African Americans

Chapter 4, Supplemental Table 38. Y-chromosome STR allele frequencies – Asians

Chapter 4, Supplemental Table 39. Y-chromosome STR allele frequencies – Caucasians

Chapter 4, Supplemental Table 40. Y-chromosome STR allele frequencies – Hispanics

Chapter 4, Supplemental Table 41. Autosomal SNP allele frequencies – African Americans

Chapter 4, Supplemental Table 42. Autosomal SNP allele frequencies – Asians

Chapter 4, Supplemental Table 43. Autosomal SNP allele frequencies – Caucasians

Chapter 4, Supplemental Table 44. Autosomal SNP allele frequencies – Hispanics

LIST OF TABLES (CONTINUED)

SUPPLEMENTAL MATERIAL

Chapter 4, Supplemental Table 45. Y-chromosome SNP allele frequencies – African Americans

Chapter 4, Supplemental Table 46. Y-chromosome SNP allele frequencies – Asians

Chapter 4, Supplemental Table 47. Y-chromosome SNP allele frequencies – Caucasians

Chapter 4, Supplemental Table 48. Y-chromosome SNP allele frequencies – Hispanics

Chapter 4, Supplemental Table 49. FST – Autosomal STRs

Chapter 4, Supplemental Table 50. FST – X-chromosome STRs (females)

Chapter 4, Supplemental Table 51. FST – X-chromosome STRs (males)

Chapter 4, Supplemental Table 52. FST – Y-chromosome STRs

Chapter 4, Supplemental Table 53. FST – Autosomal SNPs

Chapter 4, Supplemental Table 54. FST – Y-chromosome SNPs

Chapter 4, Supplemental Table 55. Linkage disequilibrium – syntenic autosomal STR loci pairs

Chapter 4, Supplemental Table 56. Linkage disequilibrium – X-chromosome STR loci pairs (part 1)

Chapter 4, Supplemental Table 57. Linkage disequilibrium – X-chromosome STR loci pairs (part 2)

Chapter 4, Supplemental Table 58. Linkage disequilibrium – X-chromosome STR loci pairs (part 3)

Chapter 4, Supplemental Table 59. Linkage disequilibrium – Y-chromosome STR loci pairs (part 1)

Chapter 4, Supplemental Table 60. Linkage disequilibrium – Y-chromosome STR loci pairs (part 2)

Chapter 4, Supplemental Table 61. Linkage disequilibrium – syntenic autosomal SNP loci pairs (part 1)

LIST OF TABLES (CONTINUED)

SUPPLEMENTAL MATERIAL

Chapter 4, Supplemental Table 62. Linkage disequilibrium – syntenic autosomal SNP loci pairs (part 2)

Chapter 4, Supplemental Table 63. Linkage disequilibrium – syntenic autosomal SNP loci pairs (part 3)

Chapter 4, Supplemental Table 64. Linkage disequilibrium – syntenic autosomal SNP loci pairs (part 4)

Chapter 4, Supplemental Table 65. Linkage disequilibrium – syntenic autosomal SNP loci pairs (part 5)

Chapter 4, Supplemental Table 66. Linkage disequilibrium – syntenic autosomal SNP loci pairs (part 6)

Chapter 4, Supplemental Table 67. Linkage disequilibrium – syntenic autosomal SNP loci pairs (part 7)

Chapter 4, Supplemental Table 68. Linkage disequilibrium – syntenic autosomal SNP loci pairs (part 8)

Chapter 4, Supplemental Table 69. Linkage disequilibrium – Y-chromosome SNP loci pairs (part 1)

Chapter 4, Supplemental Table 70. Linkage disequilibrium – Y-chromosome SNP loci pairs (part 2)

Chapter 4, Supplemental Table 71. Linkage disequilibrium – Y-chromosome SNP loci pairs (part 3)

Chapter 4, Supplemental Table 72. Linkage disequilibrium – Y-chromosome SNP loci pairs (part 4)

Chapter 4, Supplemental Table 73. Linkage disequilibrium – syntenic autosomal STR and SNP loci pairs (part 1)

Chapter 4, Supplemental Table 74. Linkage disequilibrium – syntenic autosomal STR and SNP loci pairs (part 2)

INTRODUCTION

DEVELOPMENT OF A COMPREHENSIVE MASSIVELY PARALLEL SEQUENCING PANEL OF SINGLE NUCLEOTIDE POLYMORPHISM AND SHORT TANDEM REPEAT MARKERS FOR HUMAN IDENTIFICATION

Over the past three decades, a number of robust and reliable DNA typing technologies for human identity testing have been implemented (1-3). These methods enable analyses of minute quantities of DNA and provide a resolving power such that in many cases the number of potential contributors of an evidence sample can be reduced to only a few individuals, if not a single source. The success of DNA typing has led to further applications, the most prominent of which has been its use for developing investigative leads. The potential of DNA typing for developing investigative leads and for solving future crimes came to fruition with the development of DNA databases. Many countries have established DNA databanks that contain profiles from convicted offenders, arrestees and forensic samples from unsolved cases (4,5). These databases are designed to house DNA profiles that can be used to associate individuals with forensic samples or to identify missing persons. The U.S. databank, CODIS (the Combined DNA Index System), contains more than 13,780,000 reference DNA profiles (6) (as of May 2015) and is routinely relied upon for helping to develop meaningful investigative leads. Due to the success in providing such leads, these DNA databases continue to increase in size, and the information contained within them may be used for purposes other than the direct matching of profiles, such as familial searching, the searching of DNA databases for the purpose of identifying close biological relatives of a contributor to an unknown forensic profile (7-9).

The reliance on offender, arrestee, and forensic sample DNA databases has driven and will continue to drive innovation and standards. Likewise, the demands of generating, entering, and maintaining DNA profiles in a national DNA database have fostered developments in automation and the creation of robust molecular assays. The number of reference samples from convicted felons, arrestees, detainees, and missing persons continues to increase, and there is no indication

of the demand subsiding. To meet the needs of forensic DNA typing and its infrastructure, it is imperative that forensic scientists remain vigilant and embrace new technologies that will benefit the process, as well as society, by allowing for analyses of ever-increasing numbers of reference samples, as well as more challenging samples. Such advancements will provide means to further facilitate the exoneration of the innocent, the identification of missing persons, and an enhancement of the ability to solve crimes.

One particular challenge to resolve is the selection of markers that should be used routinely by forensic laboratories. To be able to share and compare DNA results, a core set of short tandem repeat (STR), or microsatellite, loci was selected seventeen years ago (4,5). Recently, Hares (10,11), representing the position of the FBI, recommended that the core 13 STR loci for CODIS should be changed and augmented. The FBI advocated 20 STR loci to serve as the new CODIS panel of markers. Ge *et al.* (9) suggested that there were additional factors and applications for selecting a core set of markers beyond those relied upon by Hares. This alternate viewpoint was that the loci selected should be driven by the demands of casework (i.e., loci should be selected based on performance with degraded and inhibited samples, or the technology should be versatile enough to enable a variety of search strategies). Thus, there are differences of opinions on how to proceed with core marker selection. Additionally, the implementation of core markers, while useful for formalizing a common set for data exchange, inadvertently can limit progress or stifle innovation with regard to alternate markers that may serve other, more specialized forensic community needs, such as familial searching. However, these concerns can be rendered moot with the advent of massively parallel sequencing (MPS) technology.

The MPS technologies use non-Sanger-based sequencing strategies to provide DNA sequencing data with unprecedented capacity and speed at a reduced cost (12). For the past few decades, sequencing has primarily been performed by the Sanger method (13) (Figure 1). In Sanger sequencing, template DNA is replicated using a modified polymerase chain reaction (PCR). A relatively high concentration of fluorescently-labeled dideoxynucleotide triphosphates (ddNTPs) (Figure 2) are included with the deoxynucleotide triphosphates (dNTPs) during elongation. While dNTPs allow elongation of the growing strand through the formation of phosphodiester bonds, the incorporation of a ddNTP, which lacks the 3' hydroxyl group, does not allow this bond formation, causing termination of replication. This reaction results in a variety of fragments of different length, each increasing by a single base in the sequence of the template strand. These fragments then are subjected to size separation via capillary electrophoresis (CE), and the resulting electropherogram reveals a pattern that can be interpreted to deduce the nucleotide sequence.

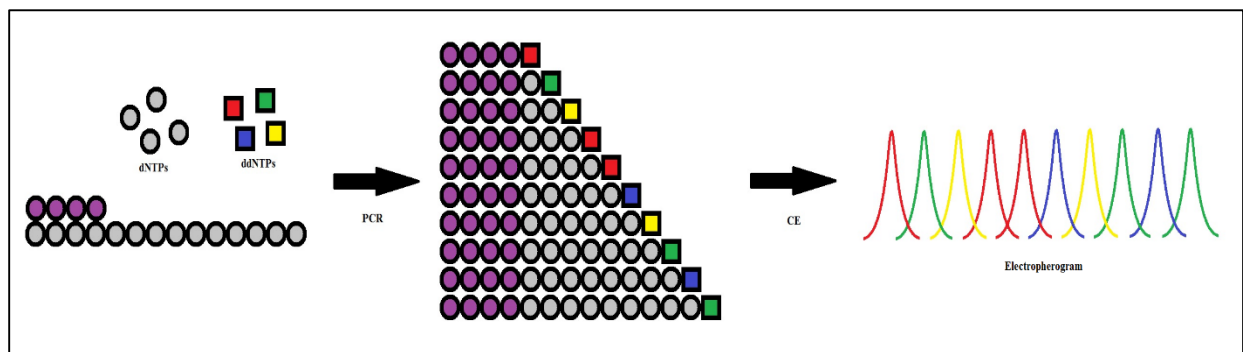


Figure 1. Sanger sequencing. A typical dye terminator Sanger sequencing process is illustrated. The template DNA strand is shown in grey, while the bound sequencing primer is depicted in purple.

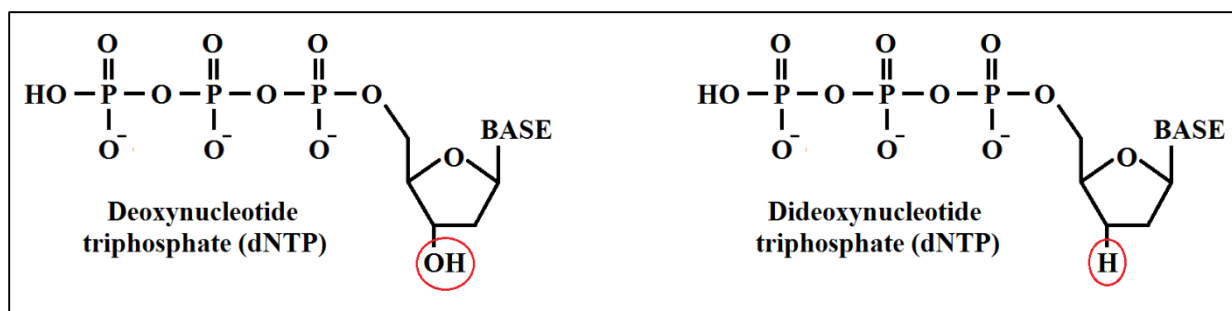


Figure 2. dNTP and ddNTP structure. ddNTPs lack the 3' hydroxyl group found in dNTPs, so phosphodiester bonds cannot be formed.

While Sanger sequencing is robust and used particularly for mitochondrial DNA (mtDNA) sequencing (14,15) and some single nucleotide polymorphism (SNP)-based assays (16,17) in a forensic context, it is labor intensive, has a relatively low throughput, and is costly on a “per nucleotide” basis. In contrast, MPS technologies sequence DNA in a massively parallel fashion with high coverage, which can result in low error, as well as a high throughput of specified targets. In fact, whole human genomes have been sequenced, with costs dropping dramatically, toward the goal of \$1000 or less. One example of MPS technology is the Illumina® MiSeq™ platform (18). During library preparation for the MiSeq™, DNA fragments are bound to adapter sequences which contain a PCR primer binding site. On the MiSeq™, these adapter-bound DNA fragments hybridize with complementary adapter sequences that are covalently attached to the floor of a small microfluidics chamber known as a flow cell. The fragments then are copied via PCR, which leaves the resulting strands bound to the flow cell, while the original strands are washed away. The bound strands then are subjected to a process called “bridge amplification” (Figure 3), whereby they are bent over and hybridize to nearby adapters before being replicated via a modified PCR. This process is repeated a number of times, and the end result is a multitude of “clusters,” dense groupings of clones of DNA fragments which are ready to be sequenced.

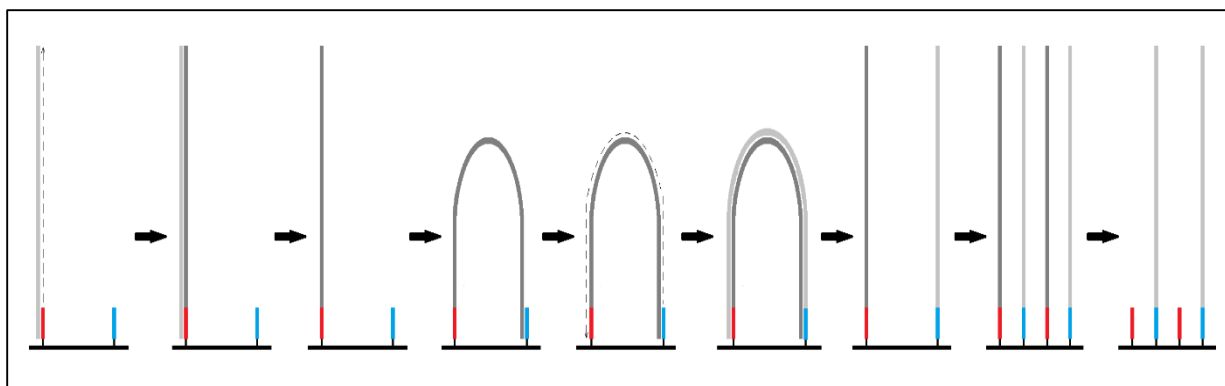


Figure 3. Bridge amplification. A typical bridge amplification process is illustrated. Forward DNA strands are shown in light grey, while reverse complement strands are depicted in dark grey. Complementary flow cell-bound adapter sequences are shown in red and blue. The process begins with an adapter-bound DNA strand that hybridizes with a complementary adapter on the flow cell. A copy is generated, and the original strand is washed away. The resulting bound strand is bent over, and its adapter sequence hybridizes with another complementary adapter on the surface of the flow cell. A copy is generated via PCR, using the bent-over DNA strand as a template. The strands then are separated, leaving two complementary adapter-bound DNA sequences attached to the flow cell. The bridge amplification process is repeated, creating clusters of the sample DNA sequence (both forward and reverse complement strands) on the flow cell's surface. The reverse complement strands are cleaved and washed away, leaving only the forward strands to be sequenced.

The MiSeq™ uses a sequencing-by-synthesis process (Figure 4), which begins when fluorescently-labeled dNTPs, with attached chemical terminators, are passed across the flow cell. The single “correct” nucleotide, which is complementary to the cloned fragment or template strands tethered to the flow cell, is incorporated into each replicating strand, and the terminator halts the elongation process. The laser excitation-induced fluorescent emission of the incorporated nucleotide tag is captured by a CCD camera for all clusters, simultaneously. Then, the terminator is cleaved, and the process begins again with the next cycle of nucleotides. The use of attached chemical terminators allows for the incorporation of only one deoxynucleotide at a time, even in homopolymeric stretches (i.e., two or more of the same nucleotide in a row).

arrestee, and family reference database files. Due to the exquisitely high throughput, a large battery of genetic markers can be analyzed simultaneously, far exceeding the current capacity of 15-27 STRs in a fluorescent multiplex/CE system (29). It is entirely possible that all forensically-relevant identified autosomal STRs, including the 24 STR loci described by Hares, a set of Y-chromosome and X-chromosome STRs, and human identity SNPs (comprising more than two hundred markers, with the potential for much more) can be typed simultaneously. Indeed, sequencing kits have been developed that contain reagents for typing 23 STR loci (30) and beyond, to include a set of Y-chromosome and X-chromosome STRs, and human identity SNPs (comprising more than two hundred markers) (31). Moreover, given the high throughput capacity afforded by MPS, many different samples can be individually labeled with short unique sequences of nucleotides, known as index barcodes, which act as identifiers for each sample so that they may be multiplexed and sequenced at the same time (Figure 5). After sequencing, the combined sequence data are parsed by the sequencing instrument's software, based on the indices, so that the individual samples can be analyzed separately. In theory, hundreds to thousands of barcodes could be synthesized, but currently, approximately 12 different reference samples could be multiplexed for simultaneous sequencing using a panel of this magnitude with this particular chemistry.

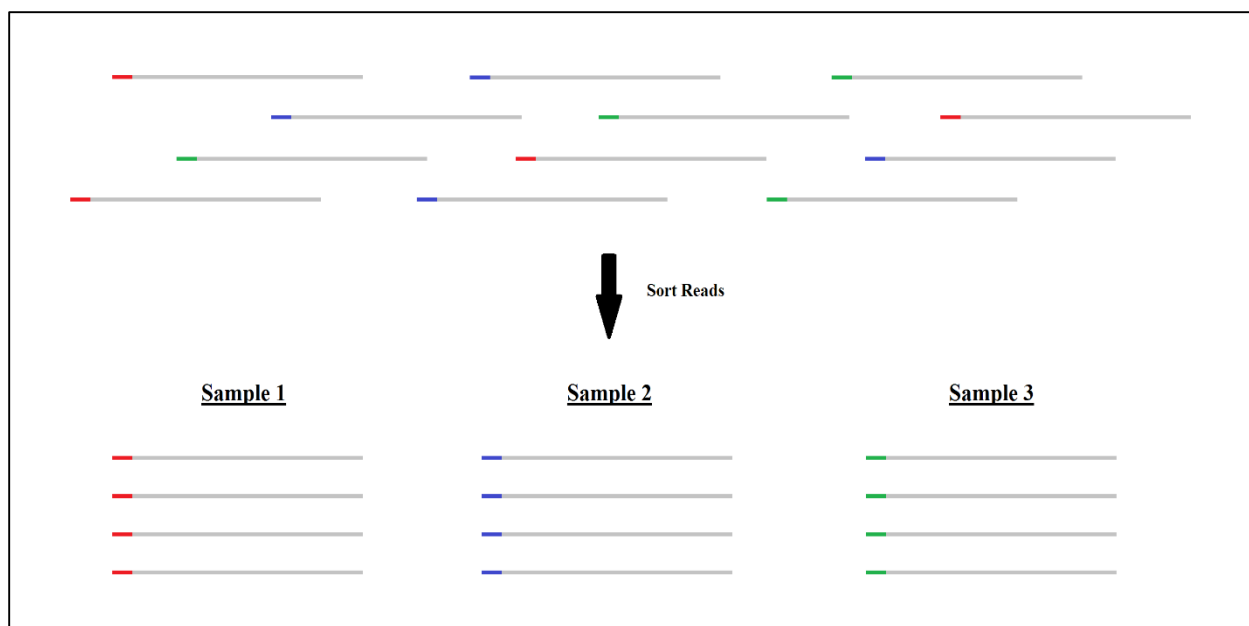


Figure 5. Index barcoding. The function of index sequences is illustrated. MPS sequence reads are shown in grey, while the index sequences, which are unique to each sample, are shown in red, green, and blue. After sequencing, the sequence reads corresponding to each sample are sorted based on these indices.

Given the increase of typing data afforded with MPS, the majority (if not all) of the profiles from evidence samples can be compared with reference profiles, regardless of the number and types of markers used in the evidentiary analyses. The profiles generated by MPS technology will be compatible with existing reference data, and the more comprehensive set of markers made possible through the use of MPS will foster investigations.

The raw sequence data yielded by MPS instruments simply consists of every DNA sequence read by the instrument for each sample. To determine genotypes for STR and SNP loci contained within these reads, the raw MPS data must be analyzed. Given the large datasets commonly sequenced by MPS and the massive amount of sequence information that is yielded, manual review of raw MPS data is not practical. Therefore, bioinformatic software is commonly used for analysis. The term “bioinformatic software” refers to a wide variety of computer

programs that perform different functions needed to organize and interpret sequence data. In the case of SNP analysis, the bioinformatics software that are used typically consist of an alignment program and a variant-calling program. Alignment programs match raw sequence data to a known reference genome sequence so that the most probable location in the genome for each sequence read can be determined. Variant-calling programs then use this information to compare the detected nucleotides in the sequence reads to the reference nucleotides at these locations, noting any differences (which are designated as SNPs). STR analysis requires the use of sophisticated software that analyzes the raw sequence data for the presence of STR repeat motifs, classifying the alleles that are present based on these motifs. Other bioinformatic software packages, such as programs that provide a graphical user interface for review of sequence data, or those that convert sequence files from one format to another, are also commonly used. Table 1 lists the various bioinformatic programs that were used in this dissertation research.

Name	Analysis Purpose	Type	Function
Burrows-Wheeler Aligner (BWA)	SNPs	Alignment	Aligns raw sequence data to a reference genome
Genome Analysis Toolkit (GATK)	SNPs	Variant-Calling	Calls SNP genotypes based on comparison with a reference genome
MiSeq™ Reporter	SNPs	Alignment and Variant-Calling	Aligns raw sequence data to a reference genome and calls SNP genotypes based on comparison with the reference genome, on the instrument
STRAIT Razor v2.0	STRs	Allele-Calling	Calls STR genotypes based on the lengths of the alleles detected in the raw sequence data
Integrative Genomics Viewer (IGV)	SNPs and STRs	Visualization	Provides a graphical user interface for visual investigation of aligned sequence data

Table 1. Bioinformatic software used in dissertation research. Various bioinformatic software packages used for this project are listed, along with their respective purposes, types, and functions.

With the increased number of forensic markers that can be incorporated in a comprehensive MPS panel, including lineage markers such as Y-chromosome STRs (which are shared by all paternally-related individuals in a given lineage), indirect searches can be performed. Familial

searching, the searching of DNA databases for the purpose of identifying close biological relatives of a contributor to an unknown forensic profile, would be highly successful and provide an increased number of investigative leads. The sheer abundance of markers will provide more robust associations and substantially reduce the number of false positive associations. These improvements will have a substantial impact on the field of missing persons identification testing. It is likely that as more kinship associations result in solved crimes, there will be motivation to further exploit familial searching via MPS profiling.

It should be noted that while CE technology reveals only the length of STR alleles, MPS also provides information about the individual arrangement of nucleotides within their repeat regions. The detection of intra-repeat SNP and novel repeat motif variants made possible by MPS will allow for a greater level of distinction than that of traditional CE methods. For instance, two individuals with the same nominal allele(s) (based on length) at a certain STR locus potentially may be distinguished by MPS if the nucleotide sequence of the allele differs between them. This level of resolution may prove invaluable in the deconvolution of genetic mixtures and also could add additional information for evolutionary studies. This untapped genetic variation has yet to be described fully.

The doctoral research described herein was based on the hypothesis that MPS, with its economies of scale, can provide a system such that reference samples can be typed for a large battery of markers, and will reveal additional genetic information that can improve the discrimination power of forensic DNA typing. The primary goal of these studies was to develop the capability of typing reference samples for a large battery of autosomal, Y-chromosome, and

X-chromosome STRs and human identity autosomal and Y-chromosome SNPs in a single multiplex analysis, and to define the population genetic variation that resides within the panel of selected markers.

The chapters of this dissertation describe the results and findings of this doctoral research. **Chapters 1 and 2** address the development and testing of STRait Razor, a novel bioinformatic software tool designed to detect STR alleles in raw sequencing data. At the time this research began, the existing MPS instruments were capable of providing extensive data, but the available software tools were inadequate for identifying forensic STR alleles. Without suitable software for STR analysis, the process of data interpretation was manual, tedious, and time-consuming, and comparison of results was difficult. Therefore, this dissertation project required the development of specialized STR-typing software for STR data generated by MPS. **Chapter 3** describes the initial evaluation of MPS for the detection of forensically-relevant human identity markers. Although this study focused only on the detection of SNP markers through the use of the PCR multiplex-based TruSeq™ Forensic Amplicon library preparation method (32), it served as a proof-of-concept trial that laid the foundation for subsequent dissertation research. **Chapter 4** explains the development and testing of a comprehensive MPS panel for the detection of human identity SNPs and STRs using the Nextera™ Rapid Capture library preparation method (33). The Nextera™ chemistry was chosen because it is a capture-based method (Figure 6). As opposed to a PCR multiplex-based method, which relies on efficient multiplex design to amplify each target for sequencing, the Nextera™ library preparation method cleaves the input DNA into relatively uniform fragments. Custom oligonucleotide probes then are used to selectively bind (or “capture”) those fragments containing the target regions of interest, while the other fragments are washed

away. This method was considered ideal for a large panel of markers because it eliminates the difficult task of multiplex design for such an extensive battery of loci and was theorized to be less prone to amplification artifacts. The results of this in-depth study illustrate that MPS is indeed capable of typing a large number of forensically-relevant markers with relative ease.

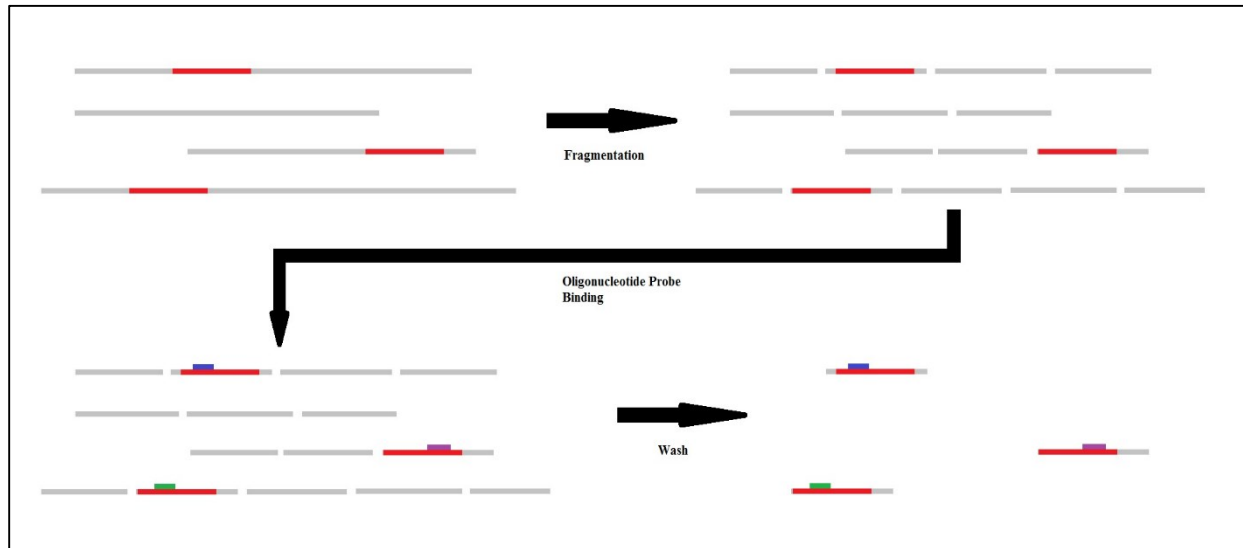


Figure 6. Capture-based library preparation. A generic capture-based library preparation method is illustrated. Input DNA is shown in grey, while regions of interest are depicted in red. Oligonucleotide probes are shown in blue, purple, and green.

Finally, although not the primary focus of this research, **Chapter 5** details the identification and description of intra-repeat sequence variants using Y-chromosome STR data, to illustrate the value of MPS and the additional diversity that exists. Future work should concentrate on describing the variation of autosomal and X-chromosome STRs and development of software tools to facilitate intra-repeat variation analyses.

REFERENCES

1. Budowle, B. and Eisenberg, A.J. 2007. Forensic Genetics. In: Emery and Rimoin's Principles and Practice of Medical Genetics, fifth edition, Vol. 1, Rimoin, D.L., Connor, J.M., Pyeritz, R.E., and Korf, B.R. (eds.), Elsevier, Philadelphia, pp.501-517.
2. Budowle, B., Planz, J.V., Campbell, R., and Eisenberg, A.J. 2005. Molecular diagnostic applications in forensic science. In: Molecular Diagnostics, Patrinos, G. and Ansorge, W., (eds.), Elsevier, Amsterdam, pp. 267-280.
3. Budowle, B. and van Daal, A. 2008. Forensically relevant SNP classes. *Biotechniques* 44:603-610.
4. Budowle, B., Moretti, T.R., Niezgoda, S.J., and Brown, B.L. 1998. CODIS and PCR-based short tandem repeat loci: Law enforcement tools. In: Second European Symposium on Human Identification 1998, Promega Corporation, Madison, Wisconsin, pp. 73-88.
5. Martin, P.D., Schmitter, H., and Schneider, P.M. 2001. A brief history of the formation of DNA databases in forensic science within Europe. *Forens. Sci. Int.* 119(2):225-231.
6. CODIS-NDIS Statistics: <http://www.fbi.gov/about-us/lab/biometric-analysis/codis/ndis-statistics>
7. Budowle, B.: Familial searching: extending the investigative lead potential of DNA typing. Profiles in DNA 13(2), 2010, Available at: www.promega.com/profiles/1302/1302_07.html.
8. Ge, J., Chakraborty, R., Eisenberg, A. and Budowle, B. 2011. Comparisons of the familial DNA database searching policies. *J. Forens. Sci.* 56(6):1448-1456.
9. Ge, J., Eisenberg, A., and Budowle, B. 2012. Developing criteria and data to determine best options for expanding the core CODIS loci. *BMC Investigative Genetics* 3:1.
10. Hares, D.R. 2015. Selection and implementation of expanded CODIS core loci in the United States. *Forensic Sci. Int. Genet.* 17:33-34.
11. Hares, D.R. 2011. Expanding the CODIS core loci in the United States. *Forensic Sci. Int. Genet.* Doi:10.1016/j.fsigen.2011.04.012.
12. DNA Sequencing Costs: <https://www.genome.gov/sequencingcosts/>
13. Sanger, F., Nicklen, S., and Coulson, A.R. 1977. DNA sequencing with chain-terminating inhibitors. *Proc. Natl. Acad. Sci. USA* 74(12):5463-5467.
14. Wilson, M.R., Polansky, D., Butler, J., DiZinno, J.A., Replogle, J., and Budowle, B. 1995. Extraction, PCR amplification, and sequencing of mitochondrial DNA from human hair shafts. *BioTechniques* 18:662-669.

15. Wilson, M.R., DiZinno, J.A., Polanskey, D., Replogle, J., and Budowle, B. 1995. Validation of mitochondrial DNA sequencing for forensic casework analysis. *Int. Journal Leg. Med.* 108:68-74.
16. Budowle, B. 2004. SNP typing strategies. *Forens. Sci. Int.* 146(suppl):S139-S142.
17. Budowle, B., Planz, J., Campbell, R., and Eisenberg, A. 2004. SNPs and microarray technology in forensic genetics: development and application to mitochondrial DNA. *Forens. Sci. Rev.* 16:22-36.
18. Illumina® MiSeq™ Specifications: http://www.illumina.com/documents/products/datasheets/datasheet_miseq.pdf.
19. Rothberg JM, Hinz W, Rearick TM, *et al.* 2011. An integrated semiconductor device enabling non-optical genome sequencing. *Nature* 475(7356):348-352.
20. Adessi, C., Matton, G., Ayala, G., Turcatti, G., Mermoud, J.J., Mayer, P., and Kawashima, E. 2000. Solid phase DNA amplification: characterisation of primer attachment and amplification mechanisms. *Nucleic Acids Res.* 28(20):E87.
21. Brenner, S., Johnson, M., Bridgham, J., *et al.* 2000. Gene expression analysis by massively parallel signature sequencing (MPSS) on microbead arrays. *Nat. Biotechnol.* 18:630-634.
22. Brenner, S., Williams, S.R., Vermaas, E.H., Storck, T., Moon, K., McCollum, C., Mao, J.I., Luo, S., Kirchner, J.J., Eletr, S., DuBridge, R.B., Burcham, T., and Albrecht, G. 2000. In vitro cloning of complex mixtures of DNA on microbeads: physical separation of differentially expressed cDNAs. *Proc. Natl. Acad. Sci. USA* 97:1665-1670.
23. Holt, K.E., Parkhill, J., Mazzoni, C.J., Roumagnac, P., Weill, F.X., Goodhead, I., Rance, R., Baker, S., Maskell, D.J., Wain, J., Dolecek, C., Achtman, M., and Dougan, G. 2008. High-throughput sequencing provides insights into genome variation and evolution in *Salmonella Typhi*. *Nat. Genet.* 40:987-993.
24. Margulies, M., Egholm, M., Altman, W.E., *et al.* 2005. Genome sequencing in microfabricated high-density picolitre reactors. *Nature* 437:376-380.
25. Quail, M.A., Kozarewa, I., Smith, F., Scally, A., Stephens, P.J., Durbin, R., Swerdlow, H., and Turner, D.J. 2008. A large genome center's improvements to the Illumina sequencing system. *Nat. Methods* 5:1005-1010.
26. Van Tassel, C.P., Smith, T.P., Matukumalli, L.K., Taylor, J.F., Schnabel, R.D., Lawley, C.T., Haudenschild, C.D., Moore, S.S., Warren, W.C., and Sonstegard, T.S. 2008. SNP discovery and allele frequency estimation by deep sequencing of reduced representation libraries. *Nat. Methods* 5:247-252.

27. Wheeler, D.A., Srinivasan, M., Egholm, M., Shen, Y., Chen, L., McGuire, A., He, W., Chen, Y.J., *et al.* 2008. The complete genome of an individual by massively parallel DNA sequencing. *Nature* 452:872-876.
28. Zeng, X., King, J.L., Stoljarova, M., Warshauer, D.H., LaRue, B.L., Sajantila, A., Patel, J., Storts, D.R., and Budowle, B. 2015. High sensitivity multiplex short tandem repeat loci analyses with massively parallel sequencing. *Forensic Sci. Int. Genet.* 16:38-47.
29. Ottaviani, E., Vernarecci, S., Asili, P., Agostino, A., and Montagna, P. 2014. Preliminary assessment of the prototype Yfiler® Plus kit in a population study of Northern Italian males. *Int. J. Legal Med.* Doi:10.1007/s00414-014-1078-x.
30. Zeng, X., King, J., Hermanson, S., Patel, J., Storts, D.R., and Budowle, B. An evaluation of the PowerSeq™ Auto system: a multiplex short tandem repeat marker kit compatible with massively parallel sequencing. *Forensic Sci. Int. Genet.* (submitted).
31. Illumina® ForenSeq™ information: <http://www.illumina.com/products/forenseq-dna-signature-kit.html>
32. Warshauer, D.H., Davis, C.P., Holt, C., Han, Y., Walichiewicz, P., Richardson, T., Stephens, K., Jager, A., King, J., and Budowle, B. 2015. Massively parallel sequencing of forensically relevant single nucleotide polymorphisms using TruSeq™ Forensic Amplicon. *Int. J. Legal Med.* 129(1):31-36.
33. Nextera™ Rapid Capture information: <http://www.illumina.com/products/nextera-rapid-capture-custom-enrichment-kit.html>

SECTION 1

Development of a Novel Bioinformatic Tool for the Detection of STR Alleles in Raw Massively Parallel Sequencing Data

CHAPTER 1

STRait Razor: A Length-Based Forensic STR Allele-Calling Tool for Use with Second Generation Sequencing Data

Published in Forensic Science International: Genetics
2013, 7:409-417

David H. Warshauer
David Lin
Kumar Hari
Ravi Jain
Carey Davis
Bobby LaRue
Jonathan King
Bruce Budowle

ABSTRACT

Recent studies have demonstrated the capability of second generation sequencing (SGS) to provide coverage of short tandem repeats (STRs) found within the human genome. However, there are relatively few bioinformatic software packages capable of detecting these markers in the raw sequence data. The extant STR-calling tools are sophisticated, but are not always applicable to the analysis of the STR loci commonly used in forensic analyses. STRait Razor is a newly developed Perl-based software tool that runs on the Linux/Unix operating system and is designed to detect forensically-relevant STR alleles in FASTQ sequence data, based on allelic length. It is capable of analyzing STR loci with repeat motifs ranging from simple to complex without the need for extensive allelic sequence data. STRait Razor is designed to interpret both single-end and paired-end data and relies on intelligent parallel processing to reduce analysis time. Users are presented with a number of customization options, including variable mismatch detection parameters, as well as the ability to easily allow for the detection of alleles at new loci. In its current state, the software detects alleles for 44 autosomal and Y-chromosome STR loci. The study described herein demonstrates that STRait Razor is capable of detecting STR alleles in data generated by multiple library preparation methods and two Illumina® sequencing instruments, with 100% concordance. The data also reveal noteworthy concepts related to the effect of different preparation chemistries and sequencing parameters on the bioinformatic detection of STR alleles.

KEYWORDS: STR · Bioinformatics · Software · Second generation sequencing · MiSeq™ · GAIIX™

INTRODUCTION

The majority of genetic analyses used in forensic casework involve the detection of short tandem repeats (STRs), which are relatively small sequences of DNA made up of repeating units of 2– 6 nucleotides [1–3]. Currently, the accepted means of detecting these markers is size separation by capillary electrophoresis (CE) [4–7]. In recent years, however, second generation sequencing (SGS) technology has advanced to the point that it can be considered a viable platform for forensic DNA analyses, including STR detection [8–11]. In addition, sequencing has long been regarded as an effective means of revealing individual base variations in DNA known as single nucleotide polymorphisms (SNPs) [11–15]. While the traditional CE-based method of STR detection reveals only the length of alleles, SGS can increase resolution and detect nucleotide variation within the repeat regions and in proximal flanking regions [16]. Furthermore, the expansive genetic coverage and read length associated with current SGS methods allow for the potential capture of genetic information related to far more forensic STR markers than are possible with conventional multiplex-based CE kits. With the ability to sequence gigabases of DNA [17,18], a properly designed assay could yield STR information in a single analysis which surpasses that of all the currently available commercial CE-based kits combined, and do so on multiple samples simultaneously. However, while the extant sequencing instruments are capable of providing such extensive data, currently available software tools for identifying forensic STR alleles within the data are only beginning to address the task.

One such tool, lobSTR [8], uses an algorithm specifically designed to identify STR alleles within SGS data. First, the software analyzes a raw FASTA/FASTQ or BAM input file, detecting

reads that contain an STR sequence and identifying the repeat motif. Next, lobSTR aligns the regions that flank the STR sequence to a modified reference sequence. Finally, the algorithm determines the identity of the allele(s) based on the number of detected repeat units between the two flanking regions, applying statistical corrections to produce the most likely allelotype. While this software is certainly a refined and reasonably accurate method, it is somewhat limited in that lobSTR, by default, identifies only a single simple repeat motif. To allow the software to detect alleles at STRs that have longer, complex repeats, such as those within the D21S11 locus, for example, the user must determine the distinct simple repeats that comprise the complex motif and instruct lobSTR to identify each of these repeats individually. The resulting data must then be interpreted altogether in order to draw conclusions. This necessity makes lobSTR less applicable for the analysis of forensic STR markers, which display varying repeat motif complexity.

Recently, a method was introduced by Bornman *et al.* [9] that allows for the detection of STR alleles in SGS data using a different strategy. This method uses the Bowtie short read aligner [19] to align raw SGS reads to an “in silico reference,” which is a user-generated FASTA file containing the full sequence of each allele at each STR locus. To reduce erroneous allele calls, reads are filtered so that only those encompassing the entire repeat region defined in the reference file are used for allelotyping. Allele calls are made using a heuristic decision model based on Fisher’s exact test, and probability values are given for each allele call. This software also is effective for identifying STR alleles in sequence data, but requires prior knowledge of allelic sequence information. As a result, novel alleles or allelic variants, or those for which there are no documented sequence data, are a limitation for this system.

STRait Razor (the STR allele identification tool – Razor) is a Perl script designed for the Linux/Unix platform that identifies alleles at forensic STR loci based on the length of the repeat sequence, a method that is conceptually similar to the length-based allele detection offered by CE. This software is capable of handling repeat motifs ranging from simple to complex, and it does not require a reference composed of extensive allelic sequence data. As a result, the allele call results are consistent with those of current CE-based methods, and it is not confounded by unexpected sequence variation within repeats. In its first iteration of development, STRait Razor is capable of detecting alleles at 44 forensically-relevant STR loci, and others can be configured readily.

Fordyce *et al.* [10] independently developed a software tool that functions similarly to STRait Razor, isolating the repeat region of interest and performing length-based allelotyping. However, the algorithm was designed for use with the Roche® Genome Sequencer FLX™ and is only able to analyze FASTA files consisting of sequence data that contain Roche® Molecular Identifier (MID) tags. Currently this software only is able to identify alleles at 5 STR loci, compared with the 44 STR loci detected by STRait Razor. While STRait Razor was only tested on raw FASTQ files output by Illumina® instruments in this study, the software should, in theory, maintain compatibility with the raw read files generated by any second generation sequencing platform.

MATERIALS AND METHODS

To test the efficiency and accuracy of STRait Razor, an initial concordance study was performed wherein allele calls made by the software were compared with CE results. Following

the University of North Texas Health Science Center Institutional Review Board approval, DNA from five Caucasian blood samples was used for this trial.

Sample preparation and CE typing

Blood samples were extracted using the Qiagen® QIAamp® DNA Mini kit (Qiagen Inc., Valencia, CA), following the manufacturer's recommendations. The quantity of extracted DNA was determined using the Applied Biosystems® Quantifiler™ Human DNA Quantification Kit (Life Technologies, Carlsbad, CA) on an Applied Biosystems® 7500 Real-Time PCR System (Life Technologies), according to the manufacturer's protocol. Amplification was performed using the reagents from the Applied Biosystems® AmpFØSTR® Identifiler® PCR Amplification Kit (Life Technologies) and the Promega® PowerPlex® 16 HS, ESI 17 Pro, and Y23 Systems (Promega Corporation, Madison, WI), on an Applied Biosystems® GeneAmp® PCR System 9700 thermal cycler (Life Technologies), according to the manufacturer's recommendations. These various kits allowed for the typing of the following STR loci: CSF1PO, D13S317, D16S539, D18S51, D19S433, D21S11, D2S1338, D3S1358, D5S818, D7S820, D8S1179, FGA, TH01, TPOX, vWA, PENTA D, PENTA E, D10S1248, D12S391, D1S1656, D22S1045, D2S441, SE33, DYS19, DYS385, DYS389I/II, DYS390, DYS391, DYS392, DYS393, DYS437, DYS438, DYS439, DYS448, DYS456, DYS458, DYS481, DYS533, DYS549, DYS570, DYS576, DYS635, DYS643, and GATA H4. Capillary electrophoresis was performed on an Applied Biosystems® 3130xl Genetic Analyzer (Life Technologies) using POP-4™ polymer (Life Technologies) and analyzed using Applied Biosystems® GeneMapper® ID v3.2 software (Life Technologies), according to the manufacturer's recommended protocol.

Sample preparation, SGS, and STRait Razor typing

The quantity of extracted DNA from the blood sample was determined using a Qubit® 2.0 Fluorometer (Life Technologies). Library preparation prior to sequencing was performed using either the Illumina® TruSeq™ Custom Enrichment protocol (Illumina, Inc., San Diego, CA) or the Agilent Technologies HaloPlex™ Target Enrichment protocol (Agilent Technologies, Inc., Santa Clara, CA). Using the DesignStudio (Illumina, Inc.) and SureDesign (Agilent Technologies, Inc.) software, respectively, custom probes were designed to target the forensically-relevant STR loci. Paired-end sequencing was carried out on the GAIIx™ and MiSeq™ sequencing platforms (Illumina, Inc.). For these trials, the read lengths employed by these instruments were 2x146 and 2x251, respectively. Sample 1 was prepared using the HaloPlex™ chemistry and sequenced on both the GAIIx™ and MiSeq™ instruments. The sample also was prepared using the TruSeq™ chemistry, and subsequently sequenced on the MiSeq™. Sample 2 was prepared using the HaloPlex™ chemistry and sequenced on the GAIIx™. Samples 3, 4, and 5 were prepared using the TruSeq™ chemistry and sequenced on the MiSeq™. Following sequencing, the GAIIx™ output *bcl* files were demultiplexed and converted to a single FASTQ file using CASAVA v1.8.2 [20]. The MiSeq™ output was automatically converted to FASTQ format by the MiSeq™ Reporter software [21]. These FASTQ files served as the input for STRait Razor. The software was designed to detect the following forensic STR loci: CSF1PO, TPOX, D2S441, D3S1358, D5S818, D13S317, D18S51, D16S539, D7S820, D8S1179, TH01, vWA, D21S11, FGA, D2S1338, D19S433, PENTA D, PENTA E, D10S1248, D12S391, D1S1656, D22S1045, DYS389I/II, DYS390, DYS456, DYS19, DYS458, DYS437, DYS438, DYS448, GATA H4, DYS391, DYS392, DYS393, DYS439, DYS481, DYS533, DYS549, DYS570, DYS576,

DYS643, DYS385, and DYS635. For these trials, allele-calling was performed using STRait Razor's default flank recognition settings (1 allowable substitution and no allowable insertions or deletions). The server used for STRait Razor analysis was a Dell™ PowerEdge™ R900 blade server, with 64 GB DDR3 ECC RAM and 4 Quad-Core Intel® Xeon® E7430 CPUs (2.13 GHz each).

STrait Razor information

The algorithm employed by STRait Razor is relatively simple (Fig. 1). First, reads containing both a leading and trailing flanking region (Tables 1 and 2) surrounding the repeat sequence are extracted from the raw FASTQ sequence file(s) using the AGREP function, an approximate string search tool that is capable of detecting inexact matches to a provided query [22]. During this process, the software allows for customizable mismatch, insertion, and deletion penalties for each flanking region. This first step ensures that the extracted reads encompass the full repeat sequence, as partial repeat sequence data cannot be used for accurate allele-calling. With the reads containing the complete repeat sequence, the regions adjacent to the repeat sequence on each side are trimmed or “shaved” away (hence the program's name), leaving only the repeats themselves. At this point, the reads are filtered based on the presence of the repeat motif specific to the STR locus, thus discarding the majority of any irrelevant sequence data that may have been inadvertently captured. Next, the number of nucleotides in each repeat sequence is counted and compared with the expected lengths of alleles at that locus, based on the repeat motif. For example, at a STR locus with a simple tetranucleotide repeat motif, a repeat sequence consisting of 48 bases would indicate the presence of a “12” allele, while a sequence of 50 bases

would indicate the presence of a “12.2” allele. Alleles can be called in this fashion regardless of intra-sequence variation or repeat motif complexity, and the length-based method of detection, to date, is concordant with CE results. The repeat region sequences of the called alleles then are sorted by length and written to a text file specific to the STR locus. With this data output, analysts can observe every nucleotide and evaluate any variations within the repeat sequences, as desired. This is not currently an automated process and requires manual inspection of the sequence data. Finally, a colon-delimited text file is generated that lists the alleles called at each STR locus, including the number of reads in which each allele was detected. These read count values can be used as a measure of abundance, similar to the RFU values in electropherograms.

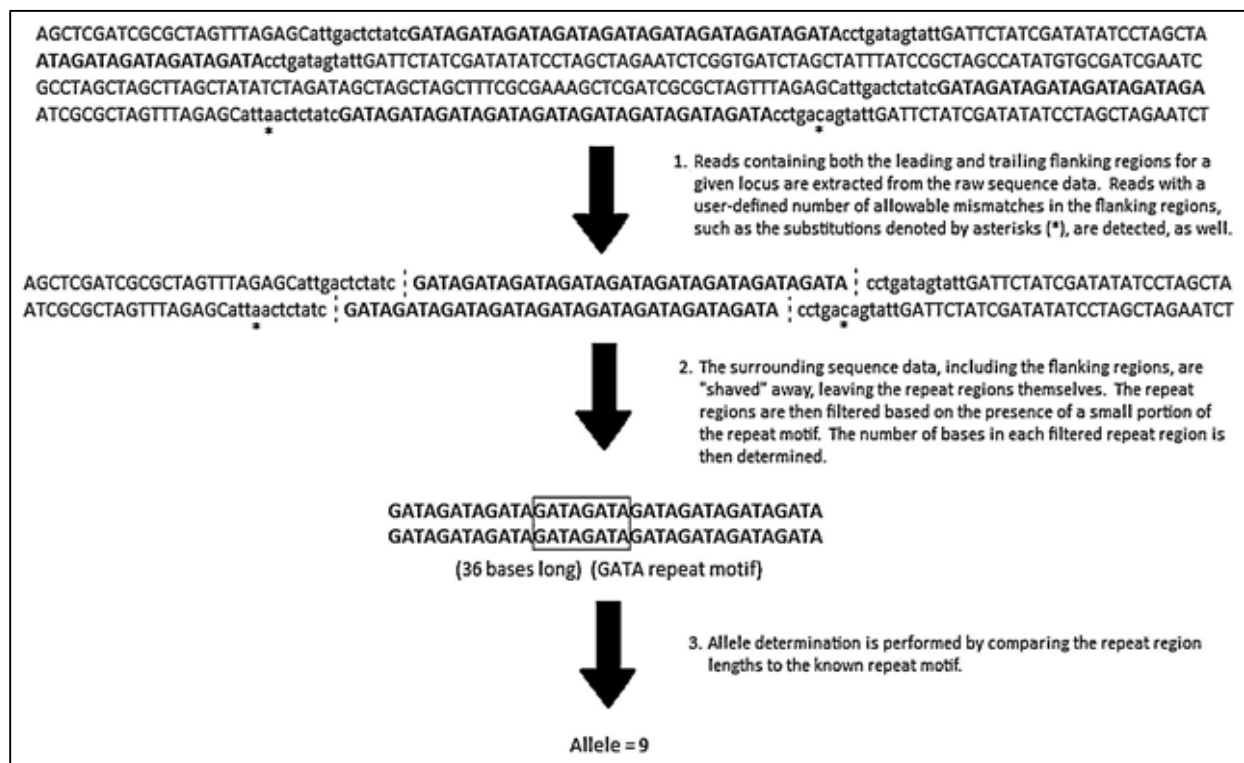


Fig. 1. STRait Razor algorithm. In this figure, each line represents an individual sequence read within the input FASTQ file. The repeat regions are shown in bold, capitalized font, while the flanking regions are shown in plain, lowercase font. Surrounding sequences are shown in plain, capitalized font.

STRait Razor is designed to analyze both single-end and paired-end data. Thus, the program accepts either single input files (Read 1) or dual input files (Reads 1 and 2), and recognizes STR loci in both forward and reverse complement forms. Each locus is analyzed separately, such that trimmed repeat regions always remain associated with their respective loci. The speed with which STRait Razor can provide allele calls is directly related to the size of the input file(s), as well as the depth of reads that contain the queried STR loci. However, the software utilizes PPSS, the (Distributed) Parallel Processing Shell Script [23], which allows STRait Razor to analyze one STR locus per available processor core in parallel, thus reducing the amount of time needed for analysis. Also, the user is able to choose whether STRait Razor detects only autosomal STR alleles, Y-chromosome STR alleles, or both. This option can further reduce analysis time and may be useful in cases wherein analysts wish to investigate only a subset of the loci recognized by STRait Razor, such as the typing of a known female reference sample. The information required by STRait Razor to detect STR alleles is provided in a modular format, and a workbook is provided to allow users to easily add other STR loci to the modules. Flanking regions queried by STRait Razor are each 12 bases long. These lengths allow for enough nucleotide diversity to make them fairly specific for the repeat regions of interest. Users may choose to use different flanking region sequences, as well as different flank lengths, according to their preferences. While most flanking regions used by STRait Razor are directly adjacent to the repeat regions, 6 sets (those for loci D1S1656, Penta D, DYS385, DYS393, DYS481, and DYS635) were selected at different proximities to allow for increased specificity.

Locus	Flanking Region Sequences	Detectable Alleles
CSF1PO	GATAGATAGATT-----AGGAAGTACTTA TAAGTACTTCTC-----AATCTATCTATC	5, 6, 6.3, 7, 7.3, 8, 8.3, 9, 9.1, 10, 10.1, 10.2, 10.3, 11, 11.1, 11.3, 12, 12.1, 13, 14, 15, 16
D10S1248	TATTGTCTTCAT-----ACTCACTCATTT AAATGAGTGAGT-----ATGAAGACAATA	8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19
D12S391	AAATCCCCTCTC-----ACCTATGCATCC GGATGCATAGGT-----GAGAGGGGATTT	15, 16, 17, 17.3, 18, 18.3, 19, 19.3, 20, 21, 22, 23, 24, 25, 26
D13S317	AGATGATTGATT-----ATGTATTTGTAA TTACAAATACAT-----AATCAATCATCT	5, 6, 7, 7.1, 8, 8.1, 9, 10, 11, 11.1, 11.3, 12, 13, 13.3, 14, 14.3, 15, 16, 17
D16S539	GACAGACAGGTG-----TCATTGAAAGAC GTCTTTCAATGA-----CACCTGTCTGTC	4, 5, 6, 7, 8, 8.3, 9, 9.3, 10, 11, 11.3, 12, 12.1, 12.2, 13, 13.1, 13.3, 14, 14.3, 15, 16
D18S51	TCCTCTCTCTTT-----GAGACAAGGTCT AGACCTTGCTCT-----AAAGAGAGAGGA	7, 8, 9, 9.2, 10, 10.2, 11, 11.1, 11.2, 12, 12.2, 12.3, 13, 13.1, 13.2, 13.3, 14, 14.2, 15, 15.1, 15.2, 15.3, 16, 16.1, 16.2, 16.3, 17, 17.1, 17.2, 17.3, 18, 18.1, 18.2, 19, 19.2, 20, 20.1, 20.2, 21, 21.1, 21.2, 22, 22.1, 22.2, 23, 23.1, 23.2, 24, 24.2, 25, 26, 27, 28.1, 28.3, 39.2
D19S433	AAGATTCTGTTG-----AGAGAGGTAGAA TTCTACCTCTCT-----CAACAGAATCTT	5.2, 6.2, 7, 8, 9, 10, 11, 11.1, 12, 12.1, 12.2, 13, 13.1, 13.2, 13.3, 14, 14.1, 14.2, 14.3, 15, 15.2, 16, 16.2, 17, 17.2, 18, 18.2, 19, 19.2, 20
D1S1656	TAAACACACACA-----CATCATACAGTT* AACTGTATGATG-----TGTGTGTGTTA*	9, 10, 11, 12, 13, 13.3, 14, 14.3, 15, 15.3, 16, 16.3, 17, 17.1, 17.3, 18, 18.3, 19, 19.3, 20, 20.3, 21
D21S11	ATAGATAGACGA-----AGGCAATCACT AGTGAATTGCCT-----TCGTCTATCTAT	12, 24, 24.2, 24.3, 25, 25.1, 25.2, 25.3, 26, 26.1, 26.2, 27, 27.1, 27.2, 27.3, 28, 28.1, 28.2, 28.3, 29, 29.1, 29.2, 29.3, 30, 30.1, 30.2, 30.3, 31, 31.1, 31.2, 31.3, 32, 32.1, 32.2, 32.3, 33, 33.1, 33.2, 33.3, 34, 34.1, 34.2, 34.3, 35, 35.1, 35.2, 35.3, 36, 36.1, 36.2, 36.3, 37, 37.2, 38, 38.2, 39, 39.2, 40.2, 41.2
D22S1045	TATTTTTATAAC-----GAGACTACTATC GATAGTAGTCTC-----GTTATAAAAATA	8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20
D2S1338	GGATTGCAGGAG-----AGGCCAAGCCAT ATGGCTTGGCCT-----CTCCTGCAATCC	11, 12, 13, 14, 15, 16, 17, 18, 19, 19.3, 20, 21, 22, 23, 23.2, 23.3, 24, 25, 26, 27, 28
D2S441	TCTATGAAAAC-----TATCATAACACC GGTGTTATGATA-----AGTTTTCATAGA	8, 9, 10, 11, 11.3, 12, 12.3, 13, 13.3, 14, 14.3, 15, 16, 17
D3S1358	AGGCTTGCATGT-----ATGAGACAGGGT ACCCTGTCTCAT-----ACATGCAAGCCT	8, 8.3, 9, 10, 11, 12, 13, 14, 14.3, 15, 15.1, 15.2, 15.3, 16, 16.2, 17, 17.1, 17.2, 18, 18.1, 18.2, 18.3, 19, 20
D5S818	ATTTATACCTCT-----TCAAAAATATTAC GTAATATTTTGA-----AGAGGTATAAAT	6, 7, 8, 9, 10, 10.1, 11, 11.1, 12, 12.3, 13, 14, 15, 16, 17, 18
D7S820	GAACGAACTAAC-----GACAGATTGATA TATCAATCTGTC-----GTTAGTTCGTTT	5, 5.2, 6, 6.2, 6.3, 7, 7.1, 7.3, 8, 8.1, 8.2, 8.3, 9, 9.1, 9.2, 9.3, 10, 10.1, 10.3, 11, 11.1, 11.3, 12, 12.1, 12.2, 12.3, 13, 13.1, 14, 14.1, 15, 16
D8S1179	CACTGTGGGGAA-----TACGAATGTACA TGTACATTCGTA-----TTCCCCACAGTG	7, 8, 9, 10, 10.2, 11, 12, 12.3, 13, 14, 15, 15.1, 15.3, 16, 17, 17.1, 18, 19, 20
FGA	GAAAGGAAGAAA-----CTAGCTTGTAAT TTTACAAGCTAG-----TTTCTTCCTTTC	12.2, 13, 13.2, 14, 14.3, 15, 15.3, 16, 16.1, 16.2, 17, 17.1, 17.2, 18, 18.1, 18.2, 19, 19.1, 19.2, 19.3, 20, 20.1, 20.2, 20.3, 21, 21.1, 21.2, 21.3, 22, 22.1, 22.2, 22.3, 23, 23.1, 23.2, 23.3, 24, 24.1, 24.2, 24.3, 25, 25.1, 25.2, 25.3, 26, 26.1, 26.2, 26.3, 27, 27.1, 27.2, 27.3, 28, 28.1, 28.2, 29, 29.1, 29.2, 30, 30.2, 31, 31.2, 32, 32.1, 32.2, 33.1, 33.2, 34.1, 34.2, 35.2, 41.1, 41.2, 42, 42.1, 42.2, 43.1, 43.2, 44, 44.2, 44.3, 45, 45.1, 45.2, 46, 46.1, 46.2, 47, 47.2, 48, 48.2, 49, 49.1, 49.2, 50.2, 50.3, 51, 51.2
Penta D	TTTATGATTCTC-----TTGAGATGGTGT* ACACCATCTCAA-----GAGAATCATAAA*	1.1, 1.2, 2.2, 3.2, 4, 5, 6, 6.4, 7, 7.1, 7.4, 8, 8.1, 9, 9.1, 9.4, 10, 10.1, 10.2, 10.3, 11, 11.1, 11.2, 12, 12.1, 12.2, 12.3, 12.4, 13, 13.2, 13.3, 13.4, 14, 14.1, 14.4, 15, 15.1, 16, 17, 18
Penta E	TCCTTACAATTT-----GAGACTGAGTCT AGACTCAGTCTC-----AAATTGTAAGGA	5, 6, 7, 8, 9, 9.1, 9.4, 10, 10.2, 11, 11.4, 12, 12.1, 12.2, 12.3, 13, 13.2, 13.4, 14, 14.4, 15, 15.2, 15.4, 16, 16.4, 17, 17.4, 18, 18.4, 19, 19.4, 20, 20.2, 20.3, 21, 22, 23, 23.4, 24, 26
TH01	CCCTTATTTCCC-----TCACCATGGAGT ACTCCATGGTGA-----GGGAAATAAGGG	3, 4, 5, 5.3, 6, 6.1, 6.3, 7, 7.1, 7.3, 8, 8.3, 9, 9.1, 9.3, 10, 10.3, 11, 12, 13.3, 14
TPOX	GAACCCCTCACTG-----TTTGGGCAAATA TATTTACCCAAA-----CAGTGAGGGTTC	4, 5, 6, 7, 7.3, 8, 9, 10, 10.1, 10.3, 11, 12, 13, 13.1, 14, 15, 16
vWA	GACTTGGATTGA-----TCCATCCATCCT AGGATGGATGGA-----TCAATCCAAGTC	10, 11, 12, 13, 14, 15, 15.2, 16, 16.1, 17, 18, 18.1, 18.2, 18.3, 19, 19.2, 20, 21, 22, 23, 24, 25

Table 1. Autosomal STR loci detected by STRait Razor. Leading and trailing flanking sequences for both forward

and reverse complement reads are listed. All flanking sequences are directly adjacent to the repeat region, except for those labeled with an asterisk (*), which were modified to allow for increased specificity or more efficient detection.

Additional alleles can be added at the discretion of the user.

Allele call comparison

The alleles detected by the CE method were compared with the allele call output files generated by STRait Razor. To be considered concordant, alleles detected via CE had to be detected by STRait Razor, based on the presence of the allelic sequence data in the input FASTQ file. It should be noted that alleles not detected by STRait Razor due to lack of respective sequence data, whether the result of kit chemistry limitations or inadequate sequencing read length, were not considered discordant.

RESULTS

Based on the various combinations of library preparation methods and sequencing platforms utilized, a total of 7 overall comparisons with CE data were made for the 5 samples. The assortment of multiplex-based CE kits used for these samples allowed for the comparison of STR alleles at all loci that are currently detectable with STRait Razor: 22 autosomal STR loci and 22 Y-STR loci, with DYS385 designated as a single locus. Across all trials, a total of 427 alleles were compared. The allele calls made by STRait Razor were completely concordant with the genotype results generated by the CE method. Of the 427 alleles compared, 403 alleles were detected, with 100% concordance (Tables 3 and 4). For the 24 alleles not detected by the software, a subsequent manual analysis of the FASTQ input files revealed that there were no sequence reads for these alleles in which the full repeat regions (including the surrounding flanking sequences) were present. These undetected alleles were not represented in the sequencing data because of the library preparation method (e.g., random shearing of genomic DNA) and/or the read length used, and thus

could not be recognized by STRait Razor. These instances do not reflect inadequacy of the software and should not be considered evidence of discordance. Intra-repeat variation was occasionally observed in the data output by STRait Razor. For example, an examination of the sequence data output file for Sample 1 revealed that the “15” allele at locus D3S1358 consisted of both the “TCTA(TCTG)₂(TCTA)₁₂” variant and the “TCTA(TCTG)₃(TCTA)₁₁” variant, at a ratio of approximately 1:1. Conversely, the sequence data output file for Sample 2, which also displays a “15” allele at this locus, revealed that only the “TCTA(TCTG)₂(TCTA)₁₂” variant was present. This difference would not be detected by traditional CE methods. The time required for analysis, using a paired-end analysis method that detected both autosomal and Y chromosome STR alleles, ranged from 16 min (for dual 395 MB input files) to 285 min (for dual 7.9 GB input files) on the 16-core server used for this study.

Locus	Flanking Region Sequences	Detectable Alleles
DYS19	TATATAGTGTTT-----TATAGTGACACT AGTGTCACTATA-----AAACACTATATA	9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19
DYS385	GAGAAAGAAAGG-----GGAGGACTATGT* ACATAGTCCTCC-----CCTTCTTTCTC*	7, 8, 9, 9.2, 10, 10.2, 11, 11.2, 11.3, 12, 12.1, 12.2, 12.3, 13, 13.1, 13.2, 13.3, 14, 14.2, 14.3, 15, 15.1, 15.2, 15.3, 16, 16.2, 16.3, 17, 17.1, 17.2, 17.3, 18, 18.1, 18.2, 19, 19.2, 19.3, 20, 21, 22, 23, 24, 25, 28
DYS389I	ATTATCTATGTA-----TCCCTCCCTCTA TAGAGGGAGGGA-----TACATAGATAAT	9, 10, 11, 12, 13, 14, 15, 16, 17
DYS389II	TCTATGTGTGTG-----TCCCTCCCTCTA TAGAGGGAGGGA-----CACACACATAGA	24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36
DYS390	ATATTCTATCTA-----TCATCTATCTAT ATAGATAGATGA-----TAGATAGAATAT	17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29
DYS391	TCTGTCTGTCTG-----TCTGCCTATCTG CAGATAGGCAGA-----CAGACAGACAGA	5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16
DYS392	TCACCATTTATT-----TTACTAAGGAAT ATTCTTAGTAA-----AATAAATGGTGA	4, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 20
DYS393	TTGTGTCAATAC-----GAGACATACCTC* GAGGTATGTCTC-----GTATTGACACAA*	7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18
DYS437	ATGCCCATCCGG-----TCATCTATCATC GATGATAGATGA-----CCGGATGGGCAT	10, 11, 12, 13, 14, 15, 16, 17, 18, 19
DYS438	GTAAACAGTATA-----TATTTGAAATGG CCATTTCAAATA-----TATACTGTTTAC	6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 18
DYS439	TGATAAATAGAA-----GAAAGTATAAGT ACTTATACTTTC-----TTCTATTATCA	5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19
DYS448	AGACATGGATA-----AGAGAGGTAAAG CTTTACCTCTCT-----TTATCCATGTCT	14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24
DYS456	TGTGATAATGTA-----ATTCCATTAGTT AACTAATGGAAT-----TACATTATCACA	9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 22, 23
DYS458	AAGAAAAGGAAG-----GGAGGGTGGGCG CGCCCACTCTCC-----CTTCCTTTTCTT	11, 12, 13, 14, 15, 16, 16.2, 17, 17.2, 18, 18.2, 19, 19.2, 20, 21, 22, 23, 24
DYS481	TTCAGCATGCTG-----GAGTCTTGCAAC* GTTGCAAGACTC-----CAGCATGCTGAA*	16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32
DYS533	TAGCTAGCTATC-----ATCATCTATCAT ATGATAGATGAT-----GATAGCTAGCTA	7, 8, 9, 10, 11, 12, 13, 14, 15
DYS549	GATTAGAAAGAT-----GAAAAAATCTAC GTAGATTTTTTC-----ATCTTTCTAATC	9, 10, 11, 12, 13, 14, 15
DYS570	CTCCAAGTTCCT-----TTTTGTAGATA TATCTACAAAAA-----AGGAACTTGGAG	12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24
DYS576	CATCTCTGAATA-----AAAAAGCCAAGA TCTTGGCTTTTT-----TATTCAGAGATG	11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22
DYS635	ATAGAATCTCTA-----TCACATTTTCTT* AAGAAAATGTGA-----TAGAGATTCTAT*	13, 16, 17, 18, 19, 20, 21, 21.3, 22, 23, 24, 25, 26, 27, 28, 29, 30
DYS643	AAACTACTGTGC-----CTTCTTTTAA TTAAAAAGAAAG-----GCACAGTAGTTT	7, 8, 9, 10, 11, 12, 13, 14, 15
GATA H4	TAGGTAGGTAGG-----ATGGATAGATTA TAATCTATCCAT-----CCTACCTACCTA	8, 9, 10, 11, 12, 13, 14, 15

Table 2. Y-chromosome STR loci detected by STRait Razor. Leading and trailing flanking sequences for both forward and reverse complement reads are listed. All flanking sequences are directly adjacent to the repeat region, except for those labeled with an asterisk (*), where the proximities were modified to allow for increased specificity or more efficient detection. Additional alleles can be added at the discretion of the user.

		Detection Method											
		Sample 1				Sample 2		Sample 3		Sample 4		Sample 5	
		CE	STRait Razor (TruSeq™ prep, GAllx™)	STRait Razor (HaloPlex™ prep, GAllx™)	STRait Razor (HaloPlex™ prep, MiSeq™)	CE	STRait Razor (HaloPlex™ prep, GAllx™)	CE	STRait Razor (TruSeq™ prep, MiSeq™)	CE	STRait Razor (TruSeq™ prep, MiSeq™)	CE	STRait Razor (TruSeq™ prep, MiSeq™)
S T R L o c a l i t y	CSFIPO	12	12 (332, 315)	12 (1135, 976)	12 (324, 322)	12	12 (2445, 4155)	10, 11	10 (89, 84), 11 (54, 75)	12	12 (304, 310)	10, 12	10 (152, 189), 12 (162, 145)
	D13S317	12	12 (104, 96)	12 (639, 0)	12 (217, 214)	8, 11	8 (2513, 4888), 11 (1995, 4377)	11, 12	11 (49, 61), 12 (50, 48)	12	12 (220, 209)	8, 11	8 (101, 107), 11 (86, 76)
	D16S539	10, 12	10 (86, 70), 12 (64, 43)	10 (886, 2896), 12 (686, 1546)	10 (181, 135), 12 (174, 138)	9, 12	9 (1733, 7394), 12 (1165, 3209)	8, 13	8 (62, 69), 13 (51, 56)	9, 12	9 (121, 110), 12 (82, 77)	11, 13	11 (117, 98), 13 (91, 73)
	D18S51	15, 21	15 (25, 29), 21 (16, 17)	15 (598, 3258), 21 (341, 541)	15 (123, 123), 21 (106, 105)	15, 16	15 (957, 5711), 16 (819, 4734)	14, 16	14 (20, 16), 16 (13, 17)	15, 18	15 (34, 22), 18 (17, 17)	12	12 (94, 79)
	D19S433	12, 14	12 (39, 19), 14 (23, 13)	12 (589, 2620), 14 (542, 1543)	12 (141, 138), 14 (126, 133)	13, 14	13 (1634, 6001), 14 (1423, 4149)	13, 14	13 (16, 5), 14 (10, 11)	13	13 (44, 33)	12, 15	12 (12, 19), 15 (28, 17)
	D21S11	29, 30	[-] , [-]	[-] , [-]	29 (70, 17), 30 (57, 14)	29, 31	[-] , [-]	29	29 (17, 9)	29	29 (38, 25)	28, 29	28 (16, 11), 29 (16, 11)
	D2S1338	18, 25	18 (19, 18), 25 (9, 0)	18 (80, 1), [-]	18 (54, 52), 25 (29, 25)	20, 25	20 (137, 1), [-]	18	18 (53, 53)	17, 24	17 (74, 67), 24 (35, 35)	20, 23	20 (70, 70), 23 (55, 37)
	D3S1358	15	15 (109, 110)	15 (449, 4079)	15 (439, 395)	15, 16	15 (1046, 4579), 16 (852, 3995)	16, 18	16 (32, 29), 18 (36, 29)	16, 18	16 (71, 74), 18 (38, 66)	14, 17	14 (99, 93), 17 (72, 84)
	D6S818	11	11 (76, 66)	[-]	11 (66, 67)	11	[-]	11, 12	11 (20, 23), 12 (30, 26)	11	11 (101, 86)	12, 13	12 (56, 62), 13 (45, 34)
	D7S820	9, 12	9 (1, 2), 12 (5, 4)	9 (0, 1754), 12 (0, 1650)	9 (70, 70), 12 (57, 57)	10	10 (0, 7733)	10, 11	10 (4, 3), 11 (3, 4)	9, 13	9 (3, 7), 13 (8, 8)	11	11 (17, 15)
	D8S1179	13	13 (98, 87)	13 (1722, 5068)	13 (335, 220)	13	13 (3941, 10527)	10, 12	10 (36, 38), 12 (24, 25)	12, 14	12 (52, 47), 14 (48, 49)	13, 16	13 (68, 78), 16 (58, 41)
	FGA	20, 21	20 (36, 25), 21 (26, 20)	20 (770, 3819), 21 (581, 3419)	20 (168, 129), 21 (146, 134)	19, 21	19 (1019, 5066), 21 (890, 4849)	23, 25	23 (12, 16), 25 (12, 13)	22, 24	22 (40, 30), 24 (18, 23)	20	20 (84, 71)
	TH01	9, 9.3	9 (160, 146), 9.3 (132, 178)	9 (3172, 5571), 9.3 (2893, 5493)	9 (260, 255), 9.3 (297, 297)	8, 9.3	8 (5701, 8471), 9.3 (4963, 8350)	7	7 (150, 162)	9.3	9.3 (287, 292)	9.3	9.3 (393, 390)
	TPOX	8, 9	8 (90, 96), 9 (99, 80)	8 (4832, 5208), 9 (4488, 4710)	8 (527, 479), 9 (475, 428)	11	11 (11043, 15943)	8, 11	8 (35, 34), 11 (32, 29)	8, 12	8 (113, 105), 12 (84, 86)	8	8 (176, 185)
	vWA	16, 17	16 (53, 37), 17 (39, 24)	16 (299, 0), 17 (213, 0)	16 (55, 55), 17 (36, 36)	15, 20	15 (669, 0), 20 (0, 3)	17	17 (49, 56)	14, 17	14 (60, 60), 17 (57, 52)	15, 17	15 (63, 65), 17 (56, 61)
	Penta D	10	10 (24, 35)	10 (388, 0)	10 (180, 0)	14, 15	14 (214, 0), [-]	9, 11	9 (11, 14), 11 (12, 10)	12, 14	12 (26, 23), 14 (25, 20)	9, 12	9 (23, 24), 12 (29, 30)
Total Alleles	Penta E	11, 12	11 (6, 7), 12 (9, 5)	11 (98, 121), 12 (123, 104)	11 (105, 87), 12 (107, 88)	10, 11	10 (290, 993), 11 (243, 191)	5, 7	5 (10, 7), 7 (6, 5)	11, 12	11 (9, 4), 12 (11, 6)	7, 14	7 (20, 19), 14 (8, 7)
	D10S1248	13, 14	13 (106, 94), 14 (69, 67)	13 (672, 4537), 14 (329, 3613)	13 (192, 199), 14 (159, 171)	13	13 (3545, 17065)	12, 13	12 (81, 73), 13 (71, 69)	13, 14	13 (134, 127), 14 (152, 134)	16	16 (211, 211)
	D12S391	15, 17	15 (50, 50), 17 (35, 31)	15 (1464, 548), 17 (1101, 447)	15 (167, 117), 17 (162, 106)	21	21 (2123, 0)	15, 24	15 (37, 31), 24 (23, 19)	17, 21	17 (90, 82), 21 (84, 79)	18, 20	18 (68, 73), 20 (58, 63)
	D1S1656	16, 18.3	16 (63, 47), 18.3 (37, 52)	16 (499, 0), 18.3 (468, 0)	16 (59, 75), 18.3 (78, 41)	11, 12	11 (2262, 2848), 12 (1872, 2284)	16.3, 18.3	16.3 (33, 24), 18.3 (25, 29)	15, 17.3	15 (79, 75), 17.3 (67, 52)	12, 16	12 (98, 83), 16 (76, 67)
	D22S1045	16, 17	16 (10, 16), 17 (7, 16)	16 (901, 2637), 17 (718, 1639)	16 (107, 105), 17 (108, 104)	15, 16	15 (2042, 5133), 16 (1415, 3535)	15, 16	15 (15, 14), 16 (13, 15)	16	16 (35, 38)	11, 16	11 (94, 94), 16 (40, 40)
	D2S441	11, 15	11 (72, 49), 15 (32, 37)	11 (291, 3863), 15 (0, 3685)	11 (124, 145), 15 (113, 167)	11.3, 14	11.3 (615, 7807), 14 (0, 7024)	11, 14	11 (74, 71), 14 (68, 64)	10, 11.3	10 (106, 100), 11.3 (149, 137)	10, 11	10 (131, 125), 11 (130, 138)
Total Alleles		38	36	34	38	37	32	40	40	37	37	38	38

Table 3. Comparison of CE allele calls and STRait Razor results – Autosomal STRs. Alleles detected by both CE and STRait Razor analysis of SGS data are shown in bold in the columns for each sample. The numbers of reads in which an allele was detected by STRait Razor are listed in parentheses next to the respective allele. The first number in parentheses represents the abundance of the allele in Read 1 of the paired-end sequencing run, while the second number represents the abundance of the allele in Read 2. Alleles not detected by STRait Razor due to lack of relevant sequence data are denoted by “[-]”.

		Detection Method											
		Sample 1				Sample 2		Sample 3		Sample 4		Sample 5	
		CE	STRait Razor (TruSeq™ prep, GAIlix™)	STRait Razor (HaloPlex™ prep, GAIlix™)	STRait Razor (HaloPlex™ prep, MiSeq™)	CE	STRait Razor (HaloPlex™ prep, GAIlix™)	CE	STRait Razor (TruSeq™ prep, MiSeq™)	CE	STRait Razor (TruSeq™ prep, MiSeq™)	CE	STRait Razor (TruSeq™ prep, MiSeq™)
S T R L o c a l i z e d	DYS19	14	14 (7, 11)	14 (231, 1221)	14 (61, 62)	15	15 (6, 2341)	14	14 (8, 18)	16	16 (13, 13)	14	14 (14, 15)
	DYS385	11, 13	11 (12, 12), 13 (12, 11)	[-] , [-]	[-] , [-]	11, 14	[-] , [-]	11, 14	11 (5, 3), 14 (4, 5)	11, 14	11 (22, 29), 14 (8, 19)	11, 14	11 (12, 13), 14 (24, 16)
	DYS389I	13	13 (102, 99)	13 (423, 1)	13 (26, 0)	13	13 (1605, 0)	13	13 (54, 47)	13	13 (137, 137)	13	13 (120, 103)
	DYS389II	29	[-]	[-]	29 (23, 0)	30	[-]	29	29 (9, 3)	28	28 (25, 20)	29	29 (13, 17)
	DYS390	24	24 (14, 16)	24 (115, 3495)	24 (115, 126)	25	25 (259, 3386)	24	24 (16, 19)	23	23 (37, 43)	23	23 (40, 37)
	DYS391	10	10 (175, 167)	10 (952, 39)	10 (99, 24)	10	10 (3179, 1431)	10	10 (73, 80)	10	10 (180, 182)	12	12 (166, 165)
	DYS392	13	13 (3, 8)	13 (885, 1965)	13 (82, 78)	11	11 (1466, 2850)	13	13 (7, 8)	13	13 (7, 8)	13	13 (11, 10)
	DYS393	13	13 (9, 2)	13 (0, 360)	13 (14, 13)	14	14 (0, 1023)	13	13 (2, 7)	13	13 (2, 3)	13	13 (10, 10)
	DYS437	15	15 (85, 70)	15 (0, 4020)	15 (247, 238)	14	14 (0, 11064)	15	15 (77, 77)	15	15 (148, 133)	14	14 (141, 146)
	DYS438	12	12 (42, 36)	12 (324, 285)	12 (62, 32)	11	11 (884, 871)	12	12 (48, 49)	12	12 (79, 68)	12	12 (96, 93)
	DYS439	12	[-]	12 (428, 2296)	12 (134, 78)	10	10 (1789, 6072)	12	12 (2, 0)	11	11 (2, 2)	13	13 (3, 1)
	DYS448	19	[-]	[-]	19 (17, 5)	19	[-]	19	19 (11, 3)	19	19 (21, 16)	18	18 (12, 11)
	DYS456	15	15 (10, 13)	15 (523, 1402)	15 (80, 54)	14	14 (2723, 6296)	15	15 (13, 10)	16	16 (11, 9)	15	15 (22, 21)
	DYS458	17	17 (10, 6)	17 (56, 258)	17 (31, 21)	15	15 (152, 1300)	19	19 (11, 9)	17	17 (13, 12)	17	17 (17, 24)
	DYS481	22	22 (19, 21)	22 (227, 1943)	22 (150, 101)	23	23 (663, 3830)	25	25 (20, 18)	23	23 (30, 43)	22	22 (50, 45)
	DYS533	12	12 (26, 36)	12 (184, 0)	12 (30, 10)	12	12 (511, 0)	12	12 (37, 37)	12	12 (34, 37)	14	14 (41, 27)
	DYS549	13	13 (44, 49)	13 (743, 0)	13 (151, 101)	12	12 (2649, 0)	13	13 (39, 38)	13	13 (46, 58)	13	13 (60, 62)
	DYS570	17	17 (44, 51)	17 (646, 0)	17 (76, 28)	20	20 (777, 0)	18	18 (64, 69)	17	17 (73, 66)	17	17 (145, 134)
	DYS576	19	19 (45, 43)	19 (0, 341)	19 (3, 11)	17	17 (0, 512)	19	19 (34, 22)	18	18 (74, 74)	18	18 (65, 57)
	DYS635	24	24 (10, 5)	24 (220, 71)	24 (18, 20)	25	25 (774, 700)	23	23 (18, 15)	24	24 (33, 29)	23	23 (30, 27)
	DYS643	10	10 (43, 25)	10 (2392, 989)	10 (221, 111)	10	10 (3314, 1407)	11	11 (10, 15)	11	11 (34, 30)	10	10 (46, 43)
	GATA H4	12	12 (23, 21)	12 (290, 2196)	12 (97, 93)	12	12 (511, 3795)	11	11 (21, 27)	11	11 (34, 47)	11	11 (33, 40)
Total Alleles		23	20	19	21	23	19	23	23	23	23	23	23

Table 4. Comparison of CE allele calls and STRait Razor results – Y-Chromosome STRs. Alleles detected by both CE and STRait Razor analysis of SGS

data are shown in bold in the columns for each sample. The numbers of reads in which an allele was detected by STRait Razor are listed in parentheses next to the respective allele. The first number in parentheses represents the abundance of the allele in Read 1 of the paired-end sequencing run, while the second number represents the abundance of the allele in Read 2. Alleles not detected by STRait Razor due to lack of relevant sequence data are denoted by “[-]”.

DISCUSSION

The results of this study show the efficiency and accuracy of STR allele detection with STRait Razor. They also reveal the close relationship between the software, the library preparation chemistries, and the sequencing platforms used to produce the sequence information.

Read length is arguably the most important factor (followed by coverage) that impacts the allelic detection capability of STRait Razor. The HaloPlex™ chemistry relies on enzymatic cleavage [24], which creates fragments with consistent start and end points. Depending on the length of the allele in question and the position of the repeat region within the resulting fragment(s), it is possible for sequencing reads to be produced that only partially span the repeat region and its associated flanking sequences (Fig. 2a). Since STRait Razor requires reads that contain all of this information, alleles may go undetected without complete repeat region traversal. An example of this phenomenon is locus D2S1338 in Sample 1 (HaloPlex™ preparation, GAIIX™ sequencing). For this locus, the “18” allele was called, but the “25” allele was too long to be covered completely by the sequencing read, and thus went undetected. When this same sample subsequently was sequenced on the MiSeq™ platform using a longer read length [17,18], the “25” allele was detected. In addition, if a repeat region is situated toward the beginning of a HaloPlex™ fragment, the allele is likely to be detected in one direction of a paired-end analysis. However, when the reads are sequenced from the opposite direction, the repeat region is oriented toward the end of the read and may not be completely encompassed (Fig. 2b). This situation can be seen in loci such as D7S820 and vWA in Sample 1, where the alleles are detected only in one set of paired-end reads and not the other. Some library preparation redesign may overcome the truncated repeat

region reads.

The TruSeq™ chemistry [25] is less prone to these issues because the random fragmentation of DNA allows for a much more diverse positioning of repeat regions within the resulting fragments (Fig. 2c). Therefore, there is a higher likelihood of at least some reads encompassing the entire repeat region. This design explains why the majority of the alleles that were not detected by STRait Razor for Sample 1 following HaloPlex™ preparation and GAIIX™ sequencing were detected normally by the program for this sample following TruSeq™ preparation on the same instrument. Despite this advantage, the non-enzymatic random fragmentation employed by the TruSeq™ chemistry may result in lower read counts for some alleles in comparison with HaloPlex™, due to the fewer resulting fragments containing the complete repeat region. In some cases, the random fragmentation method simply may not create any fragments that contain the repeat region of interest. This limitation may explain the undetected alleles in Sample 1 at loci DYS439 and DYS448 following TruSeq™ preparation and GAIIX™ sequencing. Coverage depth differences in the results of this study, however, also may be explained by the fact that the regions targeted by the TruSeq™ kit for these trials were, by design, approximately 100 times larger than those targeted by the HaloPlex™ kit. Finally, it should be noted that in cases where the larger allele of a heterozygous pair at a given locus is not detected due to read length issues, a detected stutter allele may give the impression of a true heterozygous allele. Such issues may eventually be overcome by a combination of secondary statistical allelotyping software that takes into account the allele coverage ratios, and customization of the allelic information modules utilized by STRait Razor for the specific chemistry and instrumentation used for sequencing.

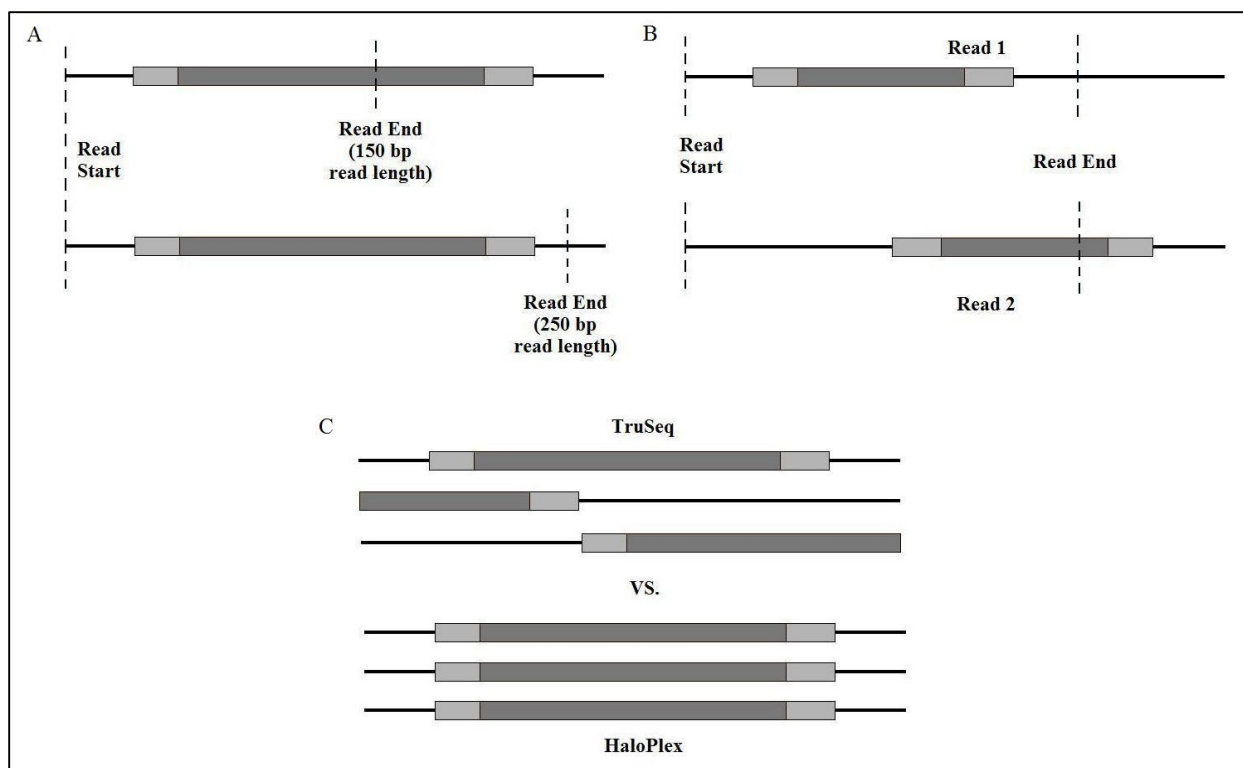


Fig. 2. Read length-related issues. In this figure, the dark gray bars represent the repeat region, while the light gray bars represent the flanking regions. The bold black lines represent surrounding sequence data. (A) Two identical fragments, such as those produced by the HaloPlex™ chemistry, are sequenced with 2 different read lengths. (B) A repeat region situated within a fragment is sequenced toward the beginning of the read in Read 1 of a paired-end sequencing run, but is sequenced toward the end of the read in Read 2 of the paired-end run. (C) A comparison of repeat region location consistency between the TruSeq™ and HaloPlex™ library preparation chemistries. This illustration is an example of a small subset of sequence reads, and does not reflect the actual proportions of reads produced by each library preparation method.

It should be noted that sequence variations (e.g., insertions or deletions) that reside outside of the flanking regions used by STRait Razor but within the primer-binding sites used by commercial STR/ CE kits may result in discordant results between SGS and CE analyses. While this discordance was not observed during the course of this study, most likely due to limited

sampling, it is not unique to SGS. Such discrepancies can occur due to different primer-binding site locations between various constructs of CE-based kits, as well.

The small portions of repeat sequence that are exactly matched to reads during the filtering portion of the algorithm were deliberately kept short (3 simple repeat units or less). This exact matching step is important, as it removes irrelevant sequence data that may have been captured due to chance flanking sequence homology. However, the use of exact matching can potentially remove repeat regions with intra-repeat variation. Thus, sequences of repeats that are shorter than the smallest known allele at each locus are used, so that they may align at various points throughout the captured reads and still allow reads with intra-repeat variation to be retained. The Penta D locus, however, requires special attention, because its smallest known allele is “1.1” [26], albeit an extremely rare allele. The repeat sequence used by STRait Razor for exact match filtering at this locus is only one repeat unit long, since using a longer sequence would not allow for intra-repeat variability detection in the smaller alleles. The use of such a small portion of repeat sequence results in less thorough filtering of the reads, but this does not appear to affect the concordance of the resulting allelic data.

While the results of this study are encouraging, it should be noted that all software has its strengths and limitations. STRait Razor is simply an alternative method of approaching an increasingly important facet of SGS data analysis. The software is being offered freely so that those working with SGS and STR detection have an additional tool to facilitate their analyses. STRait Razor is a fairly straightforward program, and the authors encourage enhancements to this software. With this phase of STRait Razor, probabilities of allele calls are not provided, nor does

the program perform any stutter filtering or attempt to differentiate between heterozygosity and homozygosity. The software only reports the allele calls and related coverage at each locus; the analyst makes the final decision in an informed manner. A separate program is currently in development that reads the colon-delimited text file(s) output by STRait Razor and detects the two allele calls with the highest count values (if any), comparing them based on a user-defined “abundance ratio” to make homozygous and heterozygous allelotype calls with confidence. Presently, STRait Razor does not filter reads based on FASTQ quality information. However, the flanking region detection and small-repeat match verification performed by STRait Razor can be considered inherent filtering steps. Quality-based filtering may be incorporated into future versions of the program. In its current format, STRait Razor is designed to detect all the known alleles at each of the tested loci, according to the allelic information listed in STRBase and the Y-Chromosome Haplotype Reference Database (YHRD) [27,28]. These alleles are defined in the modules used by the software and may be modified by the user to include other rare or undocumented variants. In the future, the software may be altered to allow for the intuitive calling of alleles based on repeat length alone, without the need for allelic definitions.

CONCLUSION

In its current state of development, STRait Razor offers forensic DNA analysts a simple and effective means of detecting STR alleles in SGS data. Its ability to do so is limited primarily by the capabilities of the library preparation chemistries and sequencing platforms used to generate the raw data. The software has been shown to properly and efficiently analyze data resulting from a variety of commercially available library preparation chemistries and sequencers. The results of

this study also indicate that the software is capable of identifying alleles in SGS sequence data with 100% concordance. STRait Razor provides multiple options for customization, and new loci can be added for detection at the discretion of the analyst. The program is freely available (see Supplementary Materials), and the scientific community is encouraged to make improvements to STRait Razor to increase its utility as a forensic tool.

Supplementary data associated with this article can be found, in the online version, at <http://dx.doi.org/10.1016/j.fsigen.2013.04.005>.

ACKNOWLEDGMENTS

This project was supported in part by Award No. 2012-DN-BX-K033, awarded by the National Institute of Justice, Office of Justice Programs, U.S. Department of Justice. The opinions, findings, and conclusions or recommendations expressed in this publication are those of the author(s) and do not necessarily reflect those of the U.S. Department of Justice.

The authors would like to thank Illumina, Inc. and Agilent Technologies, Inc. for providing their technical expertise and contributing a portion of the sample preparation and sequencing chemistries used in this study.

REFERENCES

- [1] A. Edwards, A. Civitello, H.A. Hammond, C.T. Caskey, DNA typing and genetic mapping with trimeric and tetrameric tandem repeats, *Am. J. Hum. Genet.* 49 (1991) 746–756.
- [2] A. Edwards, H.A. Hammond, L. Jin, C.T. Caskey, R. Chakraborty, Genetic variation at five trimeric and tetrameric repeat loci in four human population groups, *Genomics* 12 (1992) 241–253.
- [3] H. Ellegren, Microsatellites: simple sequences with complex evolution, *Nat. Rev. Genet.* 5 (2004) 435–445.
- [4] K. Lazaruk, P.S. Walsh, F. Oaks, D. Gilbert, B.B. Rosenblum, S. Menchen, D. Scheibler, H.M. Wenz, C. Holt, J. Wallin, Genotyping of forensic short tandem repeat (STR) systems based on sizing precision in a capillary electrophoresis instrument, *Electrophoresis* 19 (1998) 86–93.
- [5] P. Gill, D.J. Werrett, B. Budowle, R. Guerrieri, An assessment of whether SNPs will replace STRs in national DNA databases – joint considerations of the DNA working group of the European Network of Forensic Science Institutes (ENFSI) and the Scientific Working Group on DNA Analysis Methods (SWGDM), *Sci. Justice.* 44 (2004) 51–53.
- [6] P.J. Collins, L.K. Hennessy, C.S. Leibelt, R.K. Roby, D.J. Reeder, P.A. Foxall, Developmental validation of a single-tube amplification of the 13 CODIS STR loci, D2S1338, D19S433, and amelogenin: the AmpFØSTR® Identifiler® PCR Amplification Kit, *J. Forensic Sci.* 49 (2004) 1265–1277.
- [7] K. Oostdik, J. French, D. Yet, B. Smalling, C. Nolde, P.M. Vallone, E.L. Butts, C.R. Hill, M.C. Kline, T. Rinta, A.M. Gerow, S.R. Allen, C.K. Huber, J. Teske, B. Krenke, M. Ensenberger, P. Fulmer, C. Sprecher, Developmental validation of the PowerPlex® 18D system, a rapid STR multiplex for analysis of reference samples, *Forensic Sci. Int. Genet.* 7 (2013) 129–135.
- [8] M. Gymrek, D. Golan, S. Rosset, Y. Erlich, lobSTR: a short tandem repeat profiler for personal genomes, *Genome Res.* 22 (2012) 1154–1162.
- [9] D.M. Bornman, M.E. Hester, J.M. Schuetter, M.D. Kasoji, A. Minard-Smith, C.A. Barden, S.C. Nelson, G.D. Godbold, C.H. Baker, B. Yang, J.E. Walther, I.E. Tornes, P.S. Yan, B. Rodriguez, R. Bundschuh, M.L. Dickens, B.A. Young, S.A. Faith, Short-read, high-throughput sequencing technology for STR genotyping, *Biotechniques – Rapid Dispatches* (2012) 1–6.
- [10] S.L. Fordyce, M.C. Ávila-Arcos, E. Rockenbauer, C. Børsting, R. Frank-Hansen, F.T. Petersen, E. Willerslev, A.J. Hansen, N. Morling, M.T. Gilbert, High-throughput sequencing of core STR loci for forensic genetic investigations using the Roche Genome Sequencer FLX platform, *Biotechniques* 51 (2011) 127–133.

- [11] M.M. Holland, M.R. McQuillan, K.A. O'Hanlon, Second generation sequencing allows for mtDNA mixture deconvolution and high resolution detection of heteroplasmy, *Croat. Med. J.* 52 (2011) 299–313.
- [12] R. Nielsen, J.S. Paul, A. Albrechtsen, Y.S. Song, Genotype and SNP calling from next-generation sequencing data, *Nat. Rev. Genet.* 12 (2011) 443–451.
- [13] D.W. Craig, J.V. Pearson, S. Szelinger, A. Sekar, M. Redman, J.J. Corneveaux, T.L. Pawlowski, T. Laub, G. Nunn, D.A. Stephan, N. Homer, M.J. Huentelman, Identification of genetic variants using bar-coded multiplexed sequencing, *Nat. Methods* 5 (2008) 887–893.
- [14] D.C. Koboldt, K. Chen, T. Wylie, D.E. Larson, M.D. McLellan, E.R. Mardis, G.M. Weinstock, R.K. Wilson, L. Ding, VarScan: variant detection in massively parallel sequencing of individual and pooled samples, *Bioinformatics* 25 (2009) 2283–2285.
- [15] A. McKenna, M. Hanna, E. Banks, A. Sivachenko, K. Cibulskis, A. Kernytsky, K. Garimella, D. Altshuler, S. Gabriel, M. Daly, M.A. DePristo, The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data, *Genome Res.* 20 (2010) 1297–1303.
- [16] E.C. Berglund, A. Kiialainen, A.C. Syvänen, Next-generation sequencing technologies and applications for human genetic history and forensics, *Investig. Genet.* 2 (2011) 1–15.
- [17] Illumina® GAIIX™ Specifications: http://www.illumina.com/documents/products/datasheets/datasheet_genome_analyzerIIX.pdf.
- [18] Illumina® MiSeq™ Specifications: http://www.illumina.com/documents/products/datasheets/datasheet_miseq.pdf.
- [19] B. Langmead, C. Trapnell, M. Pop, S.L. Salzberg, Ultrafast and memory-efficient alignment of short DNA sequences to the human genome, *Genome Biol.* 10 (2009) R25.
- [20] CASAVA v.1.8.2: http://support.illumina.com/downloads/casava_182.ilmn.
- [21] MiSeq Reporter: http://support.illumina.com/sequencing/sequencing_software/miseq_reporter/downloads.ilmn.
- [22] AGREP: <http://laurikari.net/tre/download/>.
- [23] PPSS: <http://code.google.com/p/ppss/>.
- [24] Agilent Technologies® HaloPlex™ Specifications: <http://www.genomics.agilent.com/GenericB.aspx?pagetype=Custom&subpagetype=Custom&pageid=3081>.
- [25] Illumina TruSeq Specifications: http://www.illumina.com/Documents/%5Cproducts%5Cdatasheets%5Cdatasheet_truseq_custom_enrichment_kit.pdf.

- [26] Penta D Facts Sheet (STRBase): http://www.cstl.nist.gov/strbase/str_Penta_D.htm.
- [27] STRBase: http://www.cstl.nist.gov/strbase/str_fact.htm.
- [28] Y-Chromosome Haplotype Reference Database: <http://www.yhrd.org/Research/Loci>.

CHAPTER 2

STRait Razor v2.0: the Improved STR Allele Identification Tool – Razor

Published in Forensic Science International: Genetics
2015, 14:182-186

David H. Warshauer
Jonathan L. King
Bruce Budowle

ABSTRACT

STRait Razor (the STR Allele Identification Tool – Razor) was developed as a bioinformatic software tool to detect short tandem repeat (STR) alleles in massively parallel sequencing (MPS) raw data. The method of detection used by STRait Razor allows it to make reliable allele calls for all STR types in a manner that is similar to that of capillary electrophoresis. STRait Razor v2.0 incorporates several new features and improvements upon the original software, such as a larger default locus configuration file that increases the number of detectable loci (now including X-chromosome STRs and Amelogenin), an enhanced custom locus list generator, a novel output sorting method that highlights unique sequences for intra-repeat variation detection, and a genotyping tool that emulates traditional electropherogram data. Users also now have the option to choose whether the program detects autosomal, X-chromosome, Y-chromosome, or all STRs. Concordance testing was performed, and allele calls produced by STRait Razor v2.0 were completely consistent with those made by the original software.

KEYWORDS: STRait Razor · Massively parallel sequencing · STR · Bioinformatics · Software

INTRODUCTION

STRait Razor (the STR Allele Identification Tool – Razor) is a Perl-based bioinformatic software package for the Linux operating system that was created to provide an effective and reliable means of detecting short tandem repeat (STR) markers in massively parallel sequencing (MPS) data. While the current assortment of MPS platforms and methodologies are capable of sequencing STRs [1–4], there are very few options available for variant analysis of the raw data [1,5]. STRait Razor accurately analyzes STR sequence data from FASTQ files while avoiding the drawbacks associated with software tools that are unable to adequately detect complex repeat motifs or intra-repeat variation [1]. Although it was intended for forensic DNA typing, the software can be used for a wide variety of genetic research purposes. STRait Razor employs a simple yet effective method to analyze STRs. First, alleles are detected via matching of the leading and trailing flanking region(s) surrounding the repeat(s). This matching process allows for user-defined stringency parameters and results in the cleaving of all extraneous sequence data from around an STR, leaving only the repeat units. A final filtering step removes non-STR sequencing reads that may have been captured due to flanking sequence homology. Alleles are called by comparing the length of the repeat unit region with known allele lengths, and the total counts of each allele at each locus are ordered and stored in an output file for further analysis and comparison with other samples. STR analysis is performed in parallel, using a separate processor core for the detection of each locus. This straightforward process results in allele calls that are analogous to calls produced by traditional capillary electrophoresis (CE) methods, since the allele coverage counts are analogous to RFU values. In addition to size of the allele, registered as number of repeats, allelic sequence data provide an added benefit of highlighting nucleotide variation within the

repeat units. Since its initial release, STRait Razor has been employed by a number of laboratories with positive results. In-house needs and resulting feedback were considered strongly to develop a number of improvements to the original software. These new features include an expanded default set of detectable STR loci (autosomal, X, and Y markers), an enhanced custom locus list configuration tool, a novel output sorting method that highlights unique sequences for each allele, and a genotyping tool that emulates traditional electropherogram data. With these improvements, STRait Razor v2.0 offers users a much wider, more flexible range of analysis options and greater ease of use.

MATERIALS AND METHODS

New features

STRAit Razor v2.0 includes an expanded locus configuration file which it can use to detect a wider range of forensically-relevant STR markers. Previously, the default locus definition file included 44 STR loci (22 autosomal STRs and 22 Y-chromosome STRs). An additional 42 markers have been added, for a total of 86 markers, which include: 9 new autosomal STRs (D14S1434, D17S1301, D1S1627, D2S1776, D4S2408, D5S2500, D6S1017, D6S474, and SE33), 26 new X-chromosome STRs (DXS10011, DXS10074, DXS101, DXS10101, DXS10134, DXS10135, DXS6789, DXS6795, DXS6800, DXS6801, DXS6807, DXS6809, DXS6854, DXS7132, DXS7133, DXS7423, DXS7424, DXS8377, DXS8378, DXS981, DXS9895, DXS9902, GATA165B12, GATA172D05, GATA31E08, and HPRTB), six new Y-chromosome STRs (DYS449, DYS460, DYS505, DYS518, DYS522, and DYS612), and Amelogenin. The allelic

information contained in the locus configuration file was compiled using data from a variety of online databases, such as STRbase [6], ChrX-STR.org 2.0 [7], NCBI [8], and the Sorenson Molecular Genealogy Foundation database [9]. Users have the option of choosing the marker type to analyze with STRait Razor v2.0 by using the “-typeselection” argument (AUTO, X, Y, or ALL). This feature allows the analysis to be tailored to the specific goals of the testing and, depending on which option is selected, can reduce the time required for analysis. As with the initial version of STRait Razor, custom locus configuration files can be created for the program using the included Microsoft Excel workbook. Therefore, analysis can be performed on STR loci that are not included in the default set.

STRAit Razor v2.0 includes an enhanced locus configuration workbook for the creation of custom allelic definition files. The workbook allows for the generation of a locus configuration file containing up to 10 STR loci, although the workbook can be modified to include more. Alternatively, multiple configuration files may be generated and concatenated to produce larger, more comprehensive files. The locus configuration workbook is designed to convert locus information entered by the user for any STR locus type (autosomal, X-chromosome, or Y-chromosome) to the proper format required by STRait Razor for analysis. The workbook has been redesigned for ease of use, with features such as automatic generation of reverse complement leading and trailing flanking data based on user input, and auto-conversion of lowercase entries to the proper uppercase format.

STRAit Razor is designed to yield STR allele calls for each locus, as well as sequence data for all alleles so that intra-repeat variation can be detected. In its initial release, STRait Razor

output sequence data for each allele in a locus-specific file, sorted by repeat region length. While useful, these data were unwieldy to interpret because sequence data from each individual read were appended to a sequence file when captured, resulting in a large amount of redundant sequence information. STRait Razor v2.0 simplifies the sequence output so that only unique sequences for each allele are displayed, and the results are sorted based on the total read count for each sequence (Fig. 1). This output results in a clear and concise sequence file for each locus that can be interpreted quickly for intra-repeat nucleotide variation.

Original Sequence Data			New Sequence Data				
LOCUSA:7	28 bases	AGATAGATAGATAGATAGATAGAT		LOCUSA:7	28 bases	AGATAGATAGATAGATAGATAGAT	4
* LOCUSA:7	28 bases	AGATAGATAGGTAGATAGATAGATAGAT		* LOCUSA:7	28 bases	AGATAGATAGGTAGATAGATAGATAGAT	3
** LOCUSA:7	28 bases	AGATAGATAGATAGATAGATACATAGAT		** LOCUSA:7	28 bases	AGATAGATAGATAGATAGATACATAGAT	2
LOCUSA:7	28 bases	AGATAGATAGATAGATAGATAGATAGAT					
* LOCUSA:7	28 bases	AGATAGATAGGTAGATAGATAGATAGAT					
* LOCUSA:7	28 bases	AGATAGATAGGTAGATAGATAGATAGAT					
LOCUSA:7	28 bases	AGATAGATAGATAGATAGATAGATAGAT					
LOCUSA:7	28 bases	AGATAGATAGATAGATAGATAGATAGAT					
** LOCUSA:7	28 bases	AGATAGATAGATAGATAGATACATAGAT					

Fig. 1. Sequence data sorting. The original sequence output from STRait Razor, using an example locus named “LOCUSA” (AGAT repeat unit), is shown on the left. Sequences were appended as they were captured. Sequences denoted with (*) contain an A/G SNP, while sequences denoted with (**) contain a G/C SNP. Sequence output from STRait Razor v2.0 is shown on the right. With this update, unique sequences were identified, counted, compiled, and finally sorted based on the total read count.

The first version of STRait Razor enabled users to generate a substantial amount of information on both the alleles and underlying sequence variants. While this updated version of STRait Razor is considerably more efficient, the tremendous amount of data generated requires a toolset for subsequent analysis. To further facilitate data analysis, a set of Excel-based workbooks have been developed.

The first workbook, RazorGenotyper, converts the output files “RawSTRcallsR1” and “RawSTRcallsR2” into final genotypes. Users can set thresholds for coverage and sister allele balance to ensure that accurate genotypes are generated. Data from multiple samples can be exported and compiled via embedded macros. These exported data then can be further visualized using the second workbook.

STRait Razor Histogram Generator separates the output data of RazorGenotyper into an “allele table” of all loci. These data then are displayed as histograms showing all read variants observed (e.g., alleles, stutter, and PCR artifacts). These charts are parsed into “Autosomal,” “Y,” and “X” STR tabs. The autosomal tab also contains Amelogenin and is divided into “Core” loci (i.e., loci contained in either PowerPlex Fusion (Promega Corp., Madison, WI, USA) or GlobalFiler (ThermoFisher, San Francisco, CA, USA)) and “Additional Loci” (i.e., autosomal loci not found in either kit). Macros included, but not active, allow a user to uniformly change the axes of all charts to visualize locus-to-locus balance.

The final workbook included in the toolset, STRait Razor_SNP ID Tool, converts the *LOCUS.SEQUENCES* files into a table showing the top 20 sequence variants at each locus. First, the user must transfer the *LOCUS.SEQUENCES* file from each locus of interest into a single folder. Next, the user can run the included ‘SeqCompile.pl’ script to combine all loci into a single file. The data from this file then are pasted into the STRait Razor_SNP ID Tool. After which, data are displayed by locus showing the most relevant sequence variants. These data for all loci can be exported and compiled via embedded macros for ease of use.

Concordance testing

The accuracy and reliability of STRait Razor have been reported [1]. Therefore, concordance testing was performed to verify that STRait Razor v2.0, with its updates, provides the same allele calls as it did in its first release. The same seven sequence datasets used in the initial testing phase were re-analyzed using STRait Razor v2.0. These datasets are summarized in Table 1.

Dataset	Sample	Library Preparation Method	Sequencing Platform
1	A	Illumina® TruSeq™ Custom Enrichment	Illumina® GAIIX™
2	A	Agilent Technologies® HaloPlex™	Illumina® GAIIX™
3	A	Agilent Technologies® HaloPlex™	Illumina® MiSeq™
4	B	Agilent Technologies® HaloPlex™	Illumina® GAIIX™
5	C	Illumina® TruSeq™ Custom Enrichment	Illumina® MiSeq™
6	D	Illumina® TruSeq™ Custom Enrichment	Illumina® MiSeq™
7	E	Illumina® TruSeq™ Custom Enrichment	Illumina® MiSeq™

Table 1. Datasets used for concordance testing. The seven datasets used consisted of sequence data from 5 samples, obtained through library preparation using the TruSeq™ (Illumina, Inc., San Diego, CA) and HaloPlex™ (Agilent Technologies, Inc., Santa Clara, CA) kits and sequencing using the MiSeq™ (Illumina, Inc.) and GAIIX™ (Illumina, Inc.) platforms.

The sample preparation methods and sequencing techniques used in obtaining these datasets were described previously [1]. Allele calls made by the updated software were compared to those made by the initial version of the software, and new allele calls resulting from use of the larger default locus configuration file were noted. The time required for analysis with the wider range of detectable loci also was determined. Additionally, genotyping and histogram generation were performed to validate the effectiveness of these tools.

In addition, the allele detection capability of STRait Razor v2.0 was evaluated with MiSeq™-generated FASTQ files from a set of 12 samples (6 male and 6 female) that had previously undergone library preparation using an experimental in-house Nextera™ Rapid Capture (Illumina, Inc.) assay designed to specifically target the loci included in STRait Razor v2.0's locus configuration file. These datasets were analyzed using both STRait Razor v2.0 and the original version of the software, and the genotype results were compared.

RESULTS

STRAit Razor v2.0 yielded identical allele calls from all 7 original sequence datasets with regard to the loci previously analyzed. The read counts generated for each allele detected by STRait Razor v2.0 were 100% concordant with those indicated by the original version of the software. The testing process also demonstrated the wider range of locus detection afforded by the new expanded locus configuration file. Across these 7 datasets, STRait Razor v2.0 detected a range of 25–41 new alleles at between 19 and 35 new autosomal, X-chromosome, and Y-chromosome loci, as well as Amelogenin (Table 2).

Dataset	Additional Alleles	Additional Loci			
		Autosomal	X	Y	Amelogenin
1	26	7	9	3	+
2	25	6	10	3	-
3	34	7	17	4	-
4	27	7	10	3	-
5	35	8	16	4	+
6	40	8	21	5	+
7	41	8	21	4	+

Table 2. Additional alleles and loci detected in the original 7 datasets. New loci identified by STRait Razor v2.0

have been categorized by their respective types. (+) Indicates that amelogenin was detected in a particular dataset, while (-) indicates that amelogenin was not detected.

The results for the additional 12 datasets showed a similar pattern of enhanced allele detection with STRait Razor v2.0. The original software detected between 60 and 64 alleles at 44 loci in male samples, while between 37 and 41 alleles were identified at 22 loci in female samples, given the absence of Y-chromosome STRs in these samples (Table 3). STRait Razor v2.0, however, increased the number of alleles detected to a range of 105–111 at between 82 and 84 loci in male samples. Female samples analyzed with the updated software yielded 92–97 alleles at 56 loci. Amelogenin was detected in all 12 additional samples. Comparison of the allele calls made by STRait Razor v2.0 and the original version of STRait Razor for these samples showed that they were 100% concordant, as was the case for the initial group of seven datasets.

As noted previously, allele detection is dependent on a number of factors, including sequence read length and library preparation chemistry [1]. Lack of detection of alleles at loci included in the new locus configuration file was due to these same issues and is not indicative of improper software function. Time required for analysis is directly related to the number of loci that

are included in the locus configuration file. Thus, the new larger configuration file does increase STRait Razor v2.0's analysis time. In this study, the time required for analysis of all 86 markers for dual 400MB MiSeq™-generated FASTQ files on a 16-core server was approximately 29 min. When only Y-chromosome or X-chromosome STRs were analyzed, the analysis time dropped to approximately 9 min for each. The time required for analysis of only autosomal STRs was approximately 11 min. Shorter custom configuration files can be used to reduce analysis time, and the time required will vary depending on the specifics of each application and the computing platform utilized.

		Alleles Detected	Loci Detected			
			Autosomal	X	Y	Amelogenin
S a m p l e	1	41/96	22/30	0/25	*/*	-/+
	2	64/111	22/30	0/25	22/27	-/+
	3	37/95	22/30	0/25	*/*	-/+
	4	61/108	22/30	0/25	22/27	-/+
	5	61/107	22/31	0/25	22/27	-/+
	6	37/93	22/30	0/25	*/*	-/+
	7	63/109	22/31	0/25	22/27	-/+
	8	39/97	22/30	0/25	*/*	-/+
	9	64/110	22/30	0/24	22/27	-/+
	10	37/97	22/30	0/25	*/*	-/+
	11	37/92	22/30	0/25	*/*	-/+
	12	60/105	22/30	0/24	22/27	-/+

Table 3. Allele and locus detection comparison for the additional 12 datasets. The alleles and loci identified by the original version of STRait Razor are listed first, followed by those detected by STRait Razor v2.0. The loci are grouped based on their respective chromosomal type.

Genotypes were displayed, along with allele read counts, in a manner that was clear and easy to interpret. Histograms were generated for the alleles, stutter, and noise detected from these datasets that approximate electropherogram displays (Fig. 2). Given the similarity between these

histograms and traditional electropherograms, this option provides a simple way to visually inspect allele calls at each detected locus. In this manner, reads resulting from stutter or noise can be interpreted visually.

Finally, the sequence data for each detected allele that were output by STRait Razor v2.0 were investigated. The unique sorting process employed by the updated software allows for quick detection of intra-repeat nucleotide variation. For example, an examination of the sequence data output file for dataset 4 revealed that the homozygote “29” allele at locus D21S11 consisted of the variants “(TCTA)₄(TCTG)₆(TCTA)₃TA(TCTA)₃TCA(TCTA)₂TCCATA(TCTA)₁₁” and “(TCTA)₅(TCTG)₆(TCTA)₃TA(TCTA)₃TCA(TCTA)₂TCCATA(TCTA)₁₀”, at a ratio approaching 1:1 (Fig. 3). The latter “29” variant is not listed in STRbase. It should be noted that both variants are found in the “28” stutter allele, as well. Additional sequence variation such as this can provide increased discrimination power, especially in the case of DNA mixtures, where it can facilitate deconvolution of both unique and shared alleles. However, intra-repeat variants must be characterized and allele frequencies generated before they can be used to their full potential. Mixture deconvolution using STRait Razor v2.0 will be evaluated in future studies.

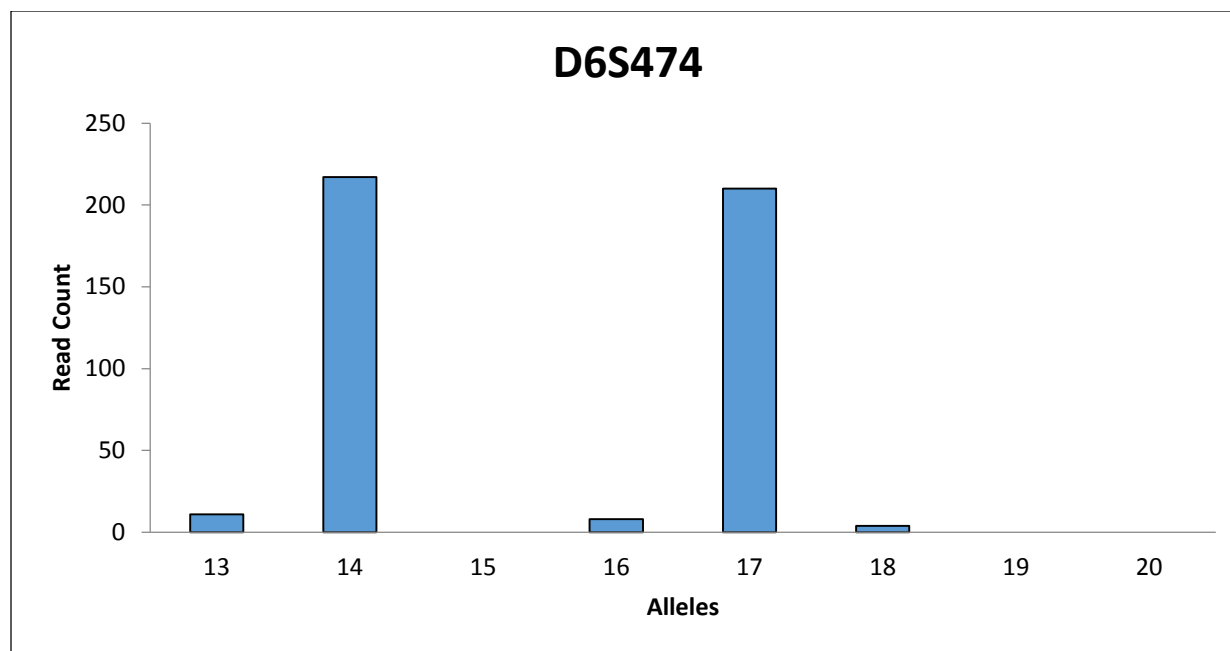


Fig. 2. Histogram generated for locus D6S474 in dataset 3 (Read 1). Histograms generated by the supplemental tool included with STRait Razor v2.0 resemble traditional electropherograms. Read counts displayed on the Y-axis are analogous to RFU values for peak heights. Here, the true alleles are “14” and “17.” Stutter peaks “13” and “16” can be seen to the left of the major allele peaks. This profile also shows the plus stutter “18” peak to the right of the major “17” allele peak.

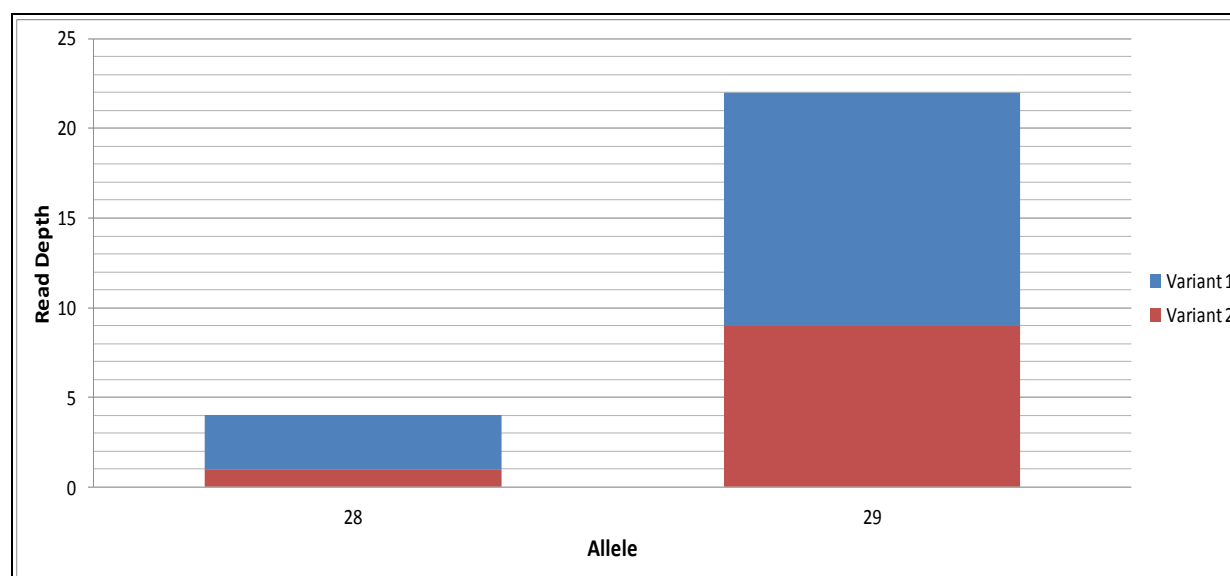


Fig. 3. Example of intra-repeat variation detected at locus D21S11 in dataset 4. This individual is homozygous for the “29” allele, Two distinct variants of this allele were detected. These variants also are seen in the stutter allele.

CONCLUSIONS

STRait Razor v2.0 includes a number of improvements over the original version of this STR detection software. This update retains the reliable and accurate allele-calling capability of the initial release, with the added benefit of a much larger range of detectable loci. The ability to detect autosomal, Y-chromosome, and now X-chromosome STRs augments the usefulness of the software and provides for a wider range of potential applications. The enhanced custom locus configuration file generator and supplemental tools, such as the genotyper and histogram generator, have made STRait Razor v2.0 much more user-friendly and facilitate ease and speed of analysis. Genotypes are determined quickly and can be reviewed by the analyst readily. Aspects such as stutter can be investigated in a manner that resembles that of current electropherogram interpretation. Finally, intra-repeat nucleotide variation is presented in a much easier way to analyze. The sorting of unique sequences for each allele by the total read count allows analysts to rapidly distinguish true allelic variants from those that may be due to simple sequencing errors or background noise.

STRait Razor v2.0 is free to use and available online ([http:// web.unthsc.edu/info/200210/molecular_and_medical_genetics/887/research_and_development_laboratory/5](http://web.unthsc.edu/info/200210/molecular_and_medical_genetics/887/research_and_development_laboratory/5)). Updates and new content will be added to the website as they are developed and tested. As always, modification and improvement to the program by other users and developers, as well as constructive feedback, are encouraged.

CONFLICT OF INTEREST

None.

ACKNOWLEDGMENTS

This work was supported in part by award no. 2012-DN-BXK033, awarded by the National Institute of Justice, Office of Justice Programs, U.S. Department of Justice. The opinions, findings, and conclusions or recommendations expressed in this publication are those of the authors and do not necessarily reflect those of the U.S. Department of Justice.

REFERENCES

- [1] D.H. Warshauer, D. Lin, K. Hari, R. Jain, C. Davis, B. LaRue, J.L. King, B. Budowle, STRait Razor: a length-based forensic STR allele-calling tool for use with second generation sequencing data, *Forensic Sci. Int. Genet.* 7 (2013) 409–417.
- [2] M. Gymrek, D. Golan, S. Rosset, Y. Erlich, lobSTR: a short tandem repeat profiler for personal genomes, *Genome Res.* 22 (2012) 1154–1162.
- [3] D.M. Bornman, M.E. Hester, J.M. Schuetter, M.D. Kasoji, A. Minard-Smith, C.A. Barden, S.C. Nelson, G.D. Godbold, C.H. Baker, B. Yang, J.E. Walther, I.E. Ternes, P.S. Yan, B. Rodriguez, R. Bundschuh, M.L. Dickens, B.A. Young, S.A. Faith, Short-read, high-throughput sequencing technology for STR genotyping, *Biotech. Rapid Dispatches* (2012) 1–6.
- [4] S.L. Fordyce, M.C. Ávila-Arcos, E. Rockenbauer, C. Børsting, R. Frank-Hansen, F.T. Petersen, E. Willerslev, A.J. Hansen, N. Morling, M.T. Gilbert, High-throughput sequencing of core STR loci for forensic genetic investigations using the Roche Genome Sequencer FLX platform, *Biotechniques* 51 (2011) 127–133.
- [5] C. Van Neste, M. Vandewoestyne, W. Van Criekinge, D. Deforce, F. Van Nieuwerburgh, My-Forensic-Loci-queries (MyFLq) framework for analysis of forensic STR data generated by massive parallel sequencing, *Forensic Sci. Int. Genet.* 9 (2014) 1–8.
- [6] STRBase: http://www.cstl.nist.gov/strbase/str_fact.htm.
- [7] ChrX-STR.org 2.0: <http://www.chrx-str.org/>.
- [8] NCBI: <http://www.ncbi.nlm.nih.gov/>.
- [9] Sorenson Molecular Genealogy Foundation Y Marker Details: http://www.smgf.org/ychromosome/marker_details.jsp.

SECTION 2

Initial Evaluation of Massively Parallel Sequencing as a Tool for Comprehensive Human Identification Marker Testing

CHAPTER 3

Massively Parallel Sequencing of Forensically Relevant Single Nucleotide Polymorphisms Using TruSeq™ Forensic Amplicon

Published in International Journal of Legal Medicine
2015, 129:31-36

David H. Warshauer
Carey P. Davis
Cydne Holt
Yonmee Han
Paulina Walichiewicz
Tom Richardson
Kathryn Stephens
Anne Jager
Jonathan King
Bruce Budowle

ABSTRACT

The TruSeq™ Forensic Amplicon library preparation protocol, originally designed to attach sequencing adapters to chromatin-bound DNA for chromatin immunoprecipitation sequencing (TruSeq™ ChIP-Seq), was used here to attach adapters directly to amplicons containing markers of forensic interest. In this study, the TruSeq™ Forensic Amplicon library preparation protocol was used to detect 160 single nucleotide polymorphisms (SNPs), including human identification SNPs (iSNPs), ancestry, and phenotypic SNPs (apSNPs) in 12 reference samples. Results were compared with those generated by a second laboratory using the same technique, as well as to those generated by whole genome sequencing (WGS). The genotype calls made using the TruSeq™ Forensic Amplicon library preparation protocol were highly concordant. The protocol described herein represents an effective and relatively sensitive means of preparing amplified nuclear DNA for massively parallel sequencing (MPS).

KEYWORDS: TruSeq™ custom amplicon · Massively parallel sequencing · Ancestry informative markers · Phenotypic SNPs

INTRODUCTION

Forensic DNA analysis is an extremely valuable tool for human identity testing in a number of situations, including criminal cases, mass disaster scenarios, and instances involving missing persons. Given their high power of discrimination and relatively small amplicon size, short tandem repeats (STRs) usually are the markers of choice for analysis of forensic biological evidence. However, there are a number of situations in which single nucleotide polymorphism (SNP) typing may provide an adjunct, alternative, or better option. The overall amplicon size of SNPs can be designed to be shorter than that of even a mini-STR marker [1] while retaining a level of discriminatory power that is comparable to STRs [2], assuming that a sufficient number of SNPs are typed. This quality makes SNP analysis a powerful tool in situations where, for example, evidentiary DNA is highly degraded.

To date, a variety of typing methodologies have been utilized for the analysis of SNP markers. These approaches include, but are not limited to, single base extension, allele-specific hybridization assays, chip-based microarrays, and mass spectrometry [3–10]. While each of these methods has its merits, they have inherent limitations. The most significant drawbacks are high-input DNA requirements, lack of quantitation, low throughput, high cost, inability to type large numbers of SNPs in a single analysis, and/or limitations requiring typing STRs and SNPs in separate reactions and separate runs.

Recently, massively parallel sequencing (MPS) has been shown to be a promising method for the detection of forensic SNP markers [2]. The number of SNPs that can be detected in a single

analysis with MPS is far greater than with the aforementioned methods, and the throughput of MPS is unparalleled. Moreover, up to 384 different samples currently can be typed simultaneously [11]. In addition, advances in sequencing technology have lowered both the cost and time required for analysis to a point that makes MPS cost-effective and competitive with other typing technologies.

Chromatin immunoprecipitation sequencing (ChIP-Seq) is a technology in which genomic DNA is cross-linked with chromatin and enriched before being subjected to MPS [12]. Traditionally, it has been used to investigate the distribution, abundance, and characteristics of DNA-bound protein targets across a genome of interest. The TruSeq™ ChIP sample preparation kit (Illumina, Inc., San Diego, CA) provides a simple workflow that allows preparation of chromatin-bound DNA for sequencing via the attachment of TruSeq™ adapters.

In this study, the TruSeq™ ChIP protocol was modified to enable library preparation of forensically relevant SNP-containing amplicons. This modified protocol, known as TruSeq™ Forensic Amplicon, was used to detect a battery of 160 human identification SNPs (iSNPs), ancestry, and phenotypic SNPs (apSNPs) in a set of 12 reference samples. The resulting data were analyzed for both sequence coverage and heterozygote allele balance. Results presented here illustrate the efficacy of this method.

MATERIALS AND METHODS

Nuclear DNA amplicons containing iSNPs and apSNPs were subjected to the TruSeq™ Forensic Amplicon protocol and subsequently sequenced on the MiSeq™ platform. Following the University of North Texas Health Science Center Institutional Review Board approval, quantitated human DNA control samples from 12 unrelated individuals (obtained from Coriell Institute for Medical Research, Camden, NJ) were used for this proof-of-concept study.

Normalization, primer design, and amplification

The 12 DNA control samples were normalized to 1 ng/μL. The normalized samples were verified to be 1 ng/μL using the Quantifiler® Human DNA Quantification Kit on the ABI 7900HT Fast Real-Time PCR System (ThermoFisher, Carlsbad, CA) following the manufacturer's recommendations.

PCR primers were designed manually using OligoAnalyzer 3.1 (Integrated DNA Technologies (IDT), Coralville, IA), Primer3, and UCSC Genome Browser [13, 14]. Two sets of desalted primer (IDT) pools were created by adding each locus-specific primer (forward and reverse) into a multiplex set. A pool of 94 iSNPs and a separate pool of aSNPs and pSNPs, totaling 56 and 10, respectively, were created (Supplemental Table 1). For the iSNP master mix, 12.5 μL of 2× Qiagen Multiplex PCR Master Mix (Qiagen Inc., Valencia, CA), 2.4 μL of the iSNP primer mix, 10.1 μL of laboratory grade water, and 1 μL of the respective normalized sample were added to each well of a 96-well plate. For the apSNP master mix, 12.5 μL of 2× Qiagen Multiplex PCR

Master Mix, 1.65 μL of the apSNP primer mix, 10.85 μL of laboratory grade water, and 1 μL of the respective normalized sample were added to each well of a 96-well plate.

The samples then were amplified using a Bio-Rad Tetrad 2 thermal cycler (Bio-Rad Laboratories, Inc., Hercules, CA) with the following PCR parameters: 95 °C for 11 min, 96 °C for 1 min, 35 cycles of 94 °C for 30 s, 58 °C for 30 s with a 0.5 °C/s ramp rate, 68 °C for 45 s with a 0.2 °C/s ramp rate, then 60 °C for 30 min and a hold at 10 °C.

Library preparation

The TruSeq™ Forensic Amplicon library preparation protocol recommends an amplified DNA input volume of 50 μL , at a concentration of 20–2000 pg/ μL (i.e., 1–100 ng total input DNA). Following these guidelines, the amplified products generated from each PCR were normalized at 0.5 ng/ μL at a volume of 50 μL in a 96-well plate, for a total of 24 wells each containing 25 ng of amplified DNA. A second laboratory (at Illumina) used 1 μL of 1 ng/ μL amplicons instead.

The TruSeq™ Forensic Amplicon library preparation process is similar to that of TruSeq™ ChIP, except that it uses PCR amplicons as starting material rather than chromatin-bound DNA. The process began with end repair, where the 5' ends of the amplicons were made blunt and phosphorylated during a 30-min incubation at 30 °C in an Applied Biosystems® GeneAmp® PCR System 9700 thermal cycler (Life Technologies). All subsequent incubation and amplification processes were carried out on this thermal cycler platform. Next, the samples were washed using AMPure XP beads and 80 % ethanol. The blunt ends then were adenylated, which prevented them

from ligating to each other during adapter ligation. Adenylation was performed by thermal cycling using the following parameters: 37 °C for 30 min, 70 °C for 5 min, and a final hold at 4 °C. Following adenylation, adapter ligation was performed, wherein TruSeq™ indexed adapters were bound to the adenylated 3' ends of the amplicons. Each sample was bound to adapters with a unique index sequence for multiplexed sequencing. Adapter ligation required a 10-min incubation at 30 °C, followed by washing using AMPure XP beads and 80% ethanol. For enrichment of adapter-bound amplicons, PCR was carried out using primers designed to amplify only those amplicons with adapters bound to them. The enrichment PCR parameters were: 98 °C for 30 s, 18 cycles of 98 °C for 10 s, 60 °C for 30 s, and 72 °C for 30 s, a final extension at 72 °C for 5 min, and a final hold at 4 °C. Enrichment PCR was followed by washing with AMPure XP beads and 80% ethanol.

Following library preparation, the adapter-ligated amplicons were quantified using the Qubit® platform (Life Technologies), according to the manufacturer's protocol. Based on the quantification results, the samples were normalized to a concentration of 10 nM with 10 mM Tris–HCl buffer at pH 8.5 with 0.1% Tween 20, as per Illumina® guidelines. A total of 5 µL of each sample were used to pool samples together for a total 10 nM sample pool of 120 µL.

MiSeq™ sequencing and data analysis

To prepare for sequencing on the MiSeq™ (Illumina), 10 µL of the 10 nM sample pool were combined with 40 µL of 10 mM Tris–HCl buffer at pH 8.5 with 0.1% Tween 20, for a resultant concentration of 2 nM. Illumina®'s library preparation guidelines for the MiSeq™ were

followed, and the concentration of the pooled sample was brought down to 12 pM using chilled HT1 buffer. Paired-end sequencing was performed, with a read length of 120 bases.

The sequencing sample sheet for these samples was created using the Illumina Experiment Manager. For this modified protocol, the “TruSeq™ Amplicon” workflow was used, and the samples were treated as custom amplicons. Once the sample sheet was created, it was edited by changing the index sequences used in the “TruSeq™ Amplicon” workflow to those used in the TruSeq™ Forensic Amplicon protocol. A custom manifest file was used for the sequencing run to define the position and names of each of the SNPs of interest. Using this manifest, MiSeq Reporter was able to produce *vcf* files for each sample which identified each SNP detected during sequencing.

Since MiSeq Reporter limits sequence coverage values for SNPs to 5000× by default, a separate method of variant-calling was required to ascertain the actual coverage at each locus of interest so that conclusions could be drawn with regard to the depth of sequencing and heterozygote balance afforded by the TruSeq™ Forensic Amplicon library preparation method. To this end, bam files were subjected to variant-calling without downsampling using the GATK [15]. Heterozygote balance was calculated by dividing the lower allele coverage value at each heterozygous SNP locus by the higher coverage value, yielding a heterozygote balance percentage.

RESULTS

Through the use of the TruSeq™ Forensic Amplicon library preparation protocol, SNP genotypes were generated for all 160 targeted iSNPs and apSNPs in 11 of the 12 samples analyzed. In sample 9, rs10776839 was not called due to low coverage (this particular SNP displayed low sequencing read depth across all samples). The amplified products of 11 of the samples tested in this study also were analyzed by a separate laboratory at Illumina. The SNP genotypes yielded were highly concordant between the two laboratories. Of the 11 samples compared, 7 were 100% concordant at all 160 SNPs. The concordance between the four remaining samples was between 98.75 and 99.38%. Discordant genotype calls can be found in Supplemental Table 2. It should be noted that discordant genotype calls between these two datasets were mainly due to differences in the calling of heterozygous versus homozygous genotypes, based on the heterozygosity thresholds used during analysis. For example, at rs2399332 in sample 2, the “T” allele displayed a coverage value of 312 reads, while the “G” allele had a coverage value of 3474 reads, which equates to a heterozygosity balance of approximately 9 % (Supplemental Table 2). The in-house heterozygosity threshold was set at 5%, while a 10% threshold was used by the second laboratory. Thus, this SNP was called in-house as a heterozygote, while the second laboratory determined that the SNP was homozygous for the “G” allele. Such results are, in effect, concordant. Indeed, if the in-house results are interpreted using a 10% heterozygosity threshold, the concordance values rise to 100% in 10 out of the 11 samples compared. The recalculated concordance value for the remaining sample (sample 10) would be 99.38%, due to a single discordance at rs2399332, where the in-house heterozygosity value was 12.9%, and thus only slightly above the 10% threshold. While the in-house heterozygosity threshold value was chosen arbitrarily for this study to simply demonstrate

proof of concept, this occurrence highlights the need for reliable thresholds developed through proper validation in each testing laboratory. Overall, the results are similar across all SNPs and differ only due to thresholds and variation of the lower signal SNP. Primer redesign for these loci may improve allele imbalance.

Whole genome sequencing (WGS)-based SNP calls were obtained from the Complete Genomics FTP site [16] for additional concordance testing. The allele calls derived from the in-house data produced by the TruSeq™ Forensic Amplicon library preparation method displayed a high concordance (96.23 to 98.74%) with the WGS data across all 12 samples. Discordance between the WGS-derived SNP calls and the in-house calls was found at a total of nine out of the 160 SNPs (rs1029047, rs1058083, rs10776839, rs10954737, rs12997453, rs182549, rs2399332, rs430046, and rs907100). It should be noted that the discordances between the in-house calls and the Illumina calls listed above were corroborated by the WGS data, consistent with the calls made by the second laboratory. This agrees with the explanation that these discordances were simply the result of threshold differences. The remaining discordant SNP loci between the WGS and in-house data appear to be discordance “hotspots” for this particular multiplex design, as all but one of the loci showed discordance in at least four of the samples tested (Table 1). Phase 3 data from the 1000 Genomes Project were available for samples 1 and 12, and a comparison showed that the phase 3 genotype calls for these samples were consistent with the WGS calls. The vast majority of the discordance (all but 3 of the total 53 discordant calls, across all samples) consisted of differences between heterozygous and homozygous allele SNP calls, which can once again be explained by differences in heterozygosity thresholds. However, a nucleotide variation within the primer binding site may have resulted in a failure to amplify one of the alleles at a given locus.

Other explanations include factors such as multiplex inefficiency, low coverage leading to skewed SNP calls, and simple alignment errors.

	1	2	3	4	5	6	7	8	9	10	11	12
rs1029047	A/T : A	A/T : A		A/T : A		A/T : A				A/T : A	A/T : A	
rs1058083					A/G : G		A/G : G	A/G : G		A/G : G		
rs10776839	G : G/T			G : G/T	G : G/T		G : G/T	G : G/T		G : G/T	G : G/T	G : T
rs10954737	T : C/T	T : C/T	T : C/T					T : C/T	T : C/T	T : C/T		
rs12997453		G : A/G										
rs182549	C/T : T	C/T : T	C/T : T	C/T : T					C/T : T	C/T : T		
rs2399332	G/T : G	G/T : G					G/T : G	G/T : G		G/T : G	G/T : G	
rs430046	C : C/T	C : C/T		C : C/T	C : C/T	C : T	C : C/T	C : C/T	C : C/T		C : C/T	C : T
rs907100						G : C/G	G : C/G	G : C/G	G : C/G	G : C/G		G : C/G

Table 1. SNP discordance (in-house vs. WGS). Discordance between the SNP calls generated in this study and

those obtained through whole genome sequencing are listed. Discordance is shown in the following format: “in-house call: WGS call.”

Overall, the heterozygote balance achieved through the use of the TruSeq™ Forensic Amplicon library preparation method was quite even. Across all samples, between 91.9 and 100% of the heterozygous loci showed allelic balance ratios of 1:2 (50% balance) or better. An example of heterozygous allele balance is shown in Fig. 1. The heterozygous loci for which allelic balance ratios dropped below 1:2 are shown in Supplemental Table 3. In some cases, allelic imbalance was explained by low coverage (e.g., rs1029047 in sample 2, which had a relatively low coverage of 281 reads and displayed a heterozygosity balance value of 12.6%), but other factors such as those noted above may explain imbalance in heterozygous loci with higher coverage values.

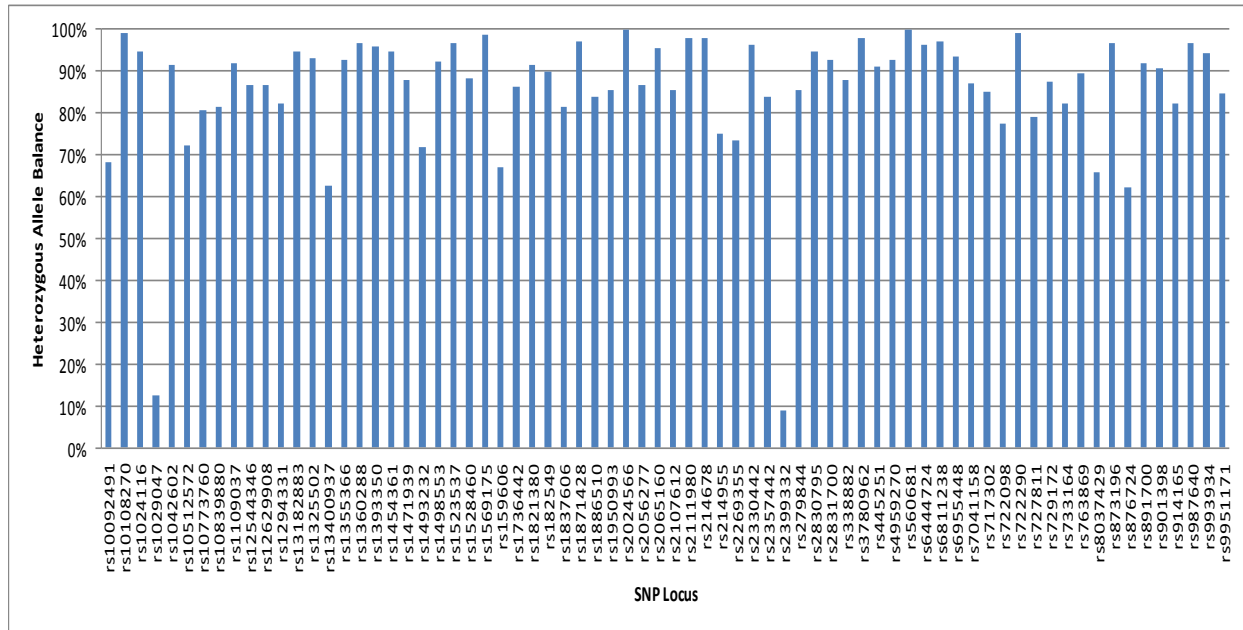


Fig. 1. Heterozygous allele balance for sample 2. Allele balance at heterozygous loci, expressed as a percentage, is shown. A value of 100 % denotes a perfect 1:1 balance of alleles. In this sample, only 2 loci (rs1029047 and rs2399332) display an allele balance value of less than 50%

The average sequencing coverage per locus across all 12 samples with both panels (i.e., effectively 24 samples) ranged from 142 $\times$ to 46,908 $\times$, and coverage was relatively consistent between samples at each locus. Figure 2 illustrates the sequence coverage across the apSNP loci, as an example. The wide range of coverage is most likely due to differences in amplification efficiency of the multiplex PCR. Further optimization is underway to reduce the coverage range.

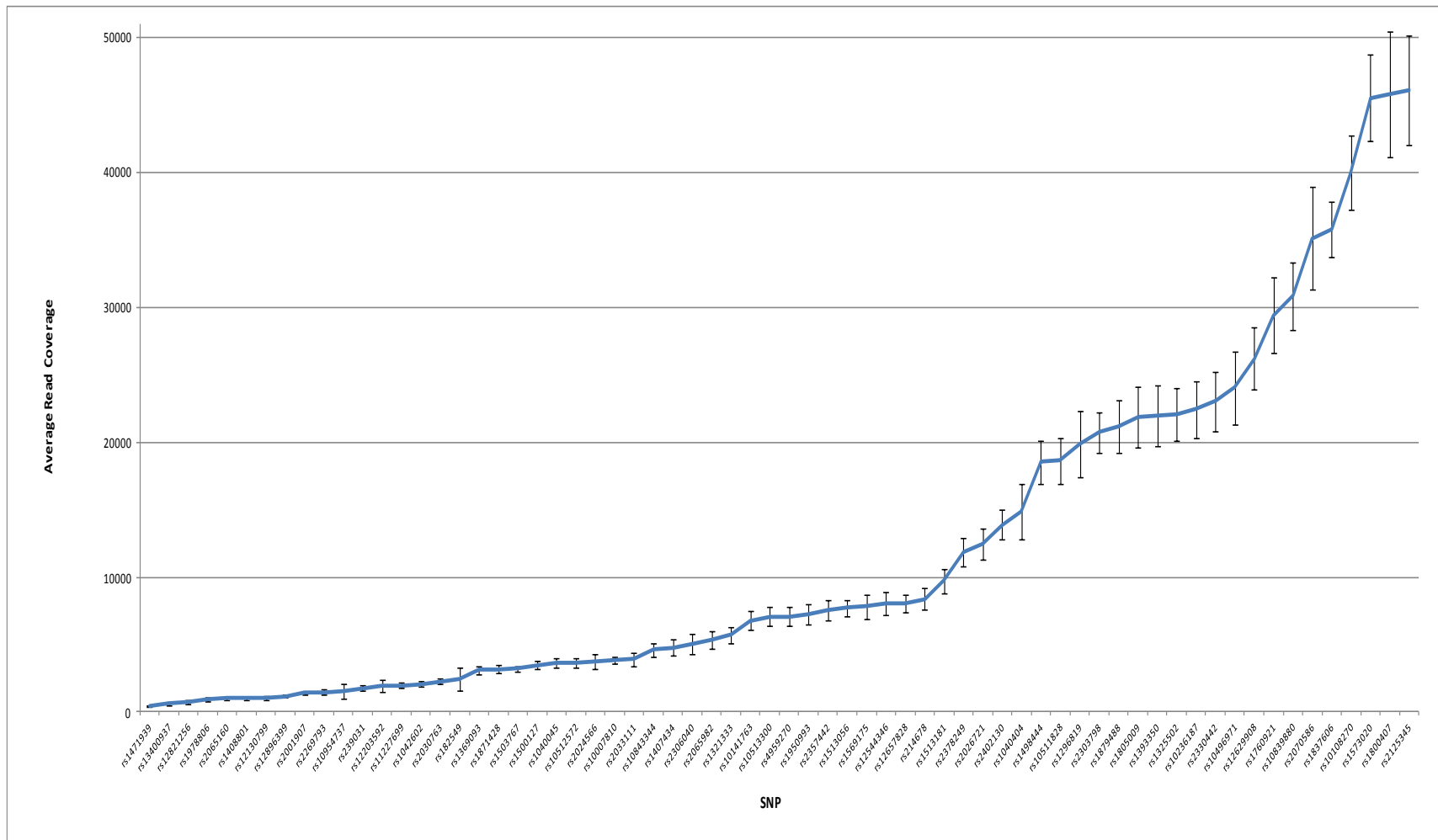


Fig. 2. Average sequence coverage for apSNP loci. The average depth of coverage across all samples for each apSNP locus is shown here. Bars represent the standard deviation

CONCLUSIONS

The results of this proof-of-concept study indicate that the TruSeq™ Forensic Amplicon library preparation protocol is an effective method of preparing amplified nuclear DNA for massively parallel sequencing. This method is less labor-intensive than alternative techniques. Unlike the TruSeq™ Custom Amplicon workflow, TruSeq™ Forensic Amplicon does not require the use of custom-designed oligonucleotide probes for library preparation. Additionally, the TruSeq™ Forensic Amplicon library preparation method is highly sensitive, with a relatively low input DNA requirement (1 ng of input DNA was amplified and 25 ng of amplified DNA were used for each sample, and at the second laboratory, 1 µL of 1 ng/µL amplicons was used, as opposed to the recommended 500 ng of input DNA recommended for the TruSeq™ Enrichment protocol). This lower DNA input is more suited for the quantities of sample DNA often encountered in forensic casework. In conjunction with a properly designed multiplex PCR, this preparation method is capable of producing reliable sequencing results with relatively even allele balance at heterozygous loci. Though not tested in this study, it is likely that the TruSeq™ Forensic Amplicon kit could be used for the preparation and detection of STR markers. The results of this proof-of-concept study suggest that this novel use of the original TruSeq™ ChIP protocol could support forensic genetic typing by MPS.

ACKNOWLEDGMENTS

This work was supported in part by award no. 2012-DN-BXK033, awarded by the National Institute of Justice, Office of Justice Programs, US Department of Justice. The opinions, findings,

and conclusions or recommendations expressed in this publication are those of the authors and do not necessarily reflect those of the US Department of Justice. The authors also would like to thank Illumina, Inc. for its support during this study.

CONFLICT OF INTEREST

C.P. Davis, C. Holt, Y. Han, P. Walichewicz, T. Richardson, K. Stephens, and A. Jager are employed by Illumina, Inc.

REFERENCES

1. Dixon LA, Murry CM, Archer EJ, Dobbins AE, Koumi P, Gill P (2005) Validation of a 21-locus autosomal SNP multiplex for forensic identification purposes. *Forensic Sci Int* 154:62–77
2. Seo SB, King JL, Warshauer DH, Davis CP, Ge J, Budowle B (2013) Single nucleotide polymorphism typing with massively parallel sequencing for human identification. *Int J Legal Med* 127:1079–1086
3. Sanchez JJ, Phillips C, Børsting C, Balogh K, Bogus M, Fondevila M, Harrison CD, Musgrave-Brown E, Salas A, Syndercombe-Court D, Schneider PM, Carracedo A, Morling N (2009) A multiplex assay with 52 single nucleotide polymorphisms for human identification. *Electrophoresis* 27:1713–1724
4. Pakstis AJ, Speed WC, Fang R, Hyland FC, Furtado MR, Kidd JR, Kidd KK (2010) SNPs for a universal individual identification panel. *Hum Genet* 127:315–324
5. Tomas C, Axler-DiPerte G, Budimlija ZM, Børsting C, Coble MD, Decker AE, Eisenberg A, Fang R, Fondevila M, Fredslund SF, Gonzalez S, Hansen AJ, Hoff-Olsen P, Haas C, Kohler P, Kriegel AK, Lindblom B, Manohar F, Maroñas O, Mogensen HS, Neureuther K, Nilsson H, Scheible MK, Schneider PM, Sonntag ML, Stangegaard M, Syndercombe-Court D, Thacker CR, Vallone PM, Westen AA, Morling N (2011) Autosomal SNP typing of forensic samples with the GenPlex™ HID System: results of a collaborative study. *Forensic Sci Int Genet* 5:369–375
6. Børsting C, Sanchez JJ, Morling N (2005) SNP typing on the NanoChip electronic microarray. *Methods Mol Biol* 297:155–168
7. Mengel-Jørgensen J, Sanchez JJ, Børsting C, Kirpekar F, Morling N (2005) Typing of multiple single-nucleotide polymorphisms using ribonuclease cleavage of DNA/RNA chimeric single-base extension primers and detection by MALDI-TOF mass spectrometry. *Anal Chem* 77:5229–5235
8. Freire-Aradas A, Fondevila M, Kriegel A-K, Phillips C, Gill P, Prieto L, Schneider PM, Carracedo Á, Lareu MV (2012) A new SNP assay for identification of highly degraded human DNA. *Forensic Sci Int Genet* 6:341–349
9. Musgrave-Brown (2007) Forensic validation of the SNPforID 52-plex assay. *Forensic Sci Int Genet* 1:186–190
10. Phillips C, Fang R, Ballard D, Fondevila M, Harrison C, Hyland F, Musgrave-Brown E, Proff C, Ramos-Luis E, Sobrino B, Carracedo A, Furtado MR, Syndercombe Court D, Schneider PM, the SNPforID Consortium (2007) Evaluation of the Genplex SNP typing system and a 49plex forensic marker panel. *Forensic Sci Int Genet* 1: 180–185

11. NuGEN (2013) Encore™ 384 Multiplex System. NuGEN. <http://www.nugeninc.com/nugen/index.cfm/products/pl/library-preparation/encore-384-multiplex-system/>. Accessed 30 January 2013
12. Park PJ (2009) ChIP-seq: advantages and challenges of a maturing technology. *Nat Rev Genet* 10:669–680
13. Untergrasser A, Cutcutache I, Koressaar T, Ye J, Faircloth BC, Remm M, Rozen SG (2012) Primer3—new capabilities and interfaces. *Nucleic Acids Res* 40(15):e115
14. Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, Haussler D (2002) The human genome browser at UCSC. *Genome Res* 12(6):996–1006
15. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M, DePristo MA (2010) The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res* 20:1297–1303
16. Complete Genomics FTP site: <ftp://ftp2.completegenomics.com>

SECTION 3

Design and Testing of an Extensive Capture-Based Assay for the Detection of Human Identification Markers

CHAPTER 4

Development of a Comprehensive Massively Parallel Sequencing Panel of Single Nucleotide Polymorphism and Short Tandem Repeat Markers for Human Identification

Submitted to Forensic Science International: Genetics
July 2015

David H. Warshauer
Xiangpei Zeng
Carey P. Davis
Jennifer Churchill
Nicole Novroski
Ranajit Chakraborty
Jonathan L. King
Bruce Budowle

ABSTRACT

A comprehensive, capture-based Nextera™ Rapid Capture panel was designed, which consisted of 84 STR (31 autosomal, 26 X-chromosome, and 27 Y-chromosome) and 275 SNP (240 autosomal and 35 Y-chromosome) markers. This panel was used to type 190 DNA samples from the U.S. Caucasian, African American, Hispanic, and Asian populations. Following sequencing on the MiSeq™, sequence data were analyzed using a combination of the onboard MiSeq™ Reporter software and the GATK, to detect SNP alleles. STRait Razor v2.0 was used to detect STR alleles. The performance of the panel was assessed, and a subset of the resulting genotype calls underwent concordance testing via comparison with genotype calls made using the ForenSeq™ DNA Signature Prep Kit. The overall performance of the capture-based comprehensive panel (based on locus performance and heterozygote balance) was similar to that of commercial PCR-based MPS kits. The detected genotypes also maintained a high degree of concordance with the ForenSeq™ data. Population genetic statistical analyses were performed to determine allele frequencies (by gene counting) and F_{ST} values, as well as to test conformity with Hardy-Weinberg equilibrium and linkage equilibrium. Y-chromosome STR haplotype diversity also was calculated, and haplogroups were predicted based on the data. The results of the analyses were consistent with expectations, considering the relatively small sample size per population group. The results of this study support that a capture-based approach can produce robust data for typing reference samples.

KEYWORDS: Comprehensive panel · massively parallel sequencing · Nextera™ Custom Enrichment · MiSeq™ · STRait Razor · STRs · SNPs

INTRODUCTION

Over the past three decades, a number of robust and reliable DNA typing technologies for human identity testing have been implemented (1-3). These methods enable analyses of minute quantities of DNA and provide a resolving power such that, in many cases, the number of potential contributors of an evidence sample can be reduced to only a few individuals, if not a single source. The success of DNA typing has led to further applications, the most prominent of which has been its use for developing investigative leads. The realization of DNA typing for developing investigative leads and for solving future crimes came to fruition with the development of DNA databases. Many countries have established DNA databanks that contain profiles from convicted offenders, arrestees and forensic samples from unsolved cases (4, 5). These databases are designed to house DNA profiles that can be used to associate individuals with forensic samples or to identify missing persons. For example, the U.S. databank, CODIS (the Combined DNA Index System), contains more than 13,780,000 reference DNA profiles (6) (as of May 2015) and is routinely relied upon for helping to develop meaningful investigative leads. Due to the success in providing such leads, these DNA databases continue to increase in size, and the information contained within them may be used for purposes other than the direct matching of profiles, such as familial searching (7-9).

The reliance on offender, arrestee, and forensic sample DNA databases has driven and will continue to drive innovation and standards. Likewise, the demands of generating, entering, and maintaining DNA profiles in a national DNA database have fostered developments in automation and the creation of robust molecular assays. The number of reference samples from convicted

felons, arrestees, detainees, and missing persons continues to increase, and there is no indication of the demand subsiding. To meet the needs of forensic DNA typing and its infrastructure, it is imperative that forensic scientists embrace new technologies that will benefit the process, as well as society, by allowing for analyses of the ever-increasing numbers of reference samples, as well as for more challenging samples. Such advancements can provide an enhancement of the ability to solve crimes, a means to further facilitate the exoneration of the innocent, and an improvement of the capability to identify missing persons.

Hares (10,11), representing the position of the FBI, recommended that the core 13 STR loci for CODIS should be changed and augmented. The FBI advocated 20 STR loci to serve as the new CODIS panel of markers. Ge *et al.* (9) suggested that there were additional factors and applications for selecting a core set of markers beyond those relied upon by Hares. This alternate viewpoint was that the loci selected should be driven by the demands of casework (i.e., loci should be selected based on performance with degraded and inhibited samples, or the system should be versatile enough to enable a variety of search strategies). Thus, there are differences of opinions on what should constitute the core marker set. However, these discussions can be rendered moot with the advent of massively parallel sequencing (MPS) technology.

MPS technologies sequence DNA in a massively parallel fashion with high coverage, which can result in low error, as well as a high throughput of specified targets (12-20). Due to the exquisitely high throughput, a large battery of genetic markers can be analyzed simultaneously, far exceeding the current capacity of 15-27 STRs in a fluorescent multiplex/capillary electrophoresis (CE) system (21), and well beyond the 20 STR loci advocated by Hares. Indeed,

sequencing kits have been developed that contain reagents for typing 23 STR loci (22) and beyond, to include a set of Y-chromosome and X-chromosome STRs, and human identity SNPs (comprising more than two hundred markers) (23). Moreover, given the high throughput capacity afforded by MPS, many different samples can be individually labeled with short unique sequences of nucleotides, known as index barcodes, and multiplexed so that they may be sequenced at the same time. Thus, MPS technology has advanced to the point that it could be used to comprehensively characterize reference samples and potentially meet the demands of DNA database typing.

Most efforts toward forensic applications of MPS to date have focused on a PCR enrichment strategy. However, in the present study, a capture approach was employed instead. The capture approach was considered under the assumption that the demands of design for target probes would be simpler than the demands of primer design in a multiplex, and the amount of template DNA required from reference samples would be less constraining than it would be for crime scene samples. In addition, using and establishing an alternate method could prove useful for concordance testing of methods using PCR enrichment. A comprehensive panel of 84 STRs (31 autosomal STRs, 26 X-chromosome STRs, and 27 Y-chromosome STRs) and Amelogenin, as well as 275 identity and Y-chromosome SNPs, was designed for the Nextera™ Rapid Capture Custom Enrichment library preparation method. The panel was used to type DNA samples from 190 Caucasian, African American, Hispanic, and Asian individuals. Sequence data generated by the MiSeq™ platform were analyzed using currently available bioinformatics tools, and the performance of the assay was evaluated. Concordance testing and population genetic analyses were performed. The results of this study indicate that MPS provides a system such that reference

samples can be typed economically for a much larger battery of diverse markers than is currently possible through the use of CE, while providing genotype data that are compatible with current database profile formats.

MATERIALS AND METHODS

Samples and Extraction

Following the University of North Texas Health Science Center Institutional Review Board approval, DNA was extracted from whole blood and saliva samples from a total of 190 unrelated individuals, consisting of 49 Caucasians (24 males and 25 females), 49 African Americans (23 males and 26 females), 49 Hispanics (18 males and 31 females), and 43 Asians (17 males and 26 females). These populations were chosen because they represent the 4 major U.S. populations. DNA extraction was performed using the Qiagen® QIAamp™ DNA Mini Kit (Qiagen, Hilden, Germany), according to the manufacturer's suggested protocol.

Panel Design

The Nextera™ Rapid Capture Custom Enrichment panel employed in this study was designed using the Illumina® DesignStudio sequencing assay design tool. Nextera™ Rapid Capture chemistry (Illumina, Inc., San Diego, CA) is based on enzymatic “tagmentation” and probe-based “capture” enrichment. The Nextera™ library chemistry was selected for this study initially because avoiding PCR enrichment should reduce amplification artifacts. In addition,

primer binding site mismatch issues would not impact multiplex design or target enrichment success. It was hypothesized that a dense probe design would ensure capture of the target loci. Lastly, laying a foundation of sequence data with an alternate enrichment system could be useful when full validation studies are undertaken.

Custom oligonucleotide probes were designed to detect 84 forensically-relevant STRs, consisting of 31 autosomal STRs, 26 X-chromosome STRs, and 27 Y-chromosome STRs (Table 1). Amelogenin also was included in the panel design, for sex determination. To improve enrichment efficiency, multiple probes were used for each STR. Custom probes also were designed to allow the assay to detect the 275 SNPs, which consisted of 240 autosomal identity SNPs and 35 Y-chromosome SNPs (Table 2).

Probes (80 bases in length) for the Nextera™ Rapid Capture Custom Enrichment Kit were designed using Design Studio (Illumina), a freely-available software. The STRs and SNPs were tabulated including details regarding chromosomal positioning, target selection (Full Region), probe density requirements (due to the alignment-specific requirements of STRs, density of these markers was set at ‘ADJACENT’), and marker information. Marker data then were uploaded to Design Studio v1.5, and probes were generated under the default conditions (with hg19 for probe reference).

STR Name	Chromosome	Capture Region (bp)
D1S1627	1	106,963,664 - 106,963,777
D1S1656	1	230,905,304 - 230,905,457
TPOX	2	1,493,392 - 1,493,662
D2S1776	2	169,645,210 - 169,645,507
D2S1338	2	218,879,514 - 218,879,706
D2S441	2	68,238,997 - 68,239,157
D3S1358	3	45,582,204 - 45,582,335
FGA	4	155,508,847 - 155,509,043
D4S2408	4	31,304,233 - 31,304,509
D5S818	5	123,111,197 - 123,111,332
CSF1PO	5	149,455,734 - 149,456,053
D5S2500	5	58,697,192 - 58,697,344
D6S474	6	112,879,105 - 112,879,267
D6S1017	6	41,677,195 - 41,677,354
SE33	6	88,986,819 - 88,987,106
D7S820	7	83,789,440 - 83,789,683
D8S1179	8	125,907,079 - 125,907,260
D10S1248	10	131,092,481 - 131,092,583
TH01	11	2,192,213 - 2,192,381
D12S391	12	12,449,929 - 12,450,154
VWA	12	6,093,103 - 6,093,254
D13S317	13	82,722,055 - 82,722,247
D14S1434	14	95,308,356 - 95,308,578
PENTAE	15	97,374,211 - 97,374,590
D16S539	16	86,386,123 - 86,386,411
D17S1301	17	72,680,934 - 72,681,088
D18S51	18	60,948,813 - 60,949,143
D19S433	19	30,417,026 - 30,417,232
D21S11	21	20,554,258 - 20,554,481
PENTAD	21	45,055,995 - 45,056,424
D22S1045	22	37,536,297 - 37,536,453

STR Name	Chromosome	Capture Region (bp)
DXS7424	X	100,618,815 - 100,618,983
DXS101	X	101,413,015 - 101,413,242
DXS7133	X	109,041,542 - 109,041,664
GATA172D05	X	113,174,983 - 113,175,103
GATA165B12	X	120,877,967 - 120,878,096
DXS6854	X	128,688,897 - 128,689,006
HPRT8	X	133,615,404 - 133,615,691
DXS10101	X	133,654,442 - 133,654,698
GATA31E08	X	140,234,254 - 140,234,502
DXS8377	X	149,566,470 - 149,566,716
DXS10134	X	149,649,915 - 149,650,436
DXS7423	X	149,710,902 - 149,711,089
DXS9902	X	15,323,615 - 15,323,787
DXS10011	X	151,188,025 - 151,188,418
DXS6795	X	23,244,499 - 23,244,783
DXS6807	X	4,743,381 - 4,743,648
DXS7132	X	64,655,335 - 64,655,623
DXS10074	X	66,976,952 - 66,977,449
DXS981	X	68,197,358 - 68,197,545
DXS9895	X	7,377,106 - 7,377,253
DXS6800	X	78,680,409 - 78,680,603
DXS10135	X	9,306,117 - 9,306,616
DXS8378	X	9,370,225 - 9,370,429
DXS6801	X	92,511,171 - 92,511,301
DXS6809	X	94,938,152 - 94,938,411
DXS6789	X	95,449,413 - 95,449,554
DYS391	Y	14,102,765 - 14,102,872
DYS635	Y	14,379,516 - 14,379,692
DYS437	Y	14,466,963 - 14,467,156
DYS439	Y	14,515,187 - 14,515,408
DYS389I/II	Y	14,612,069 - 14,612,436

STR Name	Chromosome	Capture Region (bp)
DYS438	Y	14,937,784 - 14,938,104
DYS612	Y	15,752,548 - 15,752,752
DYS390	Y	17,274,883 - 17,275,099
DYS518	Y	17,319,838 - 17,320,117
DYS643	Y	17,425,984 - 17,426,129
DYS533	Y	18,393,104 - 18,393,318
GATAH4	Y	18,743,527 - 18,743,664
DYS385a	Y	20,801,575 - 20,801,750
DYS385b	Y	20,842,449 - 20,842,600
DYS460	Y	21,050,791 - 21,050,902
DYS549	Y	21,520,077 - 21,520,317
DYS392	Y	22,633,846 - 22,634,156
DYS448	Y	24,364,963 - 24,365,273
DYS393	Y	3,131,127 - 3,131,247
DYS505	Y	3,640,749 - 3,640,923
DYS456	Y	4,270,941 - 4,271,090
DYS570	Y	6,861,114 - 6,861,370
DYS576	Y	7,053,301 - 7,053,492
DYS522	Y	7,415,372 - 7,415,724
DYS458	Y	7,867,839 - 7,867,983
DYS449	Y	8,217,984 - 8,218,232
DYS481	Y	8,426,346 - 8,426,474
DYS627	Y	8,649,929 - 8,650,266
DYS19	Y	9,521,877 - 9,522,129

Marker Name	Chromosome	Capture Region (bp)
Amelogenin	Y	6,736,678 - 6,736,894

Table 1. Panel markers – STRs and Amelogenin. The STRs included in the panel design, as well as Amelogenin, are listed. The respective chromosome and chromosomal location for which the custom oligonucleotide capture probes were designed are listed for each marker.

Quantification and Normalization

To bring the 190 DNA samples to the desired input concentration of 5 ng/μL for the Nextera™ Rapid Capture Custom Enrichment protocol, the quantity of each sample was determined using the Qubit® fluorometric quantification method (Life Technologies, Carlsbad, CA) and normalized to 10 ng/μL with a 10 mM Tris-HCl solution at pH 8.5. The samples were quantified again and normalized in the same manner to a final concentration of 5 ng/μL, to ensure that the proper amount of genomic DNA would be used for library preparation.

Library Preparation

As required by the Nextera™ Rapid Capture Custom Enrichment protocol, 10 μL of each normalized sample were used for library preparation, for a total of 50 ng of genomic DNA per sample. The samples first underwent tagmentation by the Nextera™ transposome, whereby the samples are enzymatically cleaved and bound to sequencing adapters (24), at 58° C in an Applied Biosystems® GeneAmp® PCR System 9700 thermal cycler (Life Technologies). The tagmented samples then were purified via 2 magnetic bead-based 80% ethanol washes, and the fragment sizes of a small subset of these samples were analyzed using the Agilent® 2200 TapeStation™ (Agilent Technologies, Inc., Santa Clara, CA) to ensure that the tagmentation process was successful. Dual Nextera™ sequencing indices were attached to each of the tagmented samples by amplification in an Eppendorf® Mastercycler® Pro S thermal cycler (Eppendorf, Hamburg, Germany), using the following parameters: 72° C for 3 min, 98° C for 30 sec, 10 cycles of 98° C for 10 sec, 60° C for 30 sec, and 72° for 30 sec, a final extension at 72° C for 5 min, and a final hold at 10° C. Following bead-based amplification cleanup with 80% ethanol, the quantity of each indexed sample was

determined using the Qubit® platform. The samples were normalized and pooled for sequencing, 12 at a time, such that each library contained 500 ng of each uniquely indexed sample, for a total of 6,000 ng of genomic DNA per pool. The pooled libraries were hybridized to the custom oligonucleotide probes in an Eppendorf® Mastercycler® Pro S thermal cycler, using the following parameters: 95° C for 10 min, 18 cycles of 1 min incubations, starting at 94° C, then decreasing 2° C per cycle, and a final hold at 58° C for approximately 12 hours. A streptavidin bead-based cleanup step was performed wherein the libraries were washed twice for 30 minutes with an enrichment wash solution at 50° C. A second hybridization was performed, using the same thermal cycling parameters, except that the final hold at 58° C was extended to approximately 20 hours. Following a second heated streptavidin bead-based cleanup, the libraries underwent two additional magnetic bead-based washes with 80% ethanol. The libraries were enriched through amplification in an Eppendorf® Mastercycler® Pro S thermal cycler, using the following parameters: 98° C for 30 sec, 12 cycles of 98° C for 10 sec, 60° C for 30 sec, and 72° C for 30 sec, a final extension at 72° C for 5 min, and a final hold at 10° C. A final magnetic bead-based cleanup procedure was performed, consisting of 2 washes with 80% ethanol, and the quantity of the libraries was determined using the Qubit® platform. Following quantification, each library was analyzed on the Agilent® 2200 TapeStation™ to determine the average size of the enriched fragments.

MiSeq™ Sequencing

The concentration and size, in bp, of the Nextera™ Rapid Capture Custom Enrichment libraries were used to determine their molarity. To prepare for sequencing on the MiSeq™ (Illumina), each library was normalized to 2 nM using a solution of 10 mM tris-HCl buffer at pH

8.5 with 0.1% Tween 20. Illumina®'s library preparation guidelines for the MiSeq™ were followed, and the concentration of each library was adjusted to 12 pM using chilled HT1 buffer. Paired-end sequencing was performed, with a read length of 250 bases.

Data Analysis

MiSeq Reporter, the onboard alignment and SNP-calling software used by the MiSeq™, is capable of detecting SNP markers in the raw sequence data generated by the instrument. However, the software currently does not offer the ability to display coverage values for SNP genotypes that are identical to those of a given reference genome. Thus, when using MiSeq Reporter alone, it is impossible to differentiate between a SNP that is homozygous with the reference sequence and one that has simply dropped out. For this reason, an improvised workflow was created in which the raw sequence data were aligned by MiSeq Reporter, and the resulting *bam* files then were used as input for the Genome Analysis Toolkit (GATK) (25), which was given specific instructions to display information for all SNP markers, regardless of their relationship with the reference genome.

STR allele-calling required a different strategy. STRait Razor v2.0 (26) was used to analyze the FASTQ files produced by the MiSeq™ for each sample. STRait Razor's efficient STR allele detection method allows it to type alleles found in raw sequence data based on their length, while retaining their individual nucleotide sequences. Given that this study represents exploratory research, a coverage threshold of 1X and a stutter threshold of 20% were used for STR allele-

calling. This research-implemented threshold should not be construed as a recommendation for a long-term operational protocol.

The efficacy of the panel was determined, in part, by a statistic called “relative locus performance” (RLP), which indicates how well each locus was captured and detected. RLP is calculated by dividing the coverage (i.e., read depth) of a given locus by the total coverage of all loci in the sample, and normalizing this value by averaging it across all samples. Thus, low RLP values are indicative of loci that performed less efficiently in the assay, relative to the others. Panel performance also was evaluated by studying the amount of locus dropout in the dataset. Heterozygosity balance was another indicator of the assay’s overall performance and was calculated by comparing the coverage values for both alleles of any heterozygous pair at a given marker. A heterozygosity balance value of 1 indicates that both alleles had the same coverage, while <1 indicates some degree of imbalance between the heterozygous alleles.

Concordance testing was performed on the STR and SNP markers detected in the samples by comparison with genotype information generated through the use of the ForenSeq™ DNA Signature Prep Kit (23) followed by sequencing on the MiSeq™. The manufacturer’s recommended protocols were followed for library preparation and sequencing, with the following exception: 10 μ l of the pooled, normalized libraries were used in diluting and denaturing the libraries prior to sequencing. Data generated with the ForenSeq™ DNA Signature Prep Kit was analyzed with the ForenSeq™ Universal Analysis Software (UAS) and in-house Excel-based workbooks. Discordance was defined as any instance in which an allele detected through the use of the comprehensive panel differed from one detected by the

ForenSeq™ kit. The lack of an allele at a locus during comparison (due to dropout) does not represent discordance.

Population Statistical Analyses

STR and SNP allele frequencies were determined using Arlequin (27). Tests for Hardy-Weinberg equilibrium (HWE) and linkage disequilibrium (LD), as well as determination of F_{ST} values, were performed using GDA (28). Y-chromosome STR haplotypes were used to calculate haplotype diversity values (29) and determine a haplogroup prediction for each individual in the dataset using Haplogroup Predictor (39).

RESULTS AND DISCUSSION

Panel Performance

The custom panel probes were used to analyze DNA from 190 unrelated individuals to assess the general performance of the large capture-based multiplex. STR and SNP allele calls, as well as coverage information, can be found in Supplemental Workbooks 1 and 2. Since a tremendous amount of data is generated with these analyses, summary charts were generated for RLP values and, where appropriate, heterozygosity balance (also termed here as “allele coverage ratio,” ACR). For STRs, the data were separated into autosomal, X-chromosome female, X-chromosome male, and Y-chromosome STRs (Figures 1-6). For SNPs, the data were separated into autosomal and Y-chromosome SNPs (Figures 7-9). The overall performance (based on locus

performance and heterozygote balance) is similar to that of commercial PCR-based MPS kits. For all marker systems, the depth of coverage ranged from some low-signal loci to high-signal loci. These extremes are a small subset of the total markers, and the majority are well-balanced (i.e., within 2 SD of the mean; calculations not shown). With a CE-based approach, having such a wide range of signal in a multiplex would not be feasible because the high-signal loci typically would generate substantial noise in adjacent dye channels. However, the dynamic range with MPS is much greater, as the signal (and concomitant noise) from one marker does not directly affect the signal or noise at another marker. Therefore, the range of coverage seen in this comprehensive capture panel (and commercial PCR-based MPS kits) can be accommodated more easily. The limitation on such a wide range of coverage is sample throughput. Detection of lower-performing markers will limit the number of samples that can be multiplexed to ensure that routine typing will not result in an unreasonable amount of allele and locus dropout. Future studies could improve the balance by increasing the probe density designation for the lower-signal markers from “ADJACENT” to “OVERLAPPING” and the higher-signal markers from “ADJACENT” to “INTERMEDIATE” or “STANDARD” during the initial probe design stage. A more balanced system will increase sample multiplexing capability.

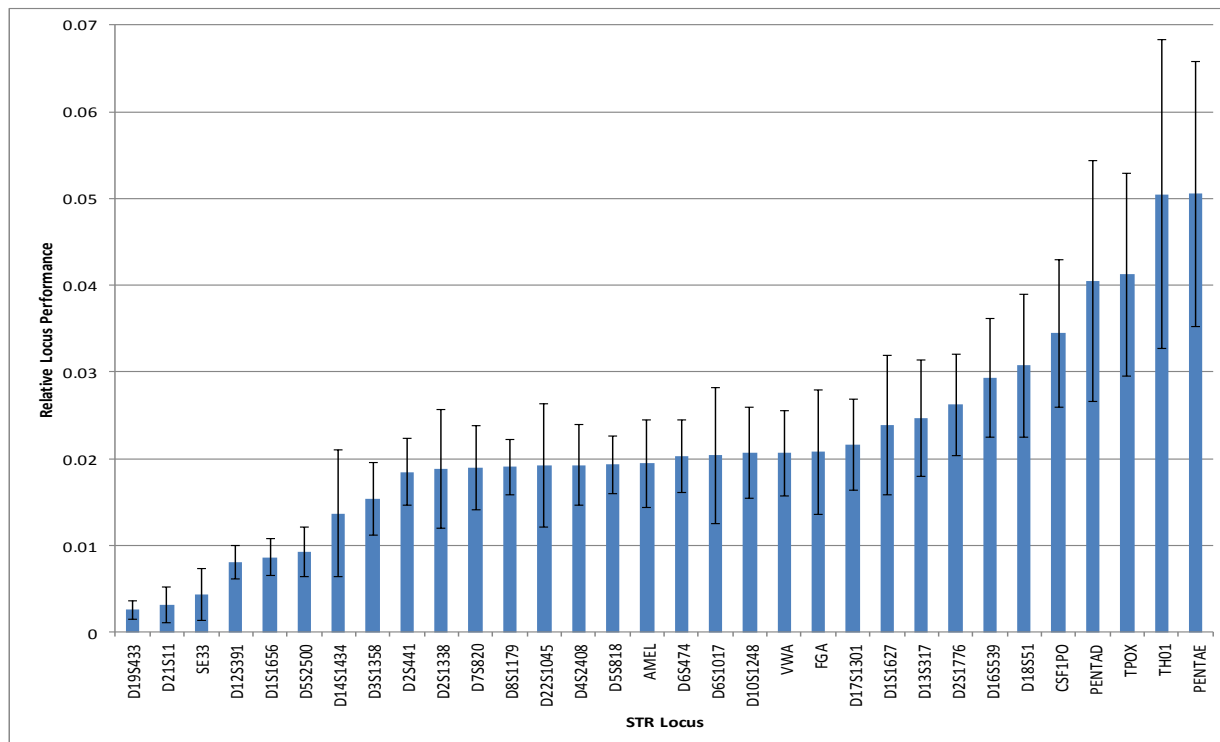


Figure 1. Relative locus performance – autosomal STRs and Amelogenin. Overall performance values, based on coverage, are shown for the 31 autosomal STRs and Amelogenin. Standard deviation is represented by an error bar.

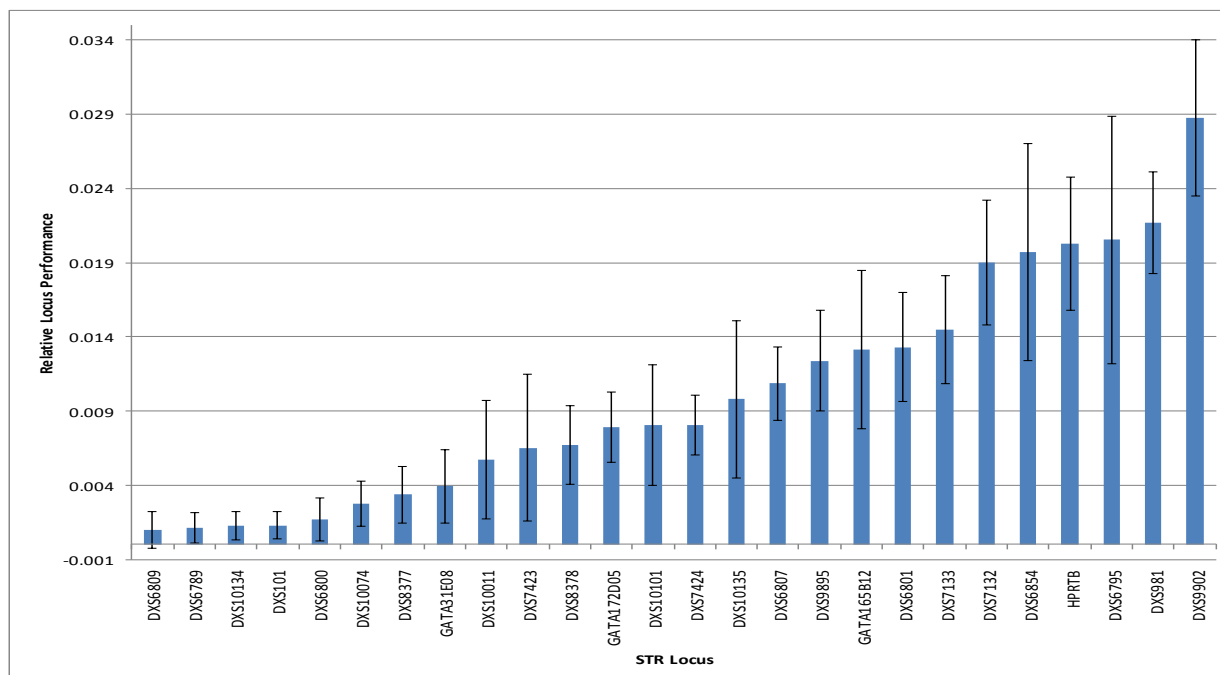


Figure 2. Relative locus performance – X-chromosome STRs (females only). Overall performance values, based on coverage, are shown for the 26 X-chromosome STRs. Standard deviation is represented by an error bar.

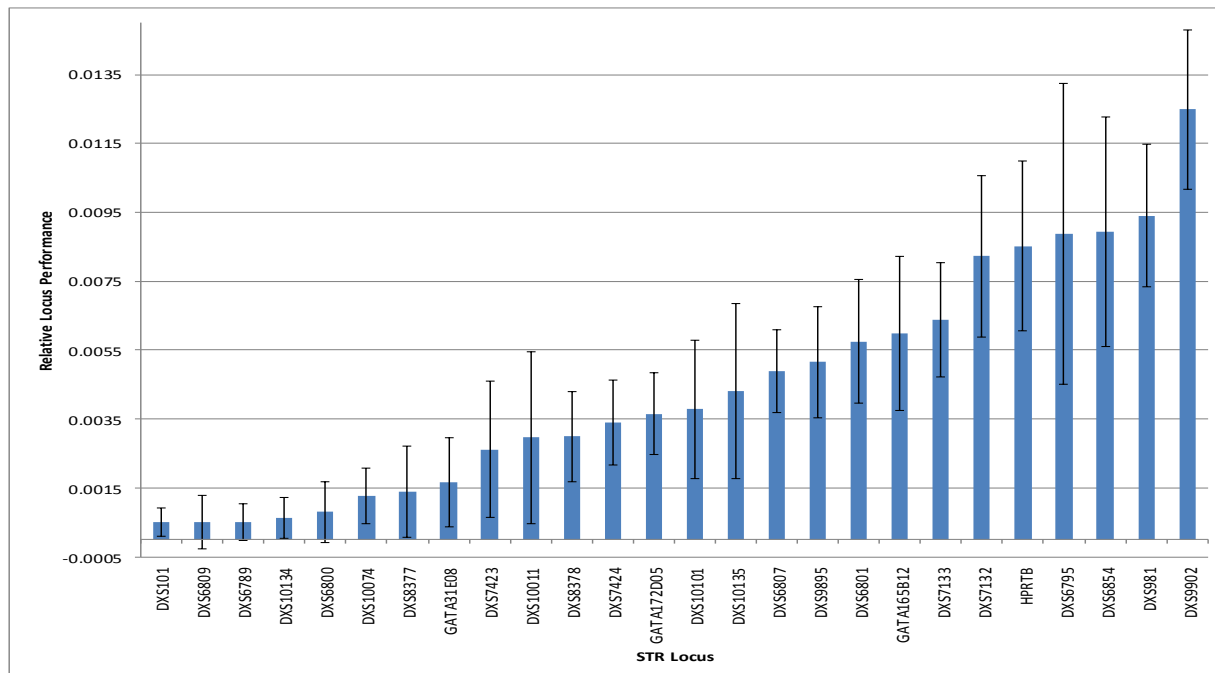


Figure 3. Relative locus performance – X-chromosome STRs (males only). Overall performance values, based on coverage, are shown for the 26 X-chromosome STRs. Standard deviation is represented by an error bar.

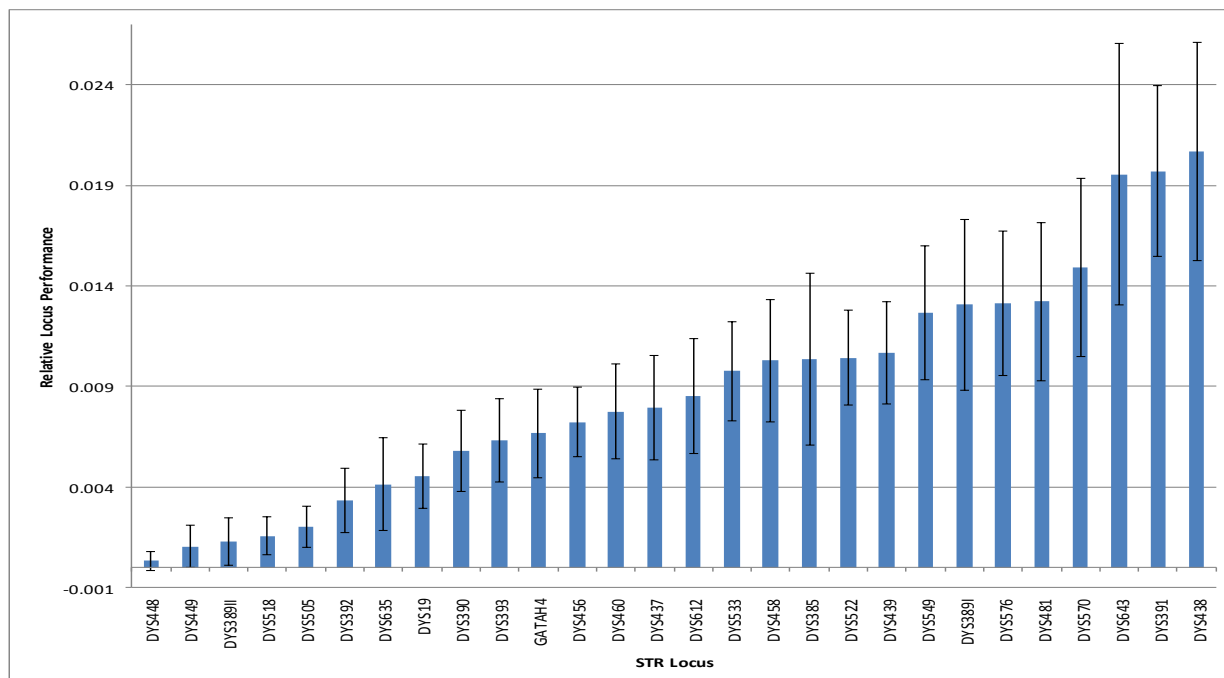


Figure 4. Relative locus performance – Y-chromosome STRs. Overall performance values, based on coverage, are shown for the 27 Y-chromosome STRs. The loci DYS389I and DYS389II are treated separately. Standard deviation is represented by an error bar.

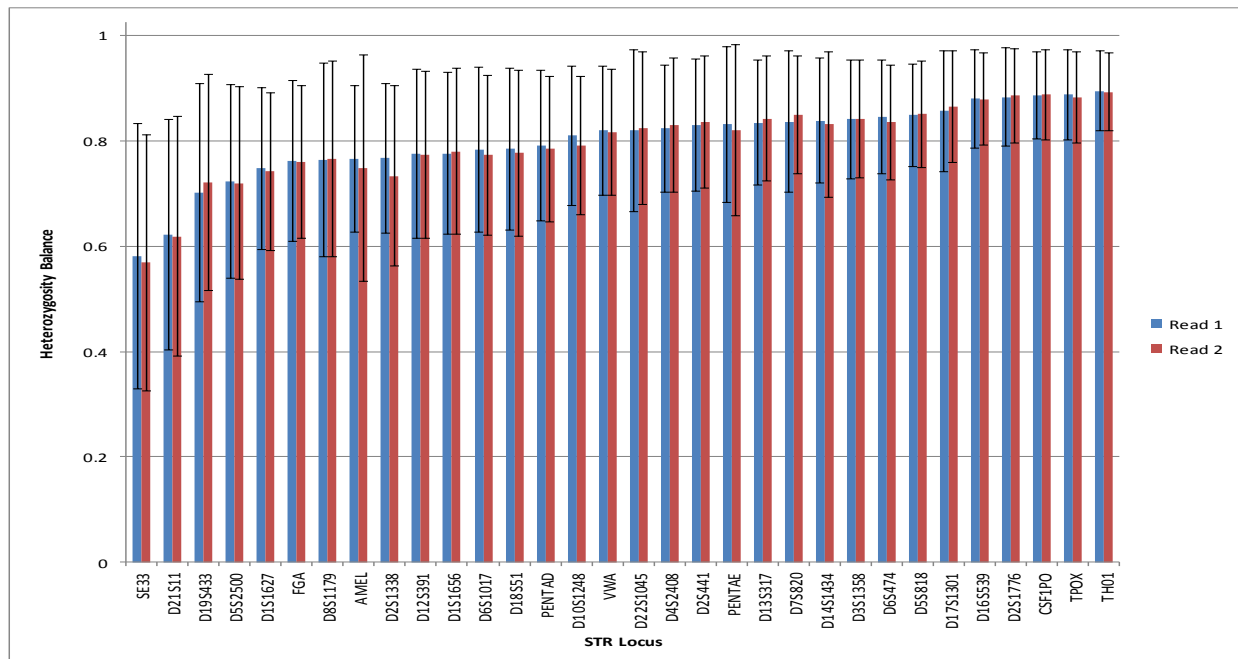


Figure 5. Heterozygosity balance – autosomal STRs and Amelogenin. ACRs for heterozygous alleles, based on coverage, are shown for the 31 autosomal STRs and Amelogenin. A heterozygosity balance value of 1 indicates perfect balance. Standard deviation is represented by an error bar.

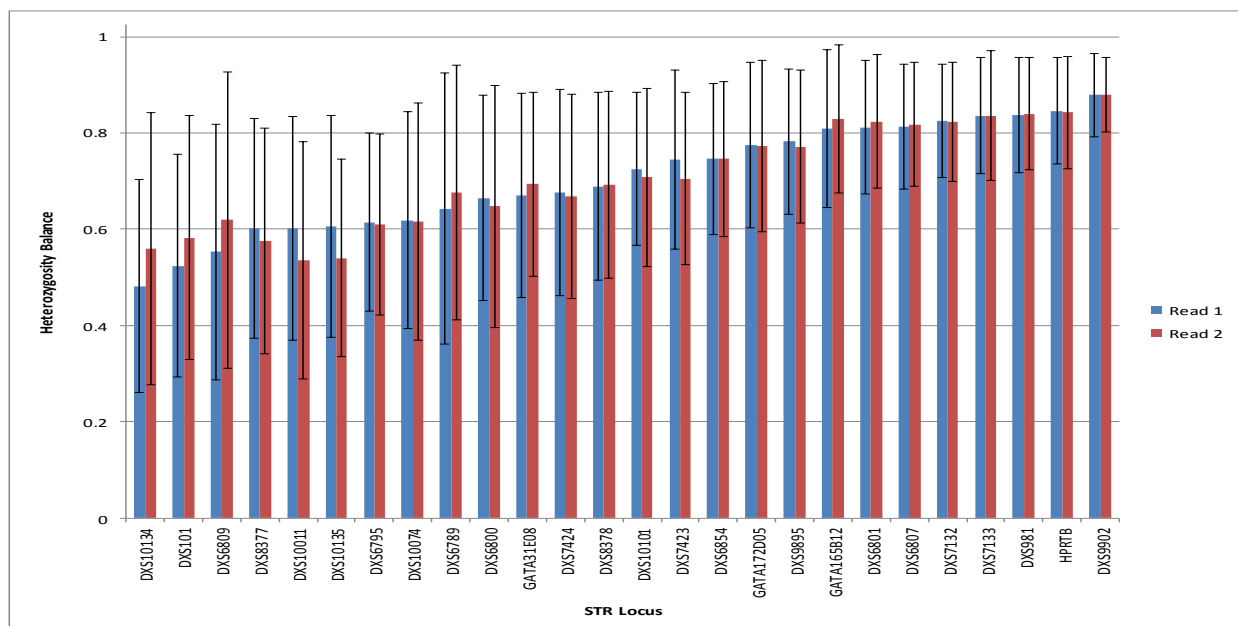


Figure 6. Heterozygosity balance – X-chromosome STRs (females only). ACRs for heterozygous alleles, based on coverage, are shown for the 26 X-chromosome STRs. A heterozygosity balance value of 1 indicates perfect balance. Standard deviation is represented by an error bar.

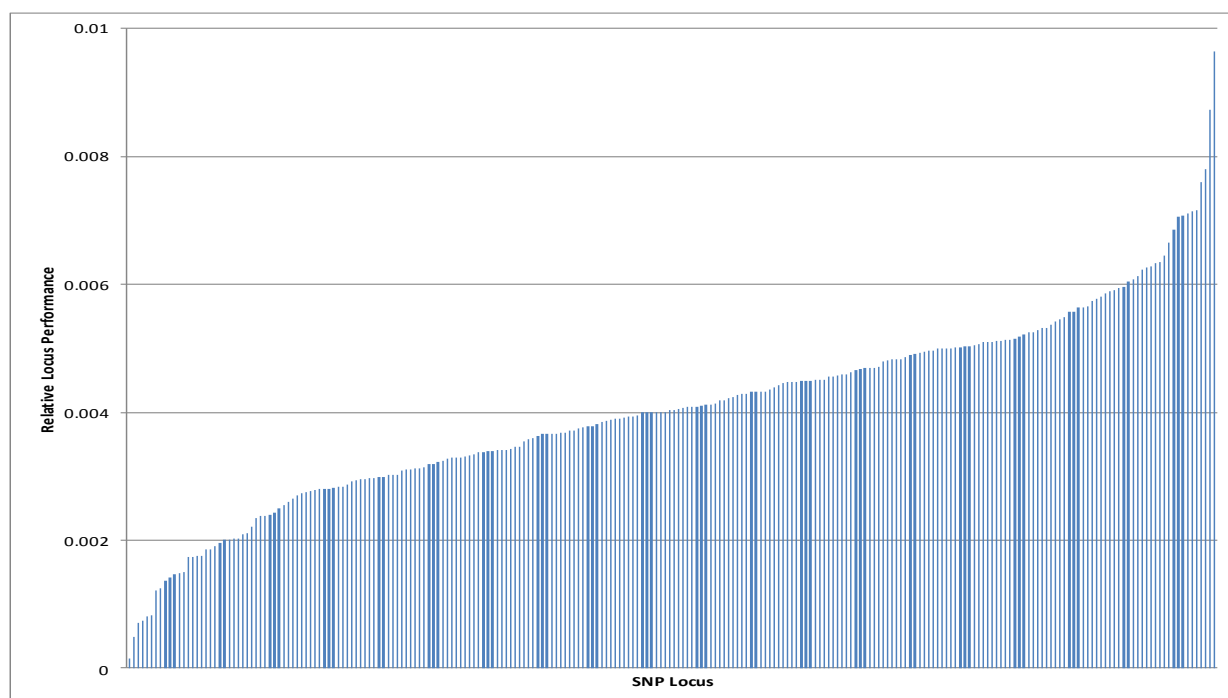


Figure 7. Relative locus performance – autosomal SNPs. Overall performance values, based on coverage, are shown for the 240 autosomal SNPs. Expanded information, including locus names and standard deviation, can be found in Supplemental Figures 1-5.

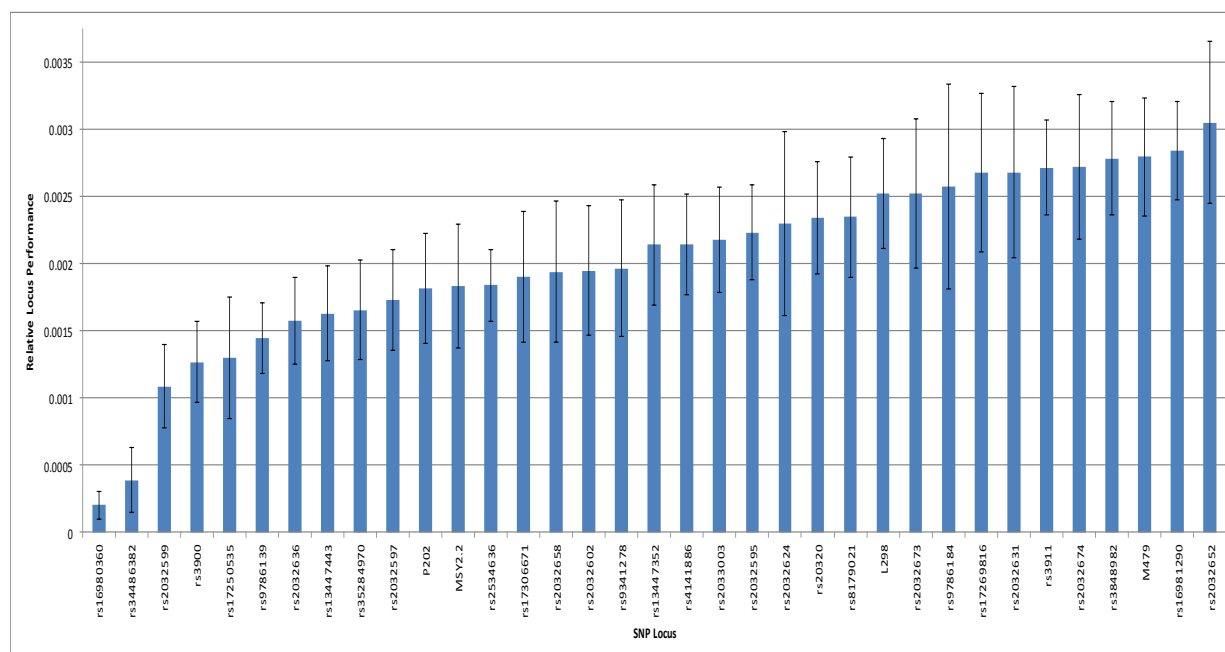


Figure 8. Relative locus performance – Y-chromosome SNPs. Overall performance values, based on coverage, are shown for the 35 Y-chromosome SNPs. Standard deviation is represented by an error bar.

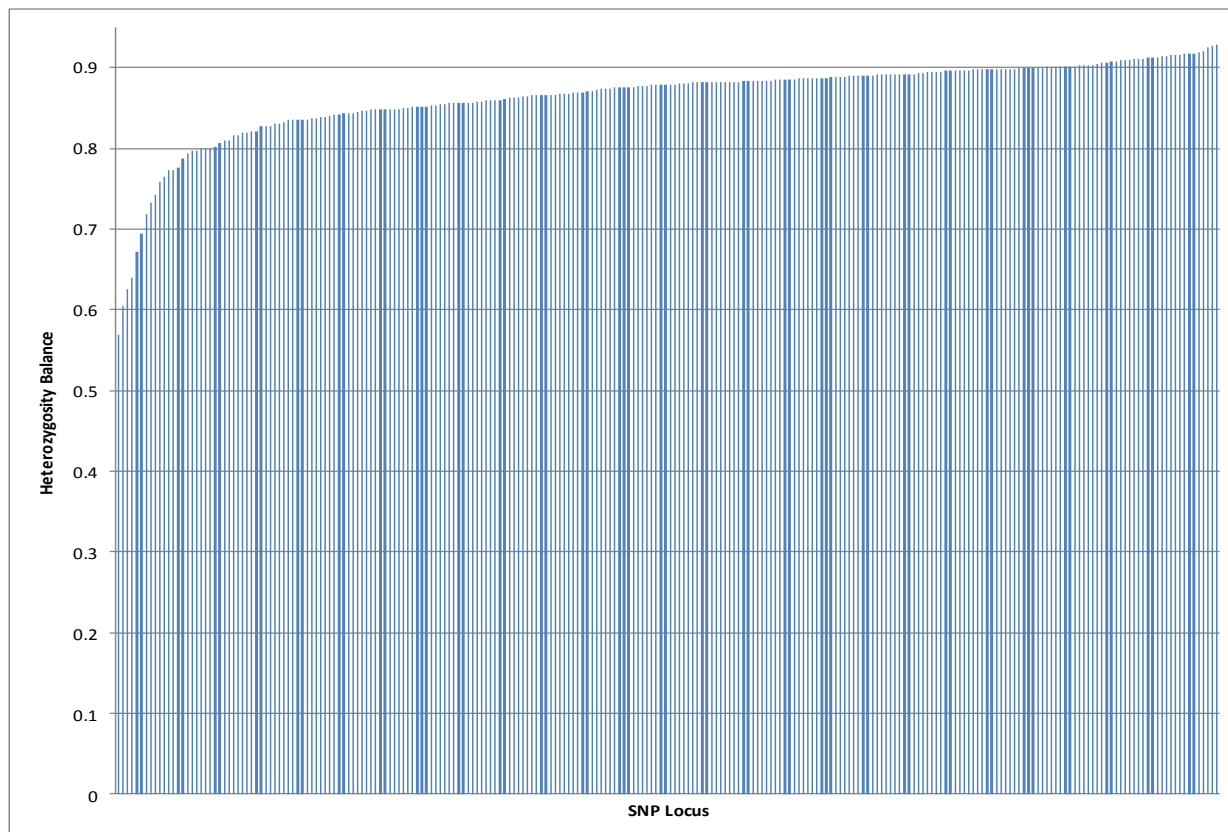


Figure 9. Heterozygosity balance – Autosomal SNPs. ACRs for heterozygous alleles, based on coverage, are shown for the 240 autosomal SNPs. A heterozygosity balance value of 1 indicates perfect balance. Expanded information, including locus names and standard deviation, can be found in Supplemental Figures 6-10.

Overall, there was very little evidence of STR locus dropout (2.7% total) with the panel. For the autosomal STRs, the D14S1434 locus accounted for 85.7% of the total locus dropout (24 out of 28 total dropouts in Reads 1 and 2 combined). For the X-chromosome STRs in females, the DXS6809 locus accounted for 29% of the total dropout (20 out of 69 total dropouts in Reads 1 and 2 combined). The next most prevalent locus dropouts were observed at the DXS10134 and DXS101 loci, which accounted for 20.3% and 15.9% of the total dropout, respectively (14 and 11 out of 69 total dropouts in Reads 1 and 2 combined, respectively). For the X-chromosome STRs in males, locus DXS10134 had the highest dropout with 25.7% of the total dropout (39 out of 152 total dropouts in Reads 1 and 2 combined), and the DXS6809 locus was second, accounting for 19.7%

of the total dropout (30 out of 152 total dropouts in Reads 1 and 2 combined). It should be noted that DXS10134 was one of the larger markers included in the panel. Dropout of large markers can occur when an allele is too long to be completely sequenced (including flanking regions) in a single read. The DXS6789 locus accounted for 11.8% of the total dropout (18 out of 152 total dropouts in Reads 1 and 2 combined), while the DXS101 locus accounted for 10.5% of the total dropout (16 out of 152 total dropouts in Reads 1 and 2 combined). For the Y-chromosome STRs, only 2 loci were notably low performers based on locus dropout: the DYS448 locus, which accounted for 49.5% of the total dropout (54 out of 109 total dropouts in Reads 1 and 2 combined), and locus DYS449, which accounted for 34.9% of the total dropout (38 out of 109 total dropouts in Reads 1 and 2 combined).

The dropout rate (0.14%) was lower for the SNP loci. Most of the low-performing autosomal SNP loci only dropped out in 1 or 2 samples (more likely due to the overall low signal in these samples). The two autosomal SNPs with the highest dropout rates were rs502776, which accounted for 18.6% of the total dropout (11 out of 59 total dropouts), and rs1406945, which accounted for 6.8% of the total dropout (4 out of 59 total dropouts). As for the Y-SNPs, rs16980360 and rs34486382 were the only ones that dropped out in more than one sample, each accounting for 25% of the total dropout (2 out of 8 total dropouts each). It should be noted that allele dropout could not be calculated for SNPs or STRs, as there is no comprehensive system with which to compare all of the markers. However, the low locus dropout rates support the notion that allelic dropout rates are similarly low. Concordance testing may provide some indication of the degree of allele dropout (see below).

Heterozygosity balance values were quite good for the vast majority of loci (i.e., >0.60), with autosomal STRs performing slightly better than X-chromosome STRs (Figures 5 and 6). The majority of identity SNPs also were well-balanced (Figure 8). Heterozygote balance with the capture panel is similar to that of CE-based systems and other commercial PCR-based MPS kits (data not shown).

Some performance analyses initially could have been misconstrued and demonstrate the need for information curation, familiarity with the limitations of the various components of a system, and an appreciation for how these limitations can impact assessment of the performance of a panel. Although not displayed as a low performer in the figures above, the STR locus D5S2500 initially appeared to have a high degree of dropout. However, the cause of the dropout was not due to the probe design. Indeed, the probes for the D5S2500 locus actually performed quite well. The coordinates for the D5S2500 locus were based on consistent data from a number of sources (30-33). These same coordinates were used to design the primers for the D5S2500 locus in the Qiagen Investigator HDplex (34). However, the flanking regions for this locus that were used by STRait Razor were derived from a different source, i.e., Hill *et al.* (35,36). There is discordance between the coordinates used to design the probes in the panel and the coordinates described by Hill *et al.* This discordance was supported by a difference in reported genotype for the 9947A cell line (the Hill *et al.* result was “14,23” and the other groups reported “15,16”) (37,38). STRait Razor was reconfigured to identify flanking regions for the coordinates used in the panel, which then resulted in no evidence of dropout for the D5S2500 locus. At this time, it is not clear which of the two sites is the correct D5S2500 locus. However, such discrepancies do point out how false conclusions can occur with regard to the performance of a marker, and during development of

methodology, some review by alternate means may be warranted. In addition, one should be aware of limitations of STR-calling software, STRait Razor included. Reads for any marker will only be identified if they are configured properly in the software. Therefore, all apparent STR dropouts were reviewed manually.

Other examples of apparent locus dropout that were not due to the chemistry of the system or a sample with overall low signal were at the loci GATA172D05, DXS981, and DYS518. For the GATA172D05 locus, the dropout was due to an error in the STRait Razor configuration file. The allelic definitions lacked 10 bases in the offset value and thus some alleles were not detected. To correct the problem, 10 bases were added to the offset value, which eliminated dropout at this locus. For the DXS981 locus, STRait Razor was configured correctly. However, the configured flanking regions were set unnecessarily far apart, and therefore a number of reads that did span the whole region between the flanks were missed. The distance between the flanking regions was shortened, which completely eliminated dropout at this locus. For the DYS518 locus, the nomenclature/repeat structure was based on old repeat motif data. The definitions for this locus were changed to comport with an alternate nomenclature, and dropout was reduced.

The genotypes detected through the use of the comprehensive panel were highly concordant with those generated by the ForenSeq™ kit. A total of 136 samples were compared, consisting of 39 Caucasians (21 males and 18 females), 50 African Americans (23 males and 27 females), and 47 Hispanics (17 males and 30 females). The ForenSeq™ data allowed for comparisons to be made at 52 STR loci (24 autosomal, 6 X-chromosome, and 22 Y-chromosome), 87 autosomal SNP loci, and the Amelogenin locus. A comparison of genotype calls for STRs

revealed a total of 99.91% concordance. Only 4 loci (D13S317, D21S11, Penta E, and DYS533) displayed discordance, and 3 of these were observed only once and in different samples. The locus D21S11 was discordant in 2 samples. In most of these cases, the detection of a discordant allele was the result of low coverage. In some instances, however, the discordance was caused by differences in the allele-calling methodology utilized by STRait Razor and the software used with the ForenSeq™ kit. These relatively few issues should be easily overcome, and efforts are underway. The SNP genotypes yielded by the comprehensive panel were 100% concordant with those detected by the ForenSeq™ kit.

			STR Loci and Locus Type			
			Autosomal			Y-Chromosome
			D13S317	D21S11	Penta E	DYS533
Samples and Populations	02486	African American	-	(32,33) : (32.2,33)	-	-
	20045		-	-	-	(12) : (8)
	95312		(9,12) : (10,12)	(31,32.2) : (30,31)	-	-
	35796	Hispanic	-	-	(12,15) : (11,12)	-

Table 3. Discordant STR loci. Discordant autosomal and Y-chromosome STR genotypes are listed. The genotype detected using the ForenSeq™ panel is given first, followed by the genotype detected using the comprehensive panel. (-) indicates no discordance.

Population Statistics

Allele counts and frequencies for the 84 STR loci and 275 SNP loci detected by this panel in the dataset of 190 samples can be found in Supplemental Tables 1-48. Tests for departures from Hardy-Weinberg equilibrium (HWE) expectations were performed for the autosomal STRs and the X-chromosome STRs detected in female samples. After Bonferroni correction, only one autosomal locus in the African American population samples displayed departure from HWE:

D14S1434 ($p < 0.0001$). Given that this locus has one of the lowest RLP values and displayed departures from HWE in 2 other populations prior to Bonferroni correction, its departure may be due to low coverage/performance, rather than population substructure or chance. When HWE testing was performed on the X-chromosome STRs in female samples, locus DXS10011 exhibited departure from expectations in the African American population samples after Bonferroni correction ($p < 0.0001$). This particular locus showed detectable departures from HWE in all population group samples prior to Bonferroni correction and also displayed a relatively low RLP value. Two other X-chromosome STRs, DXS101 ($p = 0.001563$) and GATA172D05 ($p = 0.000625$), showed departures from HWE after Bonferroni correction in the Caucasian and Asian population samples, respectively. The DXS101 locus displayed HWE departure in one other population before Bonferroni correction, although the locus did account for a substantial proportion of dropout in both male and female samples. Its departure may be due to low coverage. The GATA172D05 locus did not exhibit departures from HWE in other population groups. Its departure may be due to random chance. There were no autosomal SNP loci that significantly departed from HWE after Bonferroni correction. F_{ST} values for each of the STR and SNP alleles detected by the panel can be found in Supplemental Tables 49-54. The average F_{ST} (θ) values for the autosomal, X-chromosome female, X-chromosome male, and Y-chromosome STRs were 0.0208, 0.0289, 0.0354, and 0.0489, respectively. The average F_{ST} values for the autosomal and Y-chromosome SNPs were 0.0417 and 0.2986, respectively.

Linkage disequilibrium (LD) testing was performed on the STR loci, as well as the SNP markers. For these tests, the STRs were divided into autosomal STRs, X-chromosome female, X-chromosome male, and Y-chromosome STRs. SNPs were divided into autosomal SNPs and Y-

chromosome SNPs. LD testing also was performed on locus pairs consisting of a single autosomal STR and a single autosomal SNP. Genotype preservation was utilized for the LD testing to reduce possible effects of HWE departures. A total of 465 and 325 pairwise comparisons were made for the autosomal and X-chromosome STR loci in each population, respectively. After Bonferroni correction, there were no autosomal, X-chromosome male, or X-chromosome female locus pairs that displayed significant LD (data not shown). Supplemental Table 55 lists the LD p-values for syntenic autosomal STR locus pairs, while Supplemental Tables 56-58 display LD p-values for loci pairs that reside on the same arm of the X-chromosome. A total of 406 pairwise comparisons were made for the Y-chromosome STR loci in each population. Given the lack of recombination, it is expected that a number of Y-chromosome STR loci will show evidence of LD. After Bonferroni correction, there were 5, 6, and 8 locus pairs that displayed significant LD in the Caucasian, African American, and Hispanic population groups, respectively (Table 4). The Asian population group did not display any locus pairs with significant LD after Bonferroni correction, which may be due to small sample size. The fact that there were Y-chromosome STR loci pairs that did not show any detectable LD may indicate some degree of independence at the population level between these markers, or could simply be due to the relatively small size of the datasets. The lack of detectable significant LD for Y-chromosome STR pairs may also be due their relatively higher mutation rates (~ 0.0001 or higher). Supplemental Tables 59-60 list the LD p-values for STR loci pairs that reside on the same arm of the Y-chromosome. LD testing for the autosomal SNP loci consisted of 28,680 comparisons per population. After Bonferroni correction, there were only 15, 13, 14, and 13 loci pairs that exhibited significant LD in the Caucasian, African American, Hispanic, and Asian populations, respectively (Table 5). The number of departures are within 5% of comparisons and thus may be due to chance (although population substructure is another

possible explanation). However, 7 of the pairs displayed significant LD across all 4 population groups: rs10500617/rs10768550, rs2175957/rs8070085, rs2291395/rs4789798, rs2292972/rs3744163, rs2292972/rs689512, rs3744163/rs689512, and rs5746846/rs9606186. These locus pairs are syntenic (Supplemental Tables 61-68). Locus rs10500617 and locus rs10768550 reside 679 bases apart on the p-arm of chromosome 11. Loci rs5746846 and rs9606186 are 287 bases apart on the q-arm of chromosome 22. The rest of the locus pairs are situated on the q-arm of chromosome 17. Loci rs2175957 and rs8070085 are 55,162 bases apart, while loci rs2292972 and rs689512 are 50,086 bases apart. Loci rs2292972 and rs3744163 are 25,929 bases apart, while loci rs3744163 and rs689512 are 24,157 bases apart. Finally, locus rs2291395 and locus rs4789798 are 5,504 bases apart. Significant LD for these locus pairs is likely due to their physical proximity. A total of 595 comparisons were made for LD testing of the Y-chromosome SNP loci. As with the Y-chromosome STRs, LD was expected for these markers. After Bonferroni correction, there were 20, 48, 5, and 2 Y-chromosome SNP locus pairs that displayed significant LD in the Caucasian, African American, Hispanic, and Asian populations, respectively (Table 6). Again, the fact that there were Y-chromosome SNP loci that did not show any detectable LD may indicate some degree of independence between these markers, or simply could be due to sample size. Supplemental Tables 69-72 display LD p-values for Y-chromosome SNP locus pairs that reside on the same arm of the Y-chromosome. Finally, when the autosomal STR and SNP loci were tested together for LD, a total of 36,585 comparisons were made. After Bonferroni correction, there were only 2, 2, 4, and 3 loci pairs that displayed significant LD in the Caucasian, African American, Hispanic, and Asian populations, respectively (Table 7). None of these locus pairs were syntenic (Supplemental Tables 73 and 74), and this level of significance may be due to chance or, possibly, population substructure.

		Population			
		Caucasian	African American	Hispanic	Asian
Locus Pair	DYS19/DYS385a	0.00375	0.009687	<0.000001	0.180938
	DYS19/DYS635	0.309688	0.01375	<0.000001	0.629062
	DYS385a/DYS438	0.000625	<0.000001	0.000625	0.40375
	DYS385a/DYS635	0.03375	0.015	<0.000001	0.079062
	DYS385a/DYS643	<0.000001	0.014063	0.002188	0.004687
	DYS385b/DYS438	0.0275	<0.000001	0.180938	0.518437
	DYS390/DYS437	0.176875	<0.000001	0.121563	0.2475
	DYS390/DYS438	0.188125	<0.000001	0.001563	0.888125
	DYS391/DYS576	<0.000001	0.753125	1	0.798125
	DYS391/DYS635	<0.000001	0.233125	0.2725	0.105313
	DYS392/DYS438	<0.000001	<0.000001	<0.000001	0.1125
	DYS392/DYS576	<0.000001	0.955313	0.424687	0.881875
	DYS438/DYS522	0.019062	0.02	<0.000001	0.77875
	DYS438/DYS635	0.154062	0.001875	<0.000001	0.666562
	DYS438/DYS643	0.01875	<0.000001	0.006875	0.18875
	DYS522/DYS643	0.033438	0.36625	<0.000001	0.278438
	DYS635/DYS643	0.01125	0.15625	<0.000001	0.186875

Table 4. Significant LD results – Y-chromosome STRs. LD p-values are given for the specified loci pair in which a significant value was observed in at least one population group. Values in bold are significant, after Bonferroni correction ($p < 0.000123$).

		Population			
		Caucasian	African American	Hispanic	Asian
Locus Pair	rs1000322/rs1542931	0.38	0.891875	<0.000001	0.36375
	rs1028484/rs430046	0.809688	0.035	<0.000001	0.403438
	rs1031825/rs877228	0.284687	0.05625	<0.000001	0.297812
	rs1031825/rs927628	0.143437	0.836875	<0.000001	0.86
	rs10474080/rs2056277	0.019375	0.777188	0.484063	<0.000001
	rs10500617/rs10768550	<0.000001	<0.000001	<0.000001	<0.000001
	rs1075665/rs892503	0.325	0.446562	0.234375	<0.000001
	rs1105459/rs8070085	<0.000001	0.252188	0.073438	0.159375
	rs123714/rs500282	0.187188	<0.000001	0.918125	0.565625
	rs1355366/rs1498553	0.598437	0.94625	0.8575	<0.000001
	rs1490413/rs2399332	0.952812	<0.000001	0.706875	0.413438
	rs1498553/rs6444724	0.30125	0.971562	0.440312	<0.000001
	rs1554472/rs560681	0.814063	<0.000001	0.605938	0.186875
	rs1574542/rs2311046	<0.000001	0.260313	0.931562	0.0525
	rs1858958/rs738532	0.192812	<0.000001	0.56625	1
	rs2175957/rs8070085	<0.000001	<0.000001	<0.000001	<0.000001
	rs2255301/rs445251	0.361875	0.808125	0.155938	<0.000001
	rs2272998/rs892503	0.174375	0.760625	<0.000001	0.876563
	rs2291395/rs4789798	<0.000001	<0.000001	<0.000001	<0.000001
	rs2292972/rs3744163	<0.000001	<0.000001	<0.000001	<0.000001
	rs2292972/rs689512	<0.000001	<0.000001	<0.000001	<0.000001
	rs2311046/rs907100	<0.000001	0.69625	0.002812	0.697187
	rs2342747/rs3744163	0.391875	0.831562	0.165625	<0.000001
	rs2388618/rs354439	0.36	0.539062	<0.000001	0.33125
	rs2665846/rs740910	0.045625	<0.000001	0.310625	0.734688
	rs2920816/rs9866013	<0.000001	0.329688	0.525625	0.45375
	rs354439/rs914165	<0.000001	0.276875	0.399375	0.557187
	rs3744163/rs689512	<0.000001	<0.000001	<0.000001	<0.000001
	rs4643330/rs772436	<0.000001	0.005625	0.625313	0.482812
	rs464663/rs997750	0.766563	0.695937	<0.000001	0.64375
	rs5746846/rs9606186	<0.000001	<0.000001	<0.000001	<0.000001
	rs627119/rs6444724	0.283125	<0.000001	0.495312	0.530312
	rs6444724/rs6955448	<0.000001	0.461875	0.215938	0.848125
	rs7229946/rs916388	<0.000001	0.4525	0.974688	0.545

Table 5. Significant LD results – Autosomal SNPs. LD p-values are given for the specified loci pair in which a significant value was observed in at least one population group. Values in bold are significant, after Bonferroni correction ($p < 0.0000017$).

		Population			
		Caucasian	African American	Hispanic	Asian
Locus Pair	L298/M479	1	<0.000001	1	1
	rs16980360/rs2032652	0.002812	<0.000001	0.056563	1
	rs16980360/rs2032658	0.093438	<0.000001	1	1
	rs16980360/rs3848982	0.003125	<0.000001	0.059062	1
	rs16980360/rs3900	0.1025	<0.000001	0.350938	1
	rs16980360/rs4141886	0.005	<0.000001	0.057813	1
	rs16980360/rs9786139	0.005938	<0.000001	0.111562	1
	rs16980360/rs9786184	0.134375	<0.000001	1	1
	rs17306671/rs2032597	<0.000001	0.043437	0.220312	1
	rs17306671/rs2032624	<0.000001	1	0.0925	1
	rs17306671/rs2032631	<0.000001	1	0.005938	1
	rs17306671/rs2032658	<0.000001	1	0.092188	1
	rs17306671/rs9786184	<0.000001	1	0.084062	1
	rs2032624/rs2032631	<0.000001	<0.000001	0.011875	1
	rs2032624/rs2032652	0.100625	<0.000001	1	1
	rs2032624/rs2032658	<0.000001	<0.000001	0.000313	1
	rs2032624/rs2033003	<0.000001	<0.000001	0.013125	1
	rs2032624/rs3900	<0.000001	<0.000001	0.012188	1
	rs2032624/rs4141886	0.108438	<0.000001	1	1
	rs2032624/rs9786139	0.100937	<0.000001	0.488438	1
	rs2032624/rs9786184	<0.000001	<0.000001	<0.000001	1
	rs2032631/rs2032652	0.102813	<0.000001	0.325313	1
	rs2032631/rs2032658	<0.000001	<0.000001	0.010625	1
	rs2032631/rs2033003	<0.000001	<0.000001	<0.000001	1
	rs2032631/rs3848982	0.11125	<0.000001	0.320625	1
	rs2032631/rs3900	<0.000001	<0.000001	<0.000001	1
	rs2032631/rs9786139	0.105	<0.000001	0.093438	1
	rs2032631/rs9786184	<0.000001	<0.000001	0.013125	1
	rs2032652/rs2032658	0.102188	<0.000001	1	1
	rs2032652/rs2033003	0.095937	<0.000001	0.341562	<0.000001
	rs2032652/rs3848982	0.002812	<0.000001	0.058438	0.235313
	rs2032652/rs4141886	0.005625	<0.000001	0.054375	0.23125
	rs2032652/rs9786139	0.005	<0.000001	0.10875	0.000313
	rs2032652/rs9786184	0.125312	<0.000001	1	1
	rs2032658/rs2033003	<0.000001	<0.000001	0.010937	1
	rs2032658/rs3900	<0.000001	<0.000001	0.010937	1
	rs2032658/rs4141886	0.10875	<0.000001	1	1
	rs2032658/rs9786139	0.11	<0.000001	0.485625	1
	rs2032658/rs9786184	<0.000001	<0.000001	<0.000001	1
	rs2033003/rs3848982	0.1075	<0.000001	0.337187	0.238437
	rs2033003/rs3900	<0.000001	<0.000001	<0.000001	0.000625
	rs2033003/rs4141886	0.098125	<0.000001	0.344375	0.224688
	rs2033003/rs9786139	0.100937	0.000625	0.0925	<0.000001
	rs2033003/rs9786184	<0.000001	<0.000001	0.012812	1
	rs3848982/rs3900	0.0975	<0.000001	0.3325	0.230313
	rs3848982/rs4141886	0.003125	<0.000001	0.063125	0.059062
	rs3848982/rs9786139	0.005	<0.000001	0.111562	0.230625
	rs3848982/rs9786184	0.125938	<0.000001	1	1
	rs3900/rs4141886	0.095937	<0.000001	0.329688	0.234063
	rs3900/rs9786139	0.093125	<0.000001	0.094687	0.000313
	rs3900/rs9786184	<0.000001	<0.000001	0.011875	1
	rs4141886/rs9786139	0.00375	<0.000001	0.110937	0.223125
	rs4141886/rs9786184	0.125938	<0.000001	1	1
	rs9786139/rs9786184	0.138125	<0.000001	0.475	1

Table 6. Significant LD results – Y-chromosome SNPs. LD p-values are given for the specified locus pairs in each

population group. Values in bold are significant, after Bonferroni correction ($p < 0.000084$).

		Population			
		Caucasian	African American	Hispanic	Asian
Locus Pair	D16S539/rs354439	0.45625	<0.000001	0.212812	0.464062
	D1S1627/rs9606186	<0.000001	0.410625	0.74375	0.6875
	D21S11/rs481736	0.41125	0.965938	<0.000001	1
	D21S11/rs927628	0.843437	0.052812	0.9325	<0.000001
	D2S1776/rs500282	0.41625	0.929375	<0.000001	0.228437
	D2S441/rs1413212	0.169687	0.329375	0.631563	<0.000001
	D6S474/rs590162	0.644687	0.2725	<0.000001	0.072188
	FGA/rs987640	0.765625	<0.000001	0.766563	0.693438
	PENTAD/rs1482650	0.478438	0.445312	<0.000001	0.105313
	PENTAD/rs730249	0.295	0.976562	0.910625	<0.000001
	VWA/rs1000322	<0.000001	0.262813	0.702812	0.072188

Table 7. Significant LD results – Autosomal SNP and STR pairs. LD p-values are given for the specified loci pair in which a significant value was observed in at least one population group. Values in bold are significant, after Bonferroni correction ($p < 0.0000014$).

Each sample exhibited a unique Y-chromosome STR haplotype (Table 8). The haplotype diversity values for each population were thus >0.99 . The Y-chromosome STR haplotypes detected by the panel were used to determine haplogroups for each individual represented in the dataset (Table 8), using Haplogroup Predictor (39). These haplogroups are largely consistent with their respective population groups.

CONCLUSIONS

The data herein support that a capture-based approach can produce robust data for typing reference samples. A large set of markers and different types of markers can be typed simultaneously; thus the potential for gaining substantially more data in a single analysis is demonstrated. Given the increase of typing data afforded with MPS, the majority (if not all) of the profiles from evidence samples can be compared with reference profiles, regardless of the number and types of markers used in the evidentiary analyses. The profiles generated by MPS technology will be compatible with existing reference data, and the more comprehensive set of markers made possible through the use of MPS can foster investigations. A very few loci were low performers, and their signals likely could be increased, if desired, by increasing probe density in the design phase. Alternatively, these few markers can be removed and replaced with better performers. While locus balance is not required for a panel of markers typed by MPS, balance can improve sample multiplexing or at least allow better throughput prediction.

One motivation for using a capture-based assay was that the vagaries of PCR would impact the results less. However, the data indicate that the performance and artifacts observed with a PCR enrichment method persist with our capture-based approach. There is locus-to-locus coverage variation; stutter does occur (data not shown) mostly due to an amplification stage prior to sequencing; heterozygote balance is similar; and low level noise exists (which can be filtered out). The artifacts of locus-to-locus signal difference, stutter, and noise are not new to DNA typing and can be managed in a similar fashion as they are with CE-based systems. However, sequence data

obtained with MPS technology affords additional information that may assist in characterization of artifacts.

With the increased number of forensic markers that can be incorporated in this panel, including lineage markers such as Y-chromosome STRs, indirect searches can be performed, if desired. The success rate of familial searching would likely increase, as the sheer abundance of markers will provide more robust associations and substantially reduce the number of adventitious associations. Increased marker typing also will have a substantial impact on missing persons identification testing.

Although not part of this study, but worth considering for future studies, is that a probe-based capture system may be better suited for typing degraded samples than a PCR enrichment approach (40,41). Primers define the size of a PCR amplicon. If DNA is degraded, such that the fragments are too small to generate amplicons, no PCR product will be generated. However, a probe capture system is not as limited due to the size of the fragments of DNA in a sample. Indeed, the probe design could be increased readily from “ADJACENT” to “OVERLAPPING” to enhance capture with challenged samples. Carpenter *et al.* (40) and Templeton *et al.* (41) have described novel capture procedures that enrich highly degraded endogenous ancient genomic DNA. However, to make the current probe-based system in our study practical for analyzing challenged samples the amount of input DNA will need to be substantially reduced from the current amount of 50 ng. During the course of this project the input DNA was reduced initially from 500 ng to 50 ng, which is an order of magnitude change in template requirements. With the rapid advancements

in MPS technologies and chemistries, it is anticipated that the amount of input DNA required for capture based approaches will continue to decrease.

Lastly, our data indicate that development of the capture panel was relatively easy compared with our experience with PCR multiplex design and required fewer resources than PCR-based systems. The design is simple and the panel was relatively easy to use.

CONFLICT OF INTEREST

C.P. Davis is employed by Illumina, Inc.

ACKNOWLEDGMENTS

This work was supported in part by award no. 2012-DN-BX-K033, awarded by the National Institute of Justice, Office of Justice Programs, U.S. Department of Justice. The opinions, findings, and conclusions or recommendations expressed in this publication are those of the authors and do not necessarily reflect those of the U.S. Department of Justice. The authors also would like to thank Illumina, Inc. for its support during this study.

REFERENCES

1. Budowle, B. and Eisenberg, A.J. 2007. Forensic Genetics. In: Emery and Rimoin's Principles and Practice of Medical Genetics, fifth edition, Vol. 1, Rimoin, D.L., Connor, J.M., Pyeritz, R.E., and Korf, B.R. (eds.), Elsevier, Philadelphia, pp.501-517.
2. Budowle, B., Planz, J.V., Campbell, R., and Eisenberg, A.J. 2005. Molecular diagnostic applications in forensic science. In: Molecular Diagnostics, Patrinos, G. and Ansorge, W., (eds.), Elsevier, Amsterdam, pp. 267-280.
3. Budowle, B. and van Daal, A. 2008. Forensically relevant SNP classes. *Biotechniques* 44:603-610.
4. Budowle, B., Moretti, T.R., Niezgoda, S.J., and Brown, B.L. 1998. CODIS and PCR-based short tandem repeat loci: Law enforcement tools. In: Second European Symposium on Human Identification 1998, Promega Corporation, Madison, Wisconsin, pp. 73-88.
5. Martin, P.D., Schmitter, H., and Schneider, P.M. 2001. A brief history of the formation of DNA databases in forensic science within Europe. *Forens. Sci. Int.* 119(2):225-231.
6. CODIS-NDIS Statistics: <http://www.fbi.gov/about-us/lab/biometric-analysis/codis/ndis-statistics>
7. Budowle, B.: Familial searching: extending the investigative lead potential of DNA typing. Profiles in DNA 13(2), 2010, Available at: www.promega.com/profiles/1302/1302_07.html.
8. Ge, J., Chakraborty, R., Eisenberg, A. and Budowle, B. 2011. Comparisons of the familial DNA database searching policies. *J. Forens. Sci.* 56(6):1448-1456.
9. Ge, J., Eisenberg, A., and Budowle, B. 2012. Developing criteria and data to determine best options for expanding the core CODIS loci. *BMC Investigative Genetics* 3:1.
10. Hares, D.R. 2015. Selection and implementation of expanded CODIS core loci in the United States. *Forensic Sci. Int. Genet.* 17:33-34.
11. Hares, D.R. 2011. Expanding the CODIS core loci in the United States. *Forensic Sci. Int. Genet.* Doi:10.1016/j.fsigen.2011.04.012.
12. Rothberg JM, Hinz W, Rearick TM, et al. 2011. An integrated semiconductor device enabling non-optical genome sequencing. *Nature* 475(7356):348-352.
13. Adessi, C., Matton, G., Ayala, G., Turcatti, G., Mermoud, J.J., Mayer, P., and Kawashima, E. 2000. Solid phase DNA amplification: characterisation of primer attachment and amplification mechanisms. *Nucleic Acids Res.* 28(20):E87.

14. Brenner, S., Johnson, M., Bridgham, J., et al. 2000. Gene expression analysis by massively parallel signature sequencing (MPSS) on microbead arrays. *Nat. Biotechnol.* 18:630-634.
15. Brenner, S., Williams, S.R., Vermaas, E.H., Storck, T., Moon, K., McCollum, C., Mao, J.I., Luo, S., Kirchner, J.J., Eletr, S., DuBridge, R.B., Burcham, T., and Albrecht, G. 2000. In vitro cloning of complex mixtures of DNA on microbeads: physical separation of differentially expressed cDNAs. *Proc. Natl. Acad. Sci. USA* 97:1665-1670.
16. Holt, K.E., Parkhill, J., Mazzoni, C.J., Roumagnac, P., Weill, F.X., Goodhead, I., Rance, R., Baker, S., Maskell, D.J., Wain, J., Dolecek, C., Achtman, M., and Dougan, G. 2008. High-throughput sequencing provides insights into genome variation and evolution in *Salmonella Typhi*. *Nat. Genet.* 40:987-993.
17. Margulies, M., Egholm, M., Altman, W.E., et al. 2005. Genome sequencing in microfabricated high-density picolitre reactors. *Nature* 437:376-380.
18. Quail, M.A., Kozarewa, I., Smith, F., Scally, A., Stephens, P.J., Durbin, R., Swerdlow, H., and Turner, D.J. 2008. A large genome center's improvements to the Illumina sequencing system. *Nat. Methods* 5:1005-1010.
19. Van Tassell, C.P., Smith, T.P., Matukumalli, L.K., Taylor, J.F., Schnabel, R.D., Lawley, C.T., Haudenschild, C.D., Moore, S.S., Warren, W.C., and Sonstegard, T.S. 2008. SNP discovery and allele frequency estimation by deep sequencing of reduced representation libraries. *Nat. Methods* 5:247-252.
20. Wheeler, D.A., Srinivasan, M., Egholm, M., Shen, Y., Chen, L., McGuire, A., He, W., Chen, Y.J., et al. 2008. The complete genome of an individual by massively parallel DNA sequencing. *Nature* 452:872-876.
21. Ottaviani, E., Vernarecci, S., Asili, P., Agostino, A., Montagna, P. 2014. Preliminary assessment of the prototype Yfiler® Plus kit in a population study of Northern Italian males. *Int. J. Legal Med.* Doi:10.1007/s00414-014-1078-x.
22. Zeng, X., King, J., Hermanson, S., Patel, J., Storts, D.R., and Budowle, B. An evaluation of the PowerSeq™ Auto system: a multiplex short tandem repeat marker kit compatible with massively parallel sequencing. *Forensic Sci. Int. Genet.* (submitted).
23. Illumina® ForenSeq™ information: <http://www.illumina.com/products/forenseq-dna-signature-kit.html>
24. Adey, A., Morrison, H.G., Asan, Xun X., Kitzman, J.O., Turner, E.H., Stackhouse, B., MacKenzie, A.P., Caruccio, N.C., Zhang, X., and Shendure, J. 2010. Rapid, low-input, low-bias construction of shotgun fragment libraries by high-density *in vitro* transposition. *Genome Biol.* 11:R119.

25. McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., DePristo, M.A. 2010. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* 20(9):1297-1303.
26. Warshauer, D.H., King, J.L., and Budowle, B. 2015. STRait Razor v2.0: the improved STR Allele Identification Tool – Razor. *Forensic Sci. Int. Genet.* 14:182-186.
27. Excoffier, L. and Lischer, H.E.L. 2010. Arlequin suite ver 3.5: A new series of programs to perform population genetics analyses under Linux and Windows. *Mol. Ecol. Resour.* 10:564-567.
28. GDA (Genetic Data Analysis): <http://www.eeb.uconn.edu/people/plewis/downloads/gda-1.1.win32.zip>
29. Tajima, F. 1989. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* 123(3):585-595.
30. Marshfield Clinic: <http://research.marshfieldclinic.org/genetics/GeneticResearch/sets/Set10PrimerSequences.htm>
31. NCBI: <http://www.ncbi.nlm.nih.gov/nuccore/G08468>
32. Mentype Chimera: http://www.biotype.de/fileadmin/user/MANUALS/190213_Manual_MentypeChimera_CE.pdf
33. Thiede, C., Bornhauser, M., and Ehninger, G. 2004. Evaluation of STR informativity for chimerism testing – comparative analysis of 27 STR systems in 203 matched related donor recipient pairs. *Leukemia* 18: 248–254.
34. Qiagen Investigator HDplex Kit:
https://www.google.com/url?sa=t&rct=j&q=&esrc=s&source=web&cd=3&cad=rja&uact=8&ved=0CC8QFjAC&url=http%3A%2F%2Fwww.qiagen.com%2Fresources%2Fdownload.aspx%3Fid%3D7d1661bd-a47b-4b19-a882-357a61b48c64%26lang%3Den&ei=f9tjVI7GJoiYyASKI4GQDw&usg=AFQjCNEXCa2eHD1yg5bzJl0AbkyB9k_4gg&sig2=uJ7RG8yki7pvbp_w4aVFkw&bvm=bv.79189006,d.aWw
35. Hill, C.R., Kline, M.C., Coble, M.D., and Butler, J.M. 2008. Characterization of 26 miniSTR loci for improved analysis of degraded DNA samples. *J. Forens. Sci.* 53(1):73-80.
36. Hill, C.R., Butler, J.M., and Vallone, P.M. 2009. A 26plex autosomal STR assay to aid human identity testing. *J. Forens. Sci.* 54(5):1008-1015.
37. STRbase: http://www.cstl.nist.gov/biotech/strbase/miniSTR/miniSTR_NC_loci_types.htm

38. Becker D, Bender K, Edelmann J, Götz F, Henke L, Hering S, Hohoff C, Hoppe K, Klintschar M, Muche M, Rolf B, Szibor R, Weirich V, Jung M, and Brabetz W. 2007. New alleles and mutational events at 14 STR loci from different German populations. *Forens. Sci. Int. Genet.* 1(3-4):232-237.
39. Haplogroup Predictor: <http://www.hprg.com/hapest6/files/hapest6.htm?order=orig>
40. Carpenter, M.L., Buenrostro, J.D., Valdiosera, C., Schroeder, H., Allentoft, M.E., Sikora, M., et al. 2013. Pulling out the 1%: whole-genome capture for the targeted enrichment of ancient DNA sequencing libraries. *Amer. J. Hum. Genet.* 93(5):852-864.
41. Templeton, J.E., Brotherton, P.M., Llamas, B., Soubrier, J., Haak, W., Cooper, A., and Austin, J.J. 2013. DNA capture and next-generation sequencing can recover whole mitochondrial genomes from highly degraded samples for human identification. *Investigative Genetics*, 4(1):1-13.

SECTION 4

Detection of Intra-Repeat Nucleotide Variation in Short Tandem Repeat Alleles Through the Use of the Comprehensive Massively Parallel Sequencing Panel

CHAPTER 5

Novel Y-Chromosome Short Tandem Repeat Variants Detected Through the Use of Massively Parallel Sequencing

Submitted to Genomics, Proteomics & Bioinformatics
May 2015

David H. Warshauer
Jennifer D. Churchill
Nicole Novroski
Jonathan L. King
Bruce Budowle

ABSTRACT

Massively parallel sequencing (MPS) technology is capable of determining the sizes of short tandem repeat (STR) alleles as well as their individual nucleotide sequences. Thus, single nucleotide polymorphisms (SNPs) within the repeat regions of STRs and variations in the pattern of repeat units in a given repeat motif can be used to differentiate alleles of the same length. In this study, MPS was used to detect 28 forensically-relevant Y-chromosome STRs in a set of 41 DNA samples from the 3 major U.S. population groups. The resulting sequence data, which were analyzed with STRait Razor v2.0, revealed 46 unique allele sequence variants that have not been previously reported. Of these, 28 sequences were variations of documented sequences resulting from the presence of intra-repeat SNPs or alternative repeat unit patterns. Interestingly, 2 of the most frequently observed variants were found only in African American samples. The remaining 18 variants represented allele sequences for which there were no published data with which to compare. These findings illustrate the great potential of MPS with regard to increasing the resolving power of STR typing and emphasize the need for sample population characterization of STR alleles.

KEYWORDS: Y-STR · Sequence polymorphism · Allele variants · Massively parallel sequencing · Nextera™ · STRait Razor

INTRODUCTION

Short tandem repeat (STR) markers located on the Y-chromosome (Y-STRs) are extremely useful because of a lack of recombination. Barring mutation, all paternally-related males share the same Y-STR haplotype. As a result, Y-STRs are used in genealogical and evolutionary studies, forensic genetics casework (including paternity testing to determine the biological father of a particular male child), missing persons cases (where the Y-STR haplotype can serve as an extended reference profile for a given paternal lineage), and analyses of mixture evidence where there is substantially more female DNA than male DNA. Indeed, the variety of uses for Y-STR markers has made them the object of extensive research and application within the scientific community.

Given the value of STR markers in identity testing, efforts are underway to increase the power of discrimination associated with their respective typing and analysis methods. Primarily, an increase in power of discrimination has been accomplished through the introduction of new, highly polymorphic STRs and by developing larger multiplex panels [1-4]. Discrimination power also may be increased by further characterization beyond nominal length of the alleles at extant loci. STR alleles are typically characterized by the number of units in their repeat motifs, a distinction commonly determined by size separation by capillary electrophoresis (CE). However, other detection methods, such as Sanger sequencing and mass spectrometry, have been used to determine not only the size of STR alleles, but the nucleotide composition of their repeat regions. The emergence of massively parallel sequencing (MPS) technologies improved upon this principle by allowing for the detection of a substantially larger amount of genetic sequence information with a higher throughput, lower cost, and greater ease-of-use than previous

methods. Studies involving each of these approaches have resulted in the detection of intra-repeat single nucleotide polymorphisms (SNPs) and novel repeat motif variants that allow for a greater level of distinction than that of traditional CE methods [5-10]. For instance, two individuals with the same nominal allele(s) (based on length) at a certain locus potentially may be distinguished by MPS if the nucleotide sequence of the allele differs between them. This level of scrutiny may prove invaluable in the deconvolution of genetic mixtures and also could add additional information for evolutionary studies.

In this proof-of-principle study, MPS was used to determine the repeat sequences of 28 forensically-relevant Y-STRs across a dataset of three major US populations (n=41): Caucasian, Hispanic, and African Americans. These sequence data revealed several intra-repeat SNPs and allelic variants that have not been documented previously. The novel variants described herein are indicative of the potential of MPS with regard to identifying additional genetic diversity of Y-STRs, and support that more in depth population studies are warranted.

MATERIALS AND METHODS

Samples and Extraction

Following the University of North Texas Health Science Center Institutional Review Board approval, DNA was extracted from whole blood samples from 41 unrelated individuals, consisting of 12 Caucasian males, 16 Hispanic males, and 13 African American males. These populations were selected because they represent the three major populations in the geographic region.

Extraction was performed using the Qiagen® QIAamp™ DNA Mini Kit (Qiagen, Hilden, Germany), according to the manufacturer's suggested protocol.

Panel Design

The Nextera™ Rapid Capture Custom Enrichment panel employed in this study was designed using the Illumina® DesignStudio sequencing assay design tool. Nextera™ Rapid Capture chemistry (Illumina, Inc., San Diego, CA) is based on enzymatic “tagmentation” and probe-based “capture” enrichment. Custom oligonucleotide probes were designed to detect the following 28 forensically-significant Y-STRs: DYS19, DYS385, DYS389I/II, DYS390, DYS391, DYS392, DYS393, DYS437, DYS438, DYS439, DYS448, DYS449, DYS456, DYS458, DYS460, DYS481, DYS505, DYS518, DYS522, DYS533, DYS549, DYS570, DYS576, DYS612, DYS635, DYS643, and GATA-H4. Multiple probes were used for each Y-STR to improve enrichment efficiency.

Probes (80 bases in length) for the Nextera Rapid Capture Custom Enrichment Kit were designed using Design Studio (Illumina), a freely-available software. The STRs were tabulated including details regarding chromosomal positioning, target selection (Full Region), probe density requirements (due to the alignment-specific requirements of STRs, density of these markers was set at ‘ADJACENT’), and marker information. Marker data then were uploaded to Design Studio v1.5 and probes were generated under the default conditions (with hg19 for probe reference).

Quantification and Normalization

50 ng of genomic DNA were used as the input amount for typing. To bring the 41 extracted DNA samples to the desired input concentration of 5 ng/ μ L for the Nextera™ Rapid Capture Custom Enrichment protocol, the quantity of each DNA sample was determined using the Qubit® fluorometric quantification method (Thermo Fisher Scientific, Waltham, MA) and normalized to 10 ng/ μ L with a 10 mM Tris-HCl solution at pH 8.5. The samples then were quantified again and normalized in the same manner to a final concentration of 5 ng/ μ L, to ensure that the proper amount of genomic DNA would be used for the library preparation process.

Library Preparation

As required by the Nextera™ Rapid Capture Custom Enrichment protocol, 10 μ L of each normalized sample were used for library preparation, for a total of 50 ng of genomic DNA per sample. The samples first underwent tagmentation by the Nextera™ transposome, whereby the samples are enzymatically cleaved and bound to sequencing adapters [11], at 58° C in an Applied Biosystems® GeneAmp® PCR System 9700 thermal cycler (Thermo Fisher Scientific). The tagmented samples then were purified via 2 magnetic bead-based 80% ethanol washes, and the fragment sizes of a small subset of these samples were analyzed using the Agilent® 2200 TapeStation™ (Agilent Technologies, Inc., Santa Clara, CA) to ensure that the tagmentation process was successful. Dual Nextera™ sequencing indices then were attached to each of the tagmented samples by amplification in an Eppendorf® Mastercycler® Pro S thermal cycler (Eppendorf, Hamburg, Germany), using the following parameters: 72° C for 3 min, 98° C for 30

sec, 10 cycles of 98° C for 10 sec, 60° C for 30 sec, and 72° for 30 sec, a final extension at 72° C for 5 min, and a final hold at 10° C. Following bead-based amplification cleanup with 80% ethanol, each indexed sample was quantified using the Qubit® platform. The samples then were normalized and pooled for sequencing, 12 at a time, such that each library contained 500 ng of each uniquely indexed sample, for a total of 6,000 ng of genomic DNA per pool. It should be noted that all libraries consisted of 12 samples; some libraries included female samples that do not contribute data for this study. The pooled libraries were hybridized once to the custom oligonucleotide probes in an Eppendorf® Mastercycler® Pro S thermal cycler, using the following parameters: 95° C for 10 min, 18 cycles of 1 min incubations, starting at 94° C, then decreasing 2° C per cycle, and a final hold at 58° C for approximately 12 hours. A streptavidin bead-based cleanup step was performed wherein the libraries were washed twice for 30 minutes with an enrichment wash solution at 50° C. A second hybridization then was performed, using the same thermal cycling parameters, except that the final hold at 58° C was extended to approximately 20 hours. Following a second heated streptavidin bead-based cleanup, the libraries underwent two additional magnetic bead-based washes with 80% ethanol. The libraries then were enriched through amplification in an Eppendorf® Mastercycler® Pro S thermal cycler, using the following parameters: 98° C for 30 sec, 12 cycles of 98° C for 10 sec, 60° C for 30 sec, and 72° C for 30 sec, a final extension at 72° C for 5 min, and a final hold at 10° C. A final magnetic bead-based cleanup procedure was performed, consisting of 2 washes with 80% ethanol, and the libraries were quantified using the Qubit® platform. Following quantification, each library was analyzed on the Agilent® 2200 TapeStation™ to determine the average size of the enriched fragments.

The concentration and size, in base pairs, of the Nextera™ Rapid Capture Custom Enrichment libraries were used to determine their molarity. To prepare for sequencing on the MiSeq™ (Illumina), each library was normalized to 2 nM using a solution of 10 mM tris-HCl buffer at pH 8.5 with 0.1% Tween 20. Illumina®'s library preparation guidelines for the MiSeq™ were followed, and the concentration of each library was adjusted to 12 pM using chilled HT1 buffer. Paired-end sequencing was performed, with a read length of 250 bases.

STRait Razor v2.0 [12] was used to analyze the FASTQ files produced by the MiSeq™ for each sample. STRait Razor's STR allele detection method allows it to type alleles found in raw sequence data based on their length, while retaining their individual nucleotide sequences. The sequence data produced by STRait Razor for each of the targeted Y-STRs across all samples were analyzed using STRait Razor Sequence Analysis [12], and the unique sequences associated with each allele were identified with the STRait Razor Unique Sequences Compiler [13]. These unique sequences then were compared to the known sequences for those alleles that have been published in STRBase and the current literature [5, 6, 14-19].

Y-STR haplogroups were predicted from the repeat lengths of the STR alleles comprising the haplotype using Haplogroup Predictor (<http://www.hprg.com/hapest5/>).

RESULTS

Since nanogram and subnanogram quantities of input DNA can be typed by MPS, PCR enrichment has become the method of choice for studies involving forensic applications. However, this study employed a capture enrichment approach. The Nextera™ library chemistry was selected initially because no PCR amplification is required. Therefore, primer binding site mismatch issues would not impact multiplex design or the amplification success. It was hypothesized that a dense probe design would ensure capture of the target loci. In addition, PCR-generated errors would be reduced and thus minimize potential artifacts. Lastly, laying a foundation of sequence data with an alternate enrichment system could be useful when full validation studies are undertaken.

All 28 Y STR loci were detected with the approach described herein. The coverage ranged from 0 to 1493X, with a mean coverage of 9 to 387X. The lowest performing markers were DYS448 (mean 9X), DYS449 (mean 33X), DYS518 (mean 34X), DYS389II (mean 37X), and DYS505 (mean 38X), while the highest were DYS643 (mean 322X), DYS391 (mean 333X), and DYS438 (mean 387X).

A total of 46 unique Y-STR allele sequences that have not been previously published were detected across the 41 samples used in this study. These sequences may be divided into 2 categories: nominal allele variant sequences and novel allele sequences. For the purposes of this study, a nominal allele variant sequence is defined as any allele sequence that differs from the

previously documented sequence(s) for that particular allele. Novel allele sequences are sequences detected for alleles that have no previously published sequence data with which to compare.

Of the 46 previously undocumented allele sequences that were detected, 28 were classified as nominal allele variant sequences. These nominal variants were found in loci DYS389I/II, DYS390, DYS393, DYS449, DYS456, DYS481, DYS518, and DYS635, and have been further characterized as either SNP variants or repeat pattern variants (Table 1). Allele sequence variation may be introduced via strand slippage or one or more point variations within the repeat region. In this study, nominal variant sequences were classified as SNP variants if they displayed a repeat motif that differed from the commonly described motif, an occurrence indicative of point substitution. Repeat pattern variants are defined as allele sequences that differ from published data with regard to repeat unit arrangement, but are consistent with the reported repeat motif. Such variations may be due to strand slippage or the presence of intra-repeat SNPs, but definitive conclusions cannot be made without additional data. The unique sequences detected for allele “30” at locus DYS389II illustrate the difference between these two types of variants. The reported repeat motif of the DYS389II locus is $[TCTG]_n[TCTA]_pN_{48}[TCTG]_q[TCTA]_r$ (where n , p , q and r represent the number of repeats). One of the nominal allele variant sequences was $[TCTG]_5[TATA]_1[TCTA]_{11}N_{48}[TCTG]_3[TCTA]_{10}$, likely due to the presence of a C/A SNP in the first “TCTA” repeat unit. Since the change results in a “TATA” repeat unit that is inconsistent with the reported repeat motif, this sequence is classified as a SNP variant. The other nominal variant detected for this allele, $[TCTG]_6[TCTA]_{11}N_{48}[TCTG]_3[TCTA]_{10}$, remains consistent with the reported repeat motif, but displays a pattern of repeat units that has not been previously documented. Therefore, it is labeled a repeat pattern variant.

Locus	Reference Repeat Motif	Allele	Observed Repeat Motif	Counts	Population	Variant Type	Associated Haplogroups
DYS389I	[TCTG] ₃ [TCTA] _n	9	[TCTA] ₉	1	Caucasian	Repeat Pattern Variant	R1b
DYS389II	[TCTG] _n [TCTA] _p N ₄₈ [TCTG] ₃ [TCTA] _q	29	[TCTG] ₆ [TCTA] ₁₀ N ₄₈ [TCTG] ₃ [TCTA] ₁₀	1	Hispanic	Repeat Pattern Variant	E1b1b
		29	[TCTG] ₆ [TCTA] ₁₁ N ₄₈ [TCTG] ₃ [TCTA] ₉	1	African American	Repeat Pattern Variant	E1b1a
		30	[TCTG] ₅ [TCTA] ₁₁ N ₄₈ [TCTG] ₃ [TCTA] ₁₀	1	Hispanic	C/A SNP	J1
		30	[TCTG] ₆ [TCTA] ₁₁ N ₄₈ [TCTG] ₃ [TCTA] ₁₀	7	African American	Repeat Pattern Variant	E1b1a
		31	[TCTG] ₆ [TCTA] ₁₁ N ₄₈ [TCTG] ₃ [TCTA] ₁₁	1	Caucasian	Repeat Pattern Variant	E1b1a
		32	[TCTG] ₆ [TCTA] ₁₃ N ₄₈ [TCTG] ₃ [TCTA] ₁₀	1	African American	Repeat Pattern Variant	E1b1b
DYS390	[TCTG] ₈ [TCTA] _n [TCTG] ₁ [TCTA] ₄	21	[TCTG] ₈ [TCTA] ₈ [TCTG] ₁ [TCTA] ₄	9	African American	Repeat Pattern Variant	E1b1a
		21	[TCTG] ₈ [TCTA] ₉ [TCTG] ₁ [TCTA] ₃	1	African American	Repeat Pattern Variant	E1b1b
DYS393	[AGAT] _n	13	[cGAT] ₁ [AGAT] ₁₂	1	Caucasian	A/C SNP	R1a
DYS449	[TTTC] _n N ₅₀ [TTTC] _p	29	[TTTC] ₁₄ N ₅₀ [TTTC] ₁₅	4	2 Caucasian, 2 Hispanic	Repeat Pattern Variant	I1, I2b, O/Q, R1b
		30	[TTTC] ₁₅ N ₅₀ [TTTC] ₁₅	2	1 Hispanic, 1 African American	Repeat Pattern Variant	E1b1a, R1b
		31	[TTTC] ₁₆ N ₅₀ [TTTc] ₁ [TTTC] ₁₄	1	African American	C/A SNP	E1b1a
			[TTTC] ₁₆ N ₅₀ [TTTC] ₁₅	1	Hispanic	Repeat Pattern Variant	R1b
DYS456	[AGAT] _n	15	[AGAT] ₁ [AGAc] ₁ [AGAT] ₁₃	1	African American	T/C SNP	E1b1a
DYS481	[CTT] _n	25	[CTG] ₁ [CTT] ₂₄	1	Caucasian	T/G SNP	I2a
		26	[CTG] ₁ [CTT] ₂₅	1	Caucasian	T/G SNP	E1b1a
DYS518	[AAAG] ₃ [GAAG] ₁ [AAAG] _n [GGAG] ₁ [AAAG] ₄ N ₆ [AAAG] _p	36	[AAAG] ₃ [GAAG] ₁ [AAAG] ₁₄ [GGAG] ₁ [AAAG] ₄ N ₆ [AAAG] ₁₃	1	Hispanic	Repeat Pattern Variant	G2a
		37	[AAAG] ₃ [GAAG] ₁ [AAAG] ₁₆ [GGAG] ₁ [AAAG] ₄ N ₆ [AAAG] ₁₂	1	Caucasian	Repeat Pattern Variant	R1b
		38	[AAAG] ₃ [GAAG] ₁ [AAAG] ₁₄ [GGAG] ₁ [AAAG] ₄ N ₆ [AAAG] ₁₅	1	Hispanic	Repeat Pattern Variant	J2a
			[AAAG] ₃ [GAAG] ₁ [AAAG] ₁₅ [GGAG] ₁ [AAAG] ₄ N ₆ [AAAG] ₁₄	6	2 Caucasian, 1 Hispanic, 3 African American	Repeat Pattern Variant	E1b1a, I2a, J2b, R1b
		39	[AAAG] ₃ [GAAG] ₁ [AAAG] ₁₈ [GGAG] ₁ [AAAG] ₄ N ₆ [AAAG] ₁₂	1	Hispanic	Repeat Pattern Variant	I2b
		40	[AAAG] ₃ [GAAG] ₁ [AAAG] ₁₈ [GGAG] ₁ [AAAG] ₄ N ₆ [AAAG] ₁₃	1	African American	Repeat Pattern Variant	E1b1a
		41	[AAAG] ₃ [GAAG] ₁ [AAAG] ₁₆ [GGAG] ₁ [AAAG] ₄ N ₆ [AAAG] ₁₆	1	Caucasian	Repeat Pattern Variant	R1a
		42	[AAAG] ₃ [GAAG] ₁ [AAAG] ₁₆ [GGAG] ₁ [AAAG] ₄ N ₆ [AGAG] ₁ [AAAG] ₁₆	1	African American	A/G SNP	E1b1a
			[AAAG] ₃ [GAAG] ₁ [AAAG] ₁₉ [GGAG] ₁ [AAAG] ₄ N ₆ [AAAG] ₁₄	1	African American	Repeat Pattern Variant	E1b1a
DYS635	[TCTA] ₄ [TGTA] ₂ [TCTA] ₂ [TGTA] ₂ [TCTA] ₂ [TGTA] _n [TCTA] _p	21	[TCTA] ₄ [TGTA] ₂ [TCTA] ₂ [TGTA] ₂ [TCTA] ₅ [TCTc] ₁ [TCTA] ₅	1	African American	A/C SNP	E1b1a
		23	[TCTA] ₄ [TGTA] ₂ [TCTA] ₂ [TGTA] ₂ [TCTA] ₂ [TGTA] ₃ [TCTA] ₈	1	Hispanic	Repeat Pattern Variant	R1b

Table 1. Nominal allele variant sequences. These detected variant sequences differ from the published sequence data for these alleles. n , p , q , and r represent number of repeats per short tandem repeat unit.

The unique sequence detected for allele “9” at locus DYS389I is particularly interesting, as it completely lacks the “TCTG” repeat unit found in the locus’s repeat motif, [TCTG]_q[TCTA]_r. Instead, the variant allele, observed in only 1 Caucasian sample, consists entirely of “TCTA” repeats. The published sequence for this allele consists of 3 “TCTG” and 6 “TCTA” repeat units. Since the “TCTG” repeat unit, as defined in the reported repeat motif, is variable, its absence was not considered an inconsistency with regard to the motif, and this novel sequence is therefore deemed a repeat pattern variant. Including the aforementioned SNP variant at locus DYS389II, only 8 of the 28 Y-STR nominal variants were SNP variants. At locus DYS393, an A/C SNP in the variable “AGAT” repeat unit produced a leading “CGAT” unit in allele “13.” At locus DYS449, a C/A SNP in a “31” allele changed the repeat sequence from [TTTC]₁₆N₅₀[TTTC]₁₅, which was detected in another sample, to [TTTC]₁₆N₅₀[**TTTA**]₁[TTTC]₁₄. Similarly, a T/C SNP in allele “15” at locus DYS456 resulted in an “AGAC” repeat unit amongst the “AGAT” repeats. Additionally, a T/G SNP in the variable “CTT” repeat unit of alleles “25” and “26” at locus DYS481 resulted in the presence of a leading “CTG” repeat in both of these alleles. This SNP variation was previously characterized by Geppert *et al.* [6] in allele “21,” which also was detected in this study. An A/G SNP in allele “42” at locus DYS518 changed the allele sequence from [AAAG]₃[GAAG]₁[AAAG]₁₆[GGAG]₁[AAAG]₄N₆[AAAG]₁₇ to [AAAG]₃[GAAG]₁[AAAG]₁₆[GGAG]₁[AAAG]₄N₆[**AGAG**]₁[AAAG]₁₆. Lastly, the “21” allele at locus DYS635 (GATA-C4) exhibited the sequence [TCTA]₄[TGTA]₂[TCTA]₂[TGTA]₂[TCTA]₅[TCTC]₁[TCTA]₅ as a result of an A/C SNP that occurred in the trailing “TCTA” repeat units.

In addition to the effects of SNPs, the nominal allele sequences detected in this study highlight a high degree of allele variability at certain loci due to repeat pattern variation. Locus

DYS518, for instance, displayed multiple variants for all but 1 allele, some of which were previously characterized by D'Amato *et al.* [5]. These variations are due to differences in the quantities of the 2 variable “AAAG” repeat units at this locus. Finally, one of the detected sequence variations for the “23” allele at locus DYS635 (GATA-C4) is particularly interesting. This locus exhibits a wide range of allele variation due to the presence or absence of 2 “TGTA” repeats amongst the trailing “TCTA” repeat units, an occurrence that has been described previously in STRBase [14, 15] and Oloffson *et al.* [17]. However, the “23” allele detected in this study contained 3 “TGTA” repeats, resulting in a sequence variant that has not been characterized until now.

The majority of these nominal allele sequence variants displayed a low frequency of occurrence across the dataset, with 23 of the 28 allele variants detected in only a single sample each. However, the previously described repeat pattern variant observed for allele “30” at locus DYS389II, as well as one of the repeat pattern variants observed for allele “21” at locus DYS390, were detected in 7 and 9 samples, respectively. Interestingly, these variants occurred exclusively in African American samples, indicating that these alternative allele sequences may be population-specific and also may reflect the known greater genetic diversity in the African population. For the most part, other frequently observed sequence variants appeared to be fairly evenly parsed among at least 2 populations.

The majority of the allele sequences detected at the 28 targeted loci were consistent with previously published sequence information (data not shown). Noteworthy examples include the microvariant alleles “13.2” and “17.2” at loci DYS385 and DYS458, respectively, both of which

have been previously characterized by Myers *et al.* [18, 19]. At these loci, the microvariant alleles occur as a result of a “GA” deletion in the variable “GAAA” repeat unit. In addition to the large number of observed sequences that have been documented previously, a total of 18 novel allele sequences were detected across the 41 samples that were analyzed (Table 2). The number of samples in which these novel sequences were observed ranged from 1 to 13, although many occurred relatively infrequently across the dataset. The novel allele sequences included 2 SNP variants. At locus DYS570, a T/C SNP in allele “23” resulted in a sequence change from [TTTC]₂₃ to [TTTC]₅[TCTC]₁[TTTC]₁₇. Another T/C SNP, observed in allele “35” at locus DYS612, changed the repeat sequence from [CCT]₅[CTT]₁[TCT]₄[CCT]₁[TCT]₃₀ to [CCT]₅[CTT]₁[TCT]₄[CCT]₁[TCT]₁₇[CCT]₁[TCT]₁₂. The remaining novel sequences, such as those detected at locus DYS635 that illustrate the earlier mentioned variability at this locus, were consistent with the described repeat motifs of their respective alleles. Lastly, haplogroup assignments were made for each Y STR profile based on the number of repeats of each locus of a haplotype (Table 2). In addition, the haplogroups are listed for each of the reported allele sequences as these may prove useful for future population studies.

Locus	Reference Repeat Motif	Allele	Observed Repeat Motif	Counts	Population	Variant Type	Associated Haplogroups
DYS449	$[\text{TTC}]_n\text{N}_{50}[\text{TTC}]_p$	25	$[\text{TTC}]_{11}\text{N}_{50}[\text{TTC}]_{14}$	1	Hispanic	Repeat Pattern Variant	J1
DYS505	$[\text{TCCT}]_n$	11	$[\text{TCCT}]_{11}$	9	2 Caucasian, 5 Hispanic, 2 African American	Repeat Pattern Variant	E1b1b, G2a, I1, O/Q, R1b
		14	$[\text{TCCT}]_{14}$	2	African American	Repeat Pattern Variant	E1b1a
DYS533	$[\text{ATCT}]_n$	9	$[\text{ATCT}]_9$	1	Hispanic	Repeat Pattern Variant	G2a
		11	$[\text{ATCT}]_{11}$	13	5 Caucasian, 4 Hispanic, 4 African American	Repeat Pattern Variant	E1b1a, E1b1b, I1, J2a, O/Q, R1b
		13	$[\text{ATCT}]_{13}$	4	1 Caucasian, 2 Hispanic, 1 African American	Repeat Pattern Variant	R1b
		14	$[\text{ATCT}]_{14}$	1	Caucasian	Repeat Pattern Variant	R1b
DYS549	$[\text{GATA}]_n$	10	$[\text{GATA}]_{10}$	2	1 Caucasian, 1 African American	Repeat Pattern Variant	E1b1a, I2a
		11	$[\text{GATA}]_{11}$	6	1 Hispanic, 5 African American	Repeat Pattern Variant	E1b1a, E1b1b
DYS570	$[\text{TTC}]_n$	23	$[\text{TTC}]_5[\text{TcT}]_{11}[\text{TTC}]_{17}$	1	Hispanic	T/C SNP	E1b1b
DYS576	$[\text{AAAG}]_n$	13	$[\text{AAAG}]_{13}$	1	African American	Repeat Pattern Variant	E1b1a
		22	$[\text{AAAG}]_{22}$	1	Hispanic	Repeat Pattern Variant	R1b
DYS612	$[\text{CCT}]_3[\text{CTT}]_1[\text{TCT}]_4[\text{CCT}]_1[\text{TCT}]_n$	35	$[\text{CCT}]_3[\text{CTT}]_1[\text{TCT}]_4[\text{CCT}]_1[\text{TCT}]_{11}[\text{cCT}]_1[\text{TCT}]_{12}$	1	Hispanic	T/C SNP	J2b
DYS635	$[\text{TCTA}]_4[\text{TGTA}]_2[\text{TCTA}]_2[\text{TGTA}]_2[\text{TCTA}]_2[\text{TGTA}]_n[\text{TCTA}]_p$	24	$[\text{TCTA}]_4[\text{TGTA}]_2[\text{TCTA}]_2[\text{TGTA}]_2[\text{TCTA}]_2[\text{TGTA}]_{10}$	1	African American	Repeat Pattern Variant	R1b
		25	$[\text{TCTA}]_4[\text{TGTA}]_2[\text{TCTA}]_2[\text{TGTA}]_2[\text{TCTA}]_2[\text{TGTA}]_{11}$	2	Caucasian	Repeat Pattern Variant	R1b
		26	$[\text{TCTA}]_4[\text{TGTA}]_2[\text{TCTA}]_2[\text{TGTA}]_2[\text{TCTA}]_2[\text{TGTA}]_{12}$	1	African American	Repeat Pattern Variant	R1b
DYS643	$[\text{CTTTT}]_n$	8	$[\text{CTTTT}]_8$	1	Hispanic	Repeat Pattern Variant	J2a
		14	$[\text{CTTTT}]_{14}$	1	African American	Repeat Pattern Variant	E1b1a

Table 2. Novel allele variant sequences. These detected variant sequences represent alleles for which there are currently no published sequence data with which

to compare. n , p , q , and r represent number of repeats per short tandem repeat unit.

CONCLUSIONS

The unique allele sequence variants detected in this study have been presented to demonstrate that additional characterization of Y-STR alleles is feasible by sequencing. The results also provide some insight into the mechanism of allele variant occurrence. While SNP variants were detected, the majority of novel sequences consisted of repeat pattern variants. Although the exact mechanism of mutation for the repeat pattern variants observed in this study cannot be definitively concluded, it should be noted that the majority of STR variation has been attributed to strand slippage [20-22]. Therefore, even if a single point mutation event may seem to be the most parsimonious explanation for a repeat pattern variant, a two-step strand slippage event may be more probable. Such concepts must be taken into account when characterizing these novel variants. Regardless of their mechanism of introduction, the presence of intra-repeat SNPs and repeat pattern variations in Y-STR alleles may aid in the differentiation of males sharing the same nominal alleles, and perhaps even paternally-related males, in forensic casework samples. Given its ability to detect both the length of STR alleles and their individual nucleotide sequences, MPS technology offers more resolution with regard to STRs than traditional length-based detection methods, such as CE. CE would yield the size of an amplicon, i.e., equivalent of repeat length, which can be ascertained from sequence data simply by counting the number of nucleotides within the repeat region. To date, the vast majority of STR nominal length results have been the same between MPS and CE derived data (data not shown). While the dataset used in this study was relatively small, the large number of observed novel allele sequence variants highlights the need for characterization of Y-STR alleles in larger sample populations.

CONFLICT OF INTEREST

None.

ACKNOWLEDGMENTS

This work was supported in part by award no 2012-DN-BXK033, awarded by the National Institute of Justice, Office of Justice Programs, U.S. Department of Justice. The opinions, findings, and conclusions or recommendations expressed in this publication are those of the authors and do not necessarily reflect those of the U.S. Department of Justice. The authors also would like to thank Illumina, Inc. for its support during this study.

REFERENCES

- [1] Flores S, Sun J, King J, Budowle B. Internal validation of the GlobalFiler™ Express PCR Amplification Kit for the direct amplification of reference DNA samples on a high-throughput automated workflow. *Forensic Sci Int Genet* 2014; 10:33-9.
- [2] Oostdik K, Lenz K, Nye J, Schelling K, Yet D, Bruski S, Strong J, Buchanan C, Sutton J, Linner J, Frazier N, Young H, Matthies L, Sage A, Hahn J, Wells R, Williams N, Price M, Koehler J, Staples M, Swango KL, Hill C, Oyerly K, Duke W, Katzilierakis L, Ensenberger MG, Bourdeau JM, Sprecher CJ, Krenke B. Developmental validation of the PowerPlex® Fusion System for analysis of casework and reference samples: a 24-locus multiplex for new database standards. *Forensic Sci Int Genet* 2014; 12:69-76.
- [3] Mulero JJ, Chang CW, Calandro LM, Green RL, Li Y, Johnson CL, Hennessy LK. Development and validation of the AmpFØSTR® Yfiler™ PCR Amplification Kit: a male specific, single amplification 17 Y-STR multiplex system. *J Forensic Sci* 2006; 51:64-75.
- [4] Davis C, Ge J, Sprecher C, Chidambaram A, Thompson J, Ewing M, Fulmer P, Rabbach D, Storts D, Budowle B. Prototype PowerPlex® Y23 System: a concordance study. *Forensic Sci Int Genet* 2013;7:204-8.
- [5] D'Amato ME, Ehrenreich L, Cloete K, Benjeddou M, Davison S. Characterization of the highly discriminatory loci DYS449, DYS481, DYS518, DYS612, DYS626, DYS644 and DYS710. *Forensic Sci Int Genet* 2010;4:104-10.
- [6] Geppert M, Edelmann J, Lessig R. The Y-chromosomal STRs DYS481, DYS570, DYS576, and DYS643. *Leg Med* 2009;11:S109-S110.
- [7] Planz JV, Sannes-Lowery KA, Duncan DD, Manalili S, Budowle B, Chakraborty R, Hofstadler SA, Hall TA. Automated analysis of sequence polymorphism in STR alleles by PCR and direct electrospray ionization mass spectrometry. *Forensic Sci Int Genet* 2012;6:594-606.
- [8] Pitterl F, Schmidt K, Huber G, Zimmermann B, Delport R, Amory S, Ludes B, Oberacher H, Parson W. Increasing the discrimination power of forensic STR testing by employing high-performance mass spectrometry, as illustrated in indigenous South African and Central Asian populations. *Int J Legal Med* 2010;124:551-8.
- [9] Zeng X, King JL, Stoljarova M, Warshauer DH, LaRue BL, Sajantila A, Patel J, Storts DR, Budowle B. High sensitivity multiplex short tandem repeat loci analyses with massively parallel sequencing. *Forensic Sci Int Genet* 2015;16:38-47.
- [10] Churchill JD, Chang J, Ge J, Rajagopalan N, Lagacé R, Liao W, King JL, Budowle B. Blind study evaluation of the Ion PGM™ System for use in human identity DNA typing. *Croat Med J* (submitted).

- [11] Adey A, Morrison HG, Asan, Xun X, Kitzman JO, Turner EH, Stackhouse B, MacKenzie AP, Caruccio NC, Zhang X, Shendure J. Rapid, low-input, low-bias construction of shotgun fragment libraries by high-density *in vitro* transposition. *Genome Biol* 2010;11:R119.
- [12] Warshauer DH, King JL, Budowle B. STRait Razor v2.0: the improved STR Allele Identification Tool – Razor. *Forensic Sci Int Genet* 2015;14:182-6.
- [13] STRait Razor Unique Sequences Compiler (USC): http://web.unthsc.edu/info/200210/molecular_and_medical_genetics/887/research_and_development_laboratory/5.
- [14] STRBase Y-STR fact sheets: http://www.cstl.nist.gov/strbase/ystr_fact.htm.
- [15] SRM 2395 Human Y-Chromosome DNA Profiling Standard: <http://www.cstl.nist.gov/strbase/srm2395.htm>.
- [16] Butler JM, Decker AE, Vallone PM, Kline MC. Allele frequencies for 27 Y-STR loci with U.S. Caucasian, African American, and Hispanic samples. *Forensic Sci. Int* 2006;156:250-60.
- [17] Olofsson J, Andersen MM, Mogensen HS, Eriksen PS, Morling N. Sequence variants of allele 22 and 23 of DYS635 causing different stutter rates. *Forensic Sci Int Genet* 2012;6:e161-e162.
- [18] Myers NM, Ritchie KH, Lin AA, Hughes RH, Woodward SR, Underhill PA. Y-chromosome short tandem repeat intermediate variant alleles DYS392.2, DYS449.2, and DYS385.2 delineate new phylogenetic substructure in human Y-chromosome haplogroup tree. *Croat Med J* 2009;50:239-49.
- [19] Myers NM, Ekins JE, Lin AA, Cavalli-Sforza LL, Woodward SR, Underhill PA. Y-chromosome short tandem repeat DYS458.2 non-consensus alleles occur independently in both binary haplogroups J1-M267 and R1b3-M405. *Croat Med J* 2007;48:450-9.
- [20] Pumpernik D, Oblak B, Borstnik B. Replication slippage versus point mutation rates in short tandem repeats of the human genome. *Mol Genet Genomics* 2008;279:53-61.
- [21] Ballantyne KN, 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, Ploski R, Dobosz T, Henke L, Henke J, Furtado MR, Kayser M. Mutability of Y-chromosomal microsatellites: rates, characteristics, molecular bases, and forensic implications. *Am J Hum Genet* 2010;87:341-53.
- [22] Ge J, Budowle B, Aranda XG, Planz JV, Eisenberg AJ, Chakraborty R. Mutation rates at Y chromosome short tandem repeats in Texas populations. *Forensic Sci. Int Genet* 2009;3:179-84.

SUMMARY

DEVELOPMENT OF A COMPREHENSIVE MASSIVELY PARALLEL SEQUENCING PANEL OF SINGLE NUCLEOTIDE POLYMORPHISM AND SHORT TANDEM REPEAT MARKERS FOR HUMAN IDENTIFICATION

Massively parallel sequencing (MPS) allows for the detection of an unparalleled amount of genetic information with unprecedented speed and relative ease. This doctoral dissertation research was conducted under the hypothesis that MPS, with its economies of scale, can provide a system whereby reference samples can be typed for a large battery of markers, providing more discriminatory power for forensic DNA typing and offering increased opportunities to develop investigative leads. The primary goal of this project was to develop the capability of typing reference samples for a large battery of markers: 84 autosomal, Y-chromosome, and X-chromosome short tandem repeats (STRs), Amelogenin, and 275 human identity single nucleotide polymorphisms (SNPs), in a single multiplex analysis.

Chapter 1 described the creation of bioinformatic software called STRait Razor (the STR Allele Identification Tool – Razor). At the time this research began, the existing MPS instruments were capable of providing extensive data, but the available software tools were limited for identifying forensic STR alleles. Without suitable software for STR analysis, the process was tedious and time-consuming, and comparison of results was difficult. Therefore, this dissertation project required the development of specialized STR-typing software for MPS data. STRait Razor is a Linux-based Perl script that identifies alleles at STR loci based on the length of the repeat sequence. The software is capable of handling repeat motifs ranging from simple to complex, and it does not require a reference composed of extensive allelic sequence data. As a result, the allele call results are consistent with those of current CE-based methods, and it is not confounded by unexpected sequence variation within repeats. In its first iteration, STRait Razor was programmed to identify alleles at 44 forensically-relevant STR loci. The ability of STRait Razor to accurately detect these 44 STR alleles in raw sequence data was tested through concordance testing with five

samples that had been typed by capillary electrophoresis (CE) and subsequently sequenced using two library preparation methods on two different sequencing platforms. During its initial testing, STRait Razor achieved 100% concordance with the CE results, proving that the software was capable of analyzing the types of markers that would be included in the comprehensive MPS panel.

Since the comprehensive panel would include a much larger number of STR loci than STRait Razor had originally been configured to analyze, the software had to be improved. **Chapter 2** details the modifications and upgrades made to STRait Razor that allowed it to efficiently type a total of 86 markers, including Amelogenin. Once again, the software was tested for concordance with known typing results, and 100% concordance was achieved. One of the most important changes to the program was the development of a more efficient means of obtaining the intra-repeat sequence data associated with the detected STR alleles. The study documented in **Chapter 2** illustrated how this valuable information could be used to detect intra-repeat sequence variation within STR alleles.

With reliable data analysis tools in place, a proof-of-concept study was conducted to evaluate the efficacy of using MPS technology to detect human identity markers of forensic interest. **Chapter 3** detailed the testing of the TruSeq™ Forensic Amplicon kit, a polymerase chain reaction (PCR) multiplex-based library preparation method used to detect 160 forensically-relevant SNP loci. The SNP genotypes yielded through the use of the multiplex were largely concordant with known whole genome sequencing data, and the heterozygote balance (based on coverage) was relatively even. The results of this study indicated that MPS is capable of typing a large number of forensic markers. However, it was clear that a multiplex-based library preparation

method was not ideal for typing an extensive battery of STR loci, as it would be difficult to optimize a PCR multiplex for such a large quantity of markers.

Chapter 4 described the design and testing of a Nextera™ Rapid Capture panel consisting of 84 STRs (31 autosomal, 26 X-chromosome, and 27 Y-chromosome), Amelogenin, and 275 human identity SNPs (240 autosomal and 35 Y-chromosome). The capture-based Nextera™ library preparation method was selected to reduce limitations inherent in PCR multiple-based designs. This comprehensive MPS panel was used to analyze a total of 190 individuals from the U.S. Caucasian, African American, Hispanic, and Asian populations. The efficacy of the panel was evaluated through concordance testing with results generated by the ForenSeq™ DNA Signature Prep Kit, as well as coverage-based performance statistics and heterozygote balance. Overall, the performance of the comprehensive panel was similar to that of commercial PCR-based kits. The panel exhibited a high degree of concordance, low dropout, and good heterozygote balance. Population genetic statistics were calculated, including tests for conformity with Hardy-Weinberg equilibrium and linkage equilibrium, as well as determination of F_{ST} values. Y-chromosome STR haplotypes were used to calculate haplotype diversity and determine a haplogroup prediction for each individual in the dataset. In summary, the results of these test were consistent with expectations for the markers in question. The study described in **Chapter 4** illustrated that MPS could be used to effectively type a comprehensive battery of forensically-relevant SNP and STR markers with relative ease.

One of the benefits of MPS, as opposed to CE, is that it can provide valuable information about the individual nucleotide sequence of STR alleles rather than just their length. Due to the

fact that STRait Razor is designed to retain these data while making STR allele calls, the information can be used to identify intra-repeat nucleotide variations. **Chapter 5** illustrated this added benefit. Y-chromosome STR allelic sequence data from 41 male DNA samples (a subset of the samples used in the testing of the comprehensive MPS panel) were analyzed in an attempt to detect previously undocumented nucleotide variations. A total of 46 unique intra-repeat sequence variants were detected. Of these, 28 were variations of documented sequences, while the remaining 18 were novel. This study indicated that MPS is capable of providing additional data that may increase the discrimination power of STR typing.

CONCLUSIONS

DEVELOPMENT OF A COMPREHENSIVE MASSIVELY PARALLEL SEQUENCING PANEL OF SINGLE NUCLEOTIDE POLYMORPHISM AND SHORT TANDEM REPEAT MARKERS FOR HUMAN IDENTIFICATION

A comprehensive capture-based massively parallel sequencing (MPS) panel was developed during the course of this dissertation research. To date, this panel is the most potentially informative assay for reference sample testing for human identification. The results of these studies indicate that MPS is capable of providing robust genetic data from a large battery of forensically-relevant short tandem repeat (STR) and single nucleotide polymorphism (SNP) loci in a single analysis. This capability makes MPS one of the best methods for the typing of forensic reference samples. A far greater number of markers can be typed simultaneously with this technology as opposed to traditional methods, such as capillary electrophoresis (CE).

Given the increase of typing data afforded with MPS, the majority (if not all) of the profiles from evidence samples can be compared with reference profiles, regardless of the number and types of markers used in the evidentiary analyses. The profiles generated by MPS technology are compatible with existing reference data, and the more comprehensive set of markers made possible through the use of MPS will foster investigations.

A capture-based assay was chosen as the methodology for the comprehensive MPS panel described in this dissertation, under the assumption that it would be less susceptible to the difficulties associated with PCR-based methods. While this eliminated the need for challenging multiplex optimization, the results of the study indicate that the performance and artifacts observed with a PCR enrichment method also persist with the capture-based approach. There is locus-to-locus coverage variation, stutter does occur (mostly due to an amplification stage prior to sequencing), heterozygote balance is similar, and low-level noise exists (which can be filtered out). The phenomena of locus-to-locus signal difference, stutter, and noise are commonly encountered

in DNA typing and can be managed in a similar fashion as they are with CE-based systems. However, sequence data obtained with MPS technology affords additional information that may assist in the characterization of these phenomena.

The design and implementation of large marker panels for the typing of reference samples will reduce debates on the best core markers for forensic utility, generate innovation because focus will not be solely on a core set of autosomal STRs, promote the development of better systems that can analyze more challenging samples, and enable sharing of data across laboratories worldwide.

With the increased number of forensic markers that were incorporated in this comprehensive panel, including lineage markers such as Y-chromosome STRs and Y-chromosome SNPs, indirect database searches can be performed, if desired. The success rate of familial searching would likely increase, as the sheer abundance of markers will provide more robust associations and substantially reduce the number of adventitious associations. Increased marker typing also will have a substantial impact on missing persons identification testing.

One of the benefits of a comprehensive MPS panel is that it can reveal the nucleotide sequences of STR alleles. A bioinformatic tool designed to detect these sequence data, such as STRait Razor, can be used to investigate nucleotide differences within the repeat regions of alleles. This information can enhance the discrimination power of forensic DNA typing, improve the effectiveness of genetic analyses, and provide useful information toward the understanding of human variation.

With the rapid advancements in MPS technologies and chemistries, it is anticipated that the amount of input DNA required for capture based approaches will continue to decrease and MPS will eventually become a method of choice for analysis of forensic evidence samples. Future studies may focus on the evaluation of a comprehensive MPS panel for the purpose of analyzing samples with low levels of input DNA or degraded samples, both of which are commonly encountered in forensic evidentiary casework.

Another avenue for future research is the characterization of intra-repeat nucleotide variation in STR alleles. Studies are currently underway to determine the autosomal and X-chromosome STR sequence variants detected in the samples typed during this dissertation research, to complement the analysis of Y-chromosome STR sequence variation that already has been performed, as a proof of concept.

The work described in this dissertation was performed in accordance with all laws (both Federal and State) that apply to research, researcher conduct, and the protection of human test subjects. It was also conducted under the guidance of and in accordance with the policies of the University of North Texas Health Science Center Institutional Review Board.

SUPPLEMENTAL MATERIAL

**DEVELOPMENT OF A COMPREHENSIVE MASSIVELY
PARALLEL SEQUENCING PANEL OF SINGLE
NUCLEOTIDE POLYMORPHISM AND SHORT
TANDEM REPEAT MARKERS FOR HUMAN
IDENTIFICATION**

The original STRait Razor program (**Chapter 1**) can be found at:
<http://dx.doi.org/10.1016/j.fsigen.2013.04.005>.

STRAit Razor v2.0 (**Chapter 2**) can be found at: <https://www.unthsc.edu/graduate-school-of-biomedical-sciences/molecular-and-medical-genetics/laboratory-faculty-and-staff/strait-razor/>.

SNP	Type	SNP	Type	SNP	Type	SNP	Type
rs1005533	iSNP	rs2107612	iSNP	rs826472	iSNP	rs1369093	apSNP
rs10092491	iSNP	rs2111980	iSNP	rs873196	iSNP	rs1393350	apSNP
rs1015250	iSNP	rs214955	iSNP	rs876724	iSNP	rs1407434	apSNP
rs1024116	iSNP	rs221956	iSNP	rs891700	iSNP	rs1408801	apSNP
rs1028528	iSNP	rs2269355	iSNP	rs901398	iSNP	rs1471939	apSNP
rs1029047	iSNP	rs2342747	iSNP	rs907100	iSNP	rs1498444	apSNP
rs1031825	iSNP	rs2399332	iSNP	rs914165	iSNP	rs1500127	apSNP
rs10488710	iSNP	rs251934	iSNP	rs917118	iSNP	rs1503767	apSNP
rs10495407	iSNP	rs279844	iSNP	rs938283	iSNP	rs1513056	apSNP
rs1058083	iSNP	rs2830795	iSNP	rs964681	iSNP	rs1513181	apSNP
rs10773760	iSNP	rs2831700	iSNP	rs987640	iSNP	rs1569175	apSNP
rs10776839	iSNP	rs2920816	iSNP	rs9905977	iSNP	rs1573020	apSNP
rs1109037	iSNP	rs338882	iSNP	rs993934	iSNP	rs1760921	apSNP
rs1294331	iSNP	rs354439	iSNP	rs9951171	iSNP	rs1800407	apSNP
rs12997453	iSNP	rs3780962	iSNP	rs10007810	apSNP	rs1805009	apSNP
rs13182883	iSNP	rs430046	iSNP	rs10108270	apSNP	rs182549	apSNP
rs13218440	iSNP	rs4364205	iSNP	rs10141763	apSNP	rs1837606	apSNP
rs1335873	iSNP	rs445251	iSNP	rs10236187	apSNP	rs1871428	apSNP
rs1355366	iSNP	rs4530059	iSNP	rs1040045	apSNP	rs1879488	apSNP
rs1357617	iSNP	rs560681	iSNP	rs1040404	apSNP	rs1950993	apSNP
rs1360288	iSNP	rs576261	iSNP	rs1042602	apSNP	rs1978806	apSNP
rs1382387	iSNP	rs6444724	iSNP	rs10496971	apSNP	rs2001907	apSNP
rs1413212	iSNP	rs6811238	iSNP	rs10511828	apSNP	rs2024566	apSNP
rs1454361	iSNP	rs6955448	iSNP	rs10512572	apSNP	rs2026721	apSNP
rs1463729	iSNP	rs7041158	iSNP	rs10513300	apSNP	rs2030763	apSNP
rs1490413	iSNP	rs717302	iSNP	rs10839880	apSNP	rs2033111	apSNP
rs1493232	iSNP	rs719366	iSNP	rs10843344	apSNP	rs2065160	apSNP
rs1498553	iSNP	rs722098	iSNP	rs10954737	apSNP	rs2065982	apSNP
rs1523537	iSNP	rs722290	iSNP	rs11227699	apSNP	rs2070586	apSNP
rs1528460	iSNP	rs727811	iSNP	rs12130799	apSNP	rs2125345	apSNP
rs159606	iSNP	rs729172	iSNP	rs12203592	apSNP	rs214678	apSNP
rs1736442	iSNP	rs733164	iSNP	rs12544346	apSNP	rs2269793	apSNP
rs1821380	iSNP	rs735155	iSNP	rs12629908	apSNP	rs2303798	apSNP
rs1886510	iSNP	rs737681	iSNP	rs12657828	apSNP	rs2306040	apSNP
rs1979255	iSNP	rs740598	iSNP	rs12821256	apSNP	rs2330442	apSNP
rs2016276	iSNP	rs740910	iSNP	rs12896399	apSNP	rs2357442	apSNP
rs2040411	iSNP	rs7520386	iSNP	rs1296819	apSNP	rs2378249	apSNP
rs2046361	iSNP	rs763869	iSNP	rs1321333	apSNP	rs239031	apSNP
rs2056277	iSNP	rs8037429	iSNP	rs1325502	apSNP	rs2402130	apSNP
rs2076848	iSNP	rs8078417	iSNP	rs13400937	apSNP	rs4959270	apSNP

Chapter 3, Supplemental Table 1. iSNPs and apSNPs included in the TruSeq™ Forensic Amplicon multiplexes.
The Forensic Amplicon multiplexes consisted of 94 iSNPs and 66 apSNPs.

	1	2	3	4	5	6	7	8	9	10	11	12
rs1058083							A/G : G (6.7% : 2.2%) [746 : 227]	A/G : G (5.6% : 1.0%) [699 : 1349]		A/G : G (5.8% : 0.5%) [587 : 572]		
rs2399332		G/T : G (9.0% : 6.1%) [3790 : 2552]					G/T : G (7.5% : 5.9%) [4171 : 2084]	G/T : G (9.9% : 5.7%) [3915 : 10453]		G/T : G (12.9% : 5%) [3103 : 5817]		

Chapter 3, Supplemental Table 2. SNP discordance (in-house vs. Illumina). Discordance between the SNP calls generated in this study and those obtained through Illumina's testing of the protocol are shown. Discordance is shown in the following format: "in-house call : Illumina call". Observed heterozygosity balance values are displayed in parentheses, in the format "in-house heterozygosity value : Illumina heterozygosity value". Overall read coverage values for each SNP locus are displayed in brackets, in the format "in-house coverage : Illumina coverage".

	1	2	3	4	5	6	7	8	9	10	11	12
rs10092491	48.8% (183)											
rs1029047	16.8% (195)	12.6% (281)		19.2% (274)	23.9% (270)	19.5% (239)	36.2% (312)	11.8% (312)	36.2% (331)	17.8% (251)	17% (379)	
rs1058083					4.9% (644)		6.7% (746)	5.6% (699)		5.8% (587)		
rs12997453	11.8% (76)				14.3% (80)	22.9% (59)						
rs1493232										46.3% (2445)		
rs182549				48.3% (3771)			23.2% (4014)					17.2% (3173)
rs2076848												28.9% (7233)
rs2399332	11.7% (3492)	8.9% (3790)					7.4% (4171)	9.9% (3915)		12.9% (3103)	9.9% (3535)	
rs338882							46.4% (1707)					
rs876724										48.2% (123)		

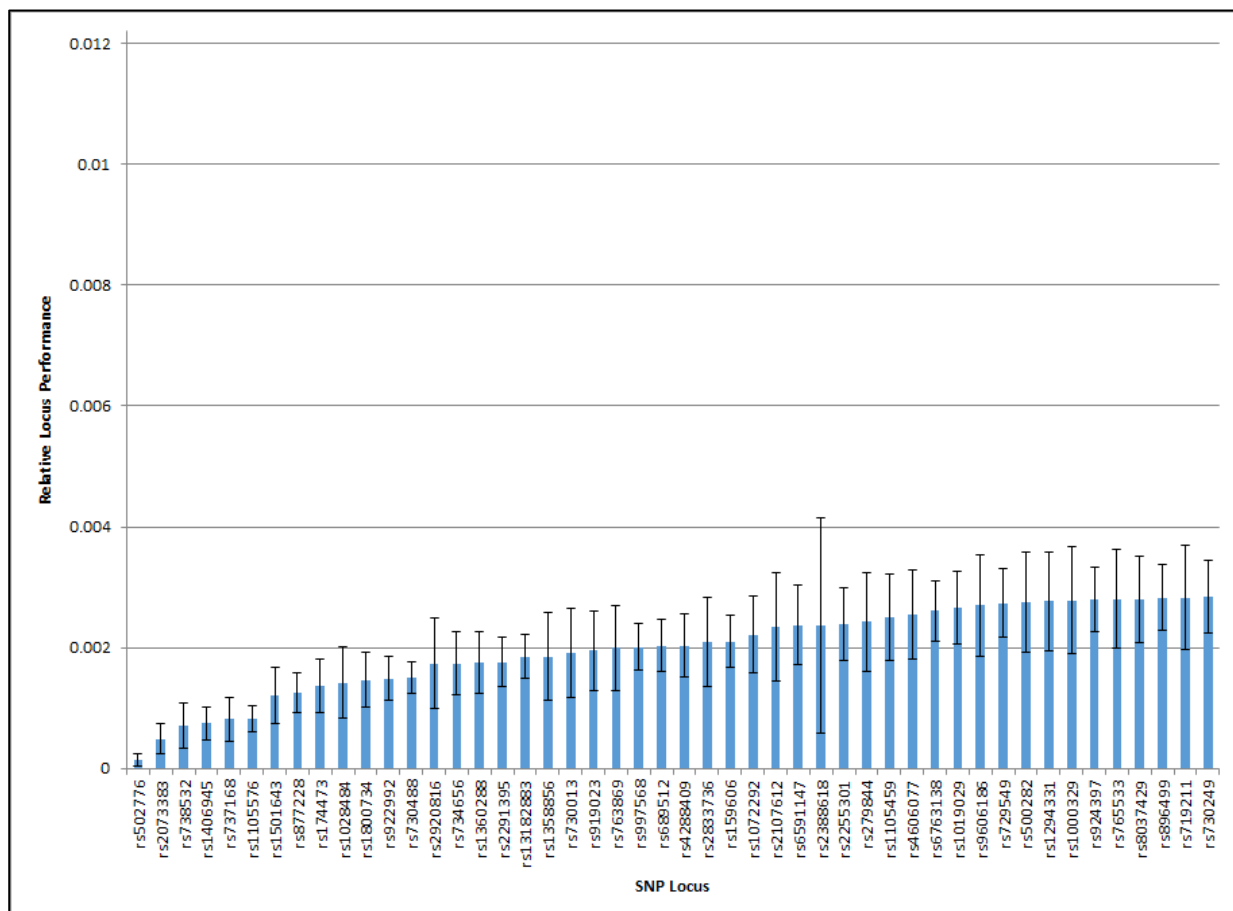
Chapter 3, Supplemental Table 3. Heterozygous SNP loci with allelic balance values below 50%. An allelic balance value of 50% is equivalent to an allelic balance ratio of 1:2. Numbers in parentheses are sequence coverage values.

Supplemental Workbooks 1 and 2 (**Chapter 4**) can be found at:

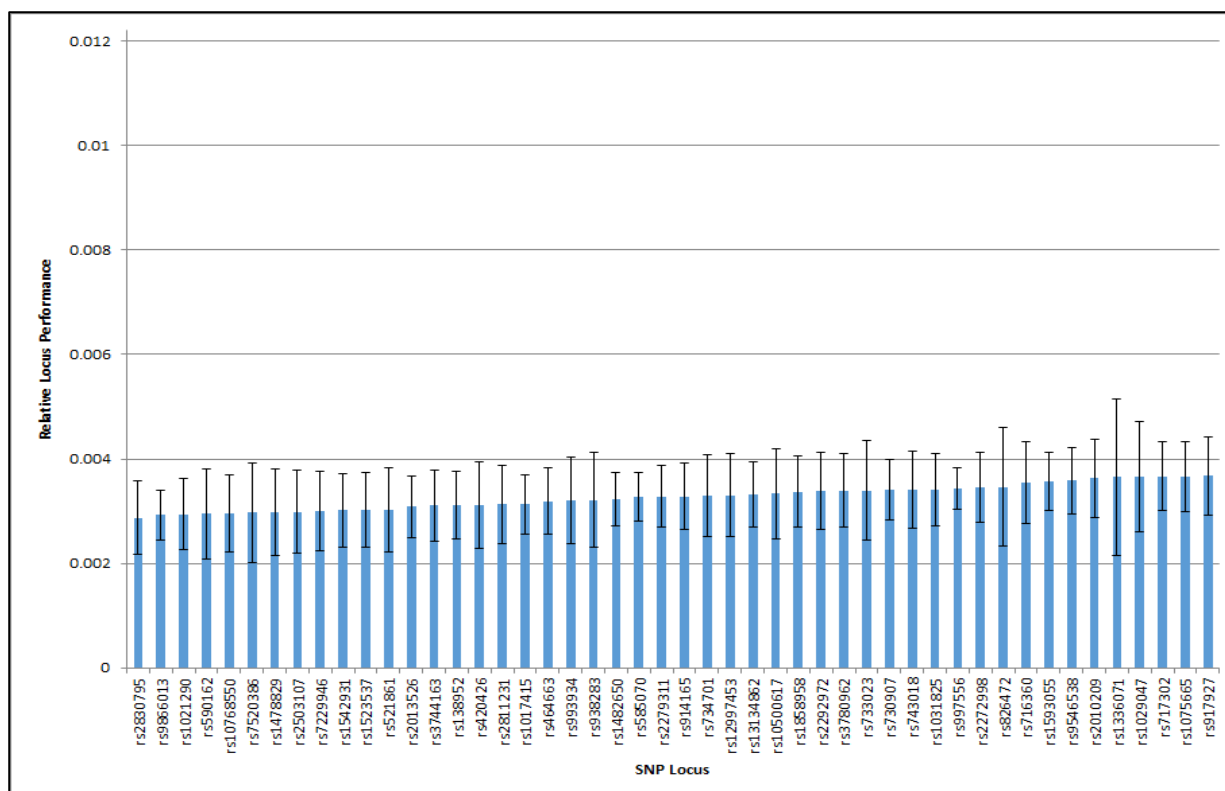
<https://app.box.com/s/eyb8bc8743n60bszmzzpc5o70jcbtgb7>

and

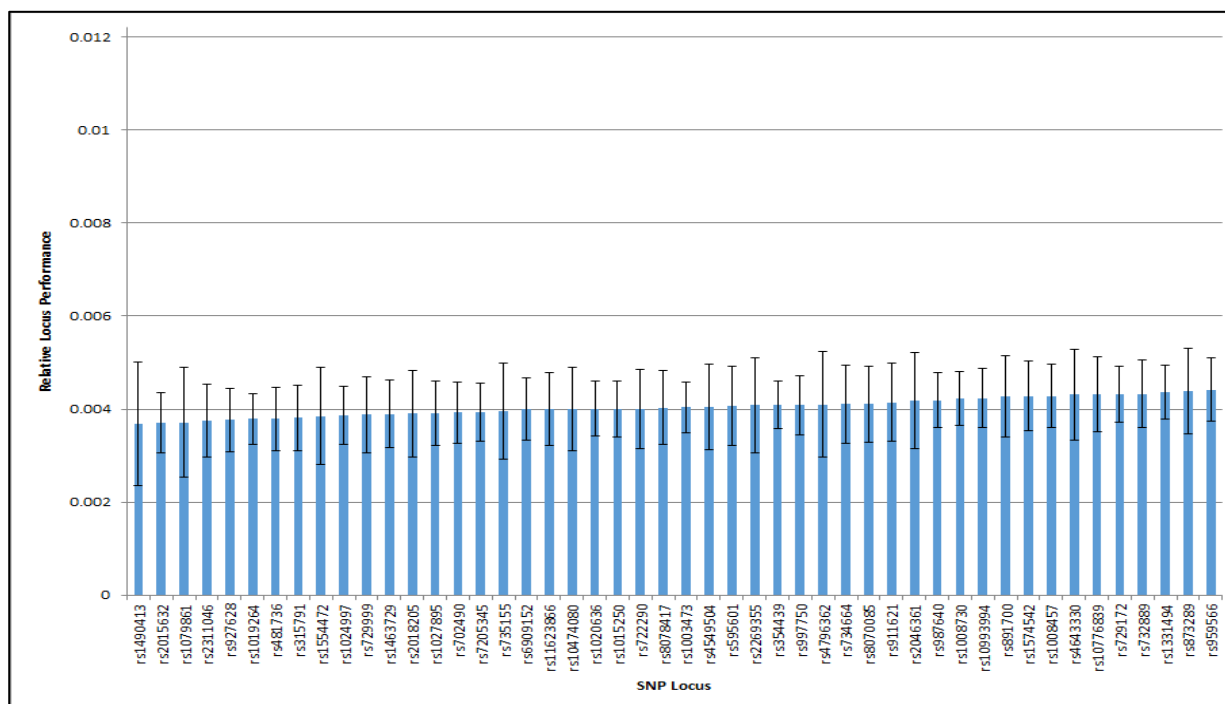
<https://app.box.com/s/pjm35mvaan6h10vk68jiu922j49yhuum>



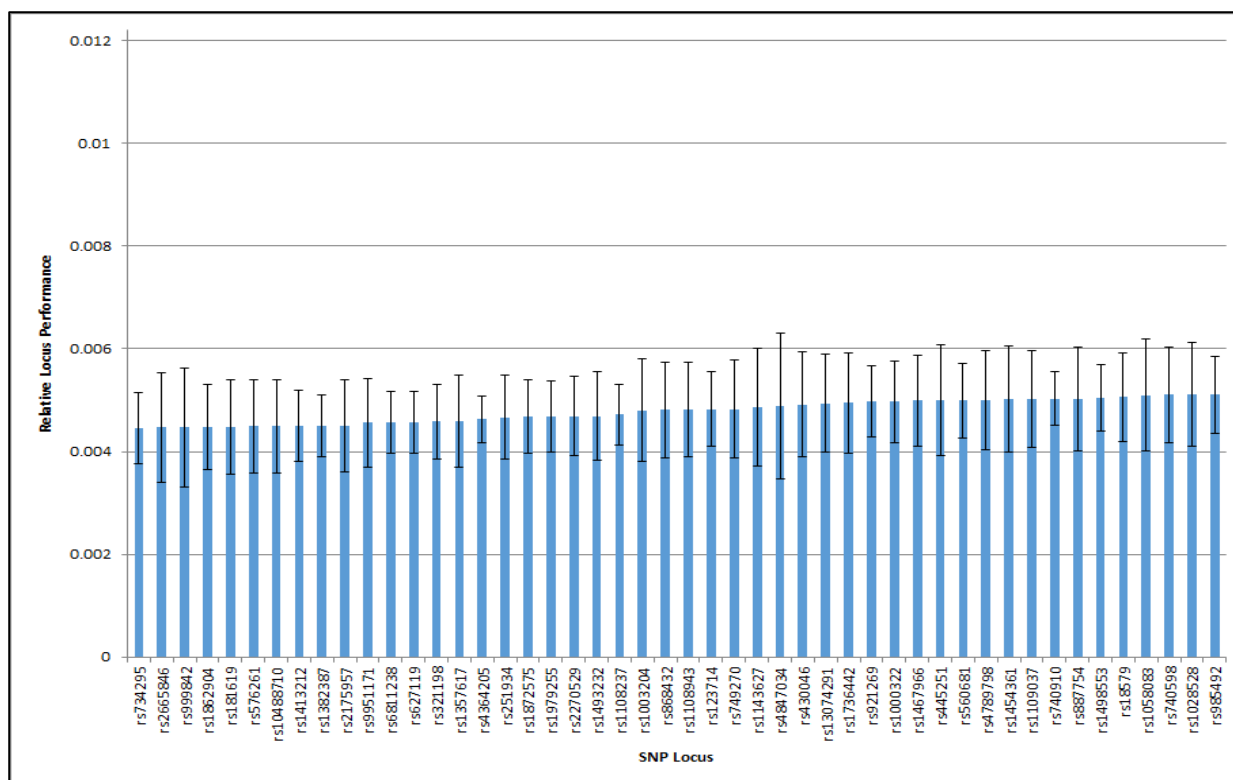
Chapter 4, Supplemental Figure 1. Relative locus performance – autosomal SNPs (part 1). Overall performance values, based on coverage, are shown for a subset of the 240 autosomal SNPs.



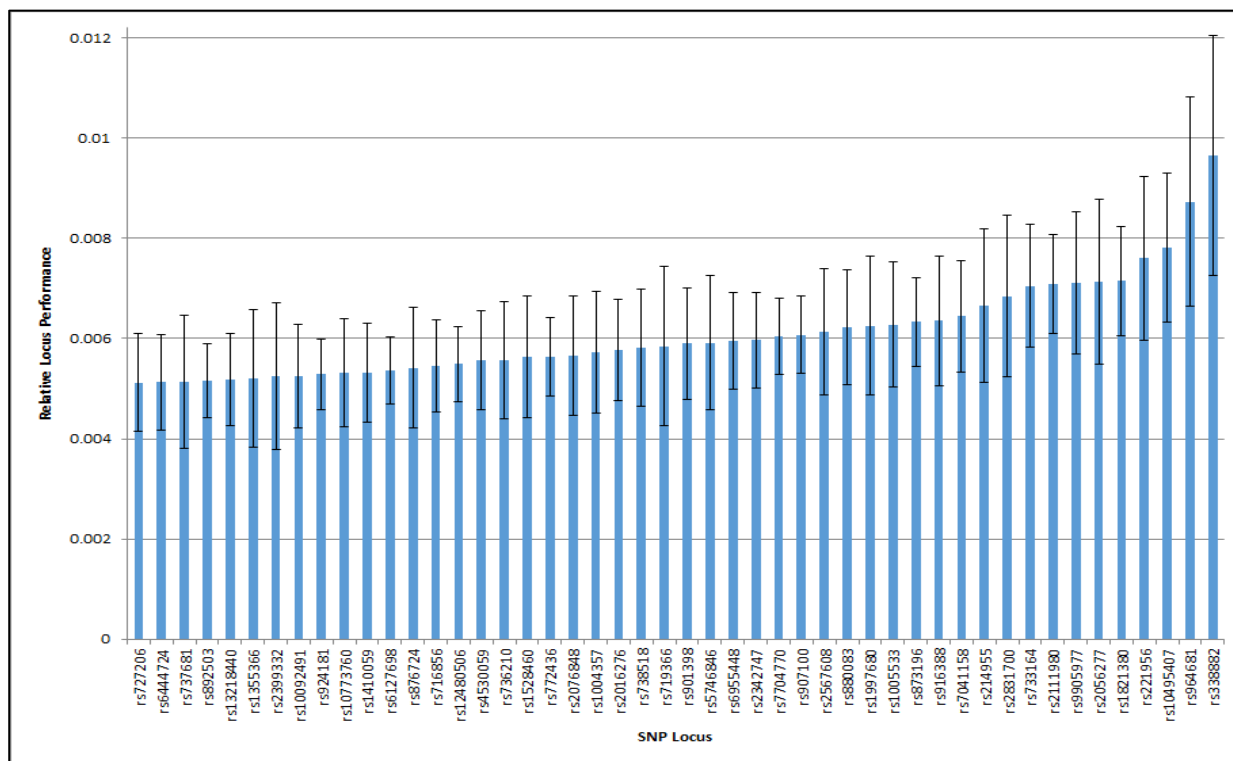
Chapter 4, Supplemental Figure 2. Relative locus performance – autosomal SNPs (part 2). Overall performance values, based on coverage, are shown for a subset of the 240 autosomal SNPs.



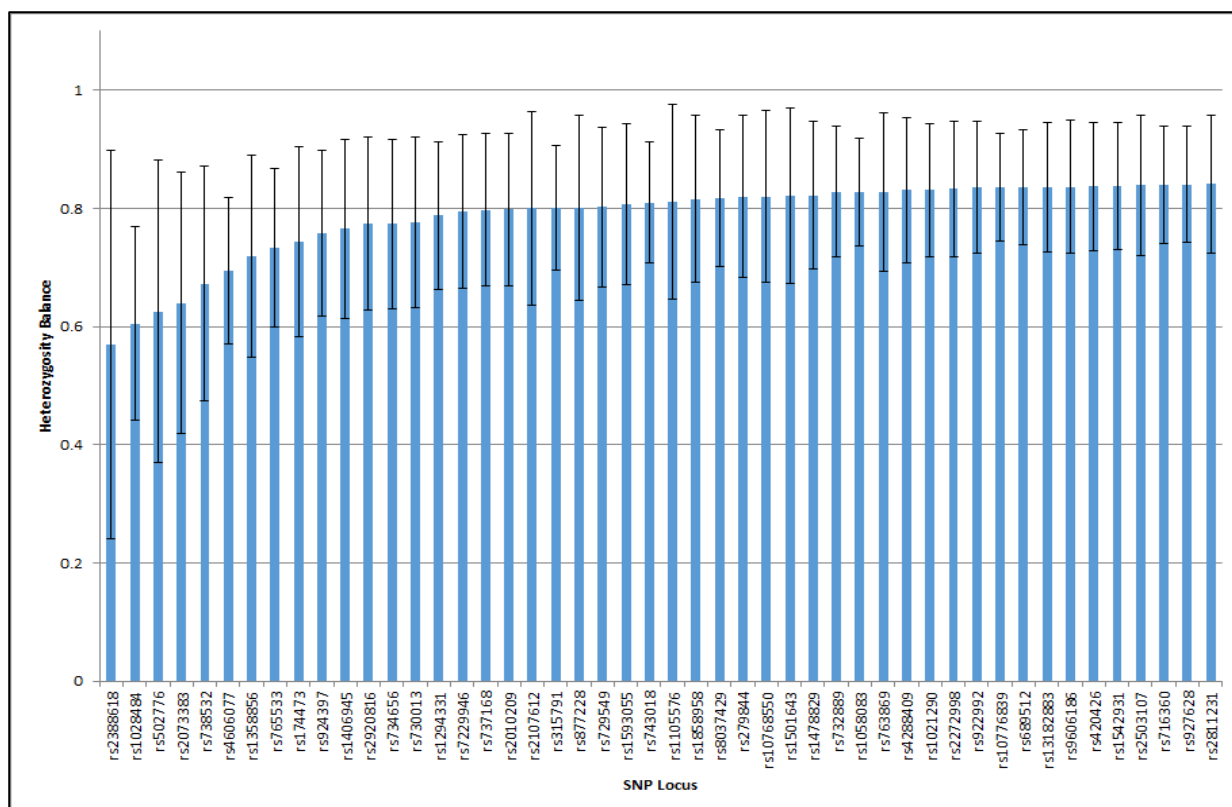
Chapter 4, Supplemental Figure 3. Relative locus performance – autosomal SNPs (part 3). Overall performance values, based on coverage, are shown for a subset of the 240 autosomal SNPs.



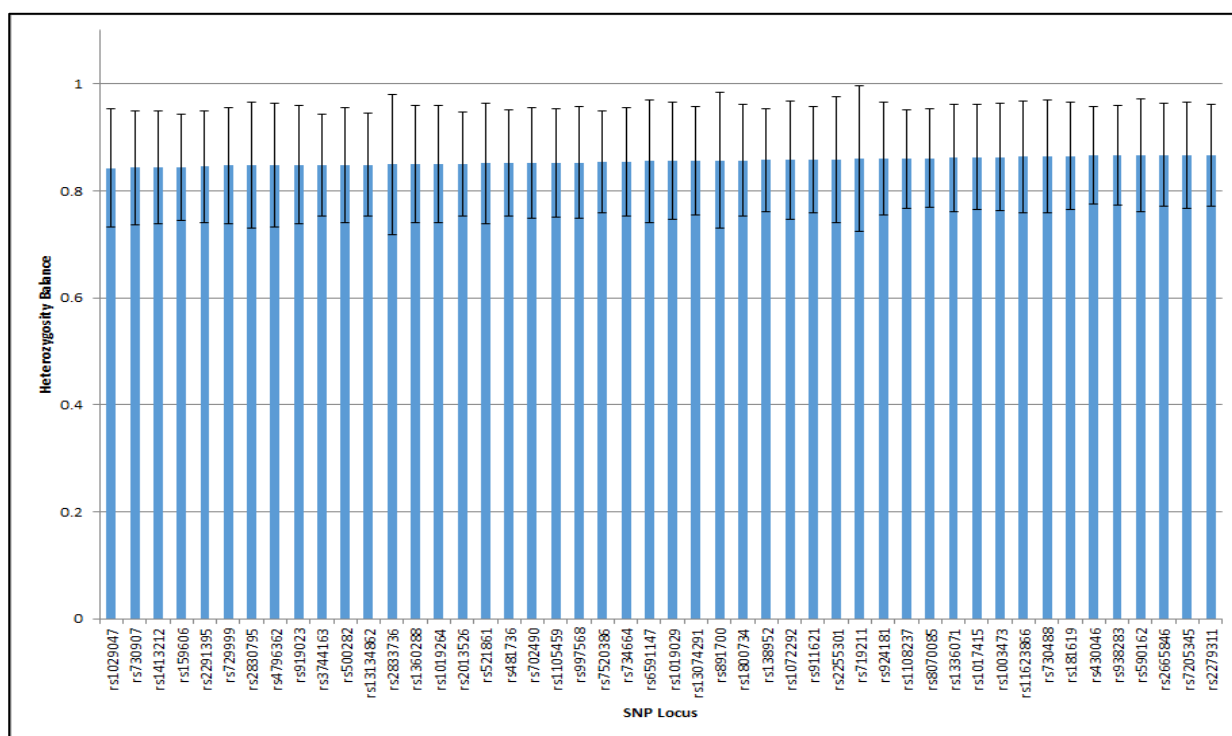
Chapter 4, Supplemental Figure 4. Relative locus performance – autosomal SNPs (part 4). Overall performance values, based on coverage, are shown for a subset of the 240 autosomal SNPs.



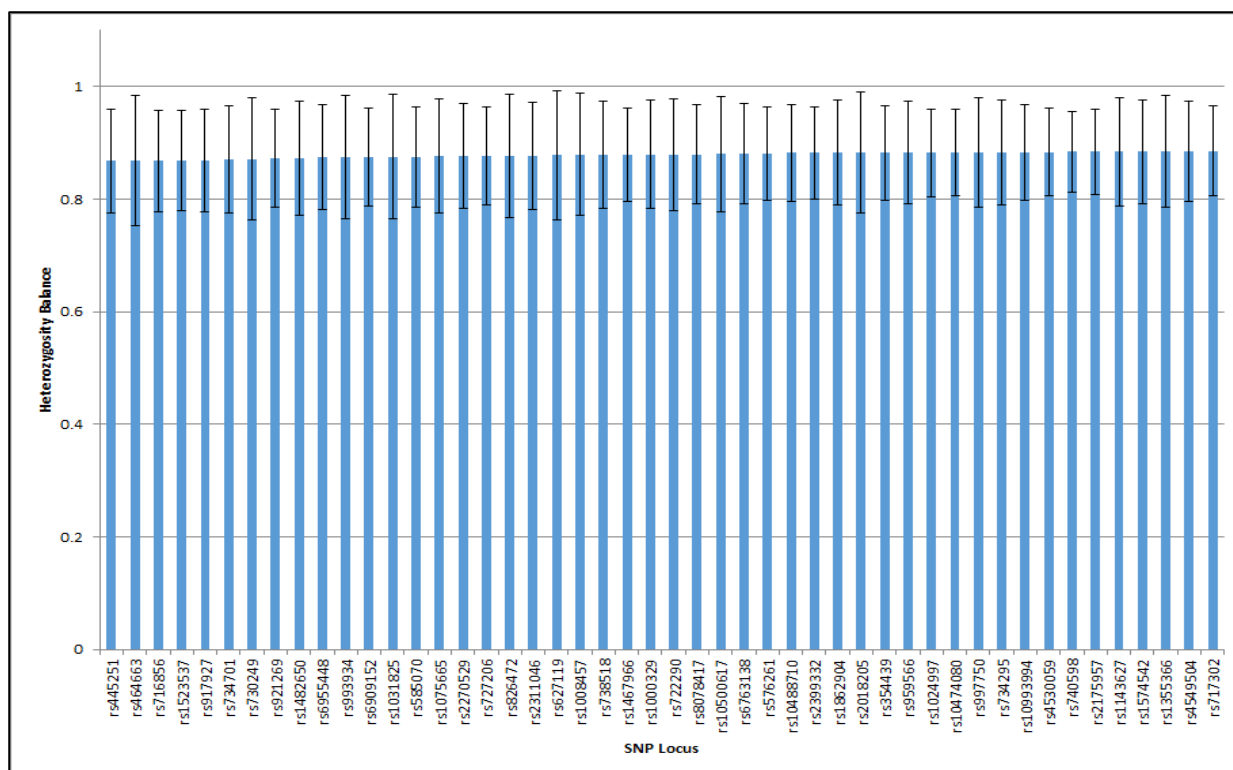
Chapter 4, Supplemental Figure 5. Relative locus performance – autosomal SNPs (part 5). Overall performance values, based on coverage, are shown for a subset of the 240 autosomal SNPs.



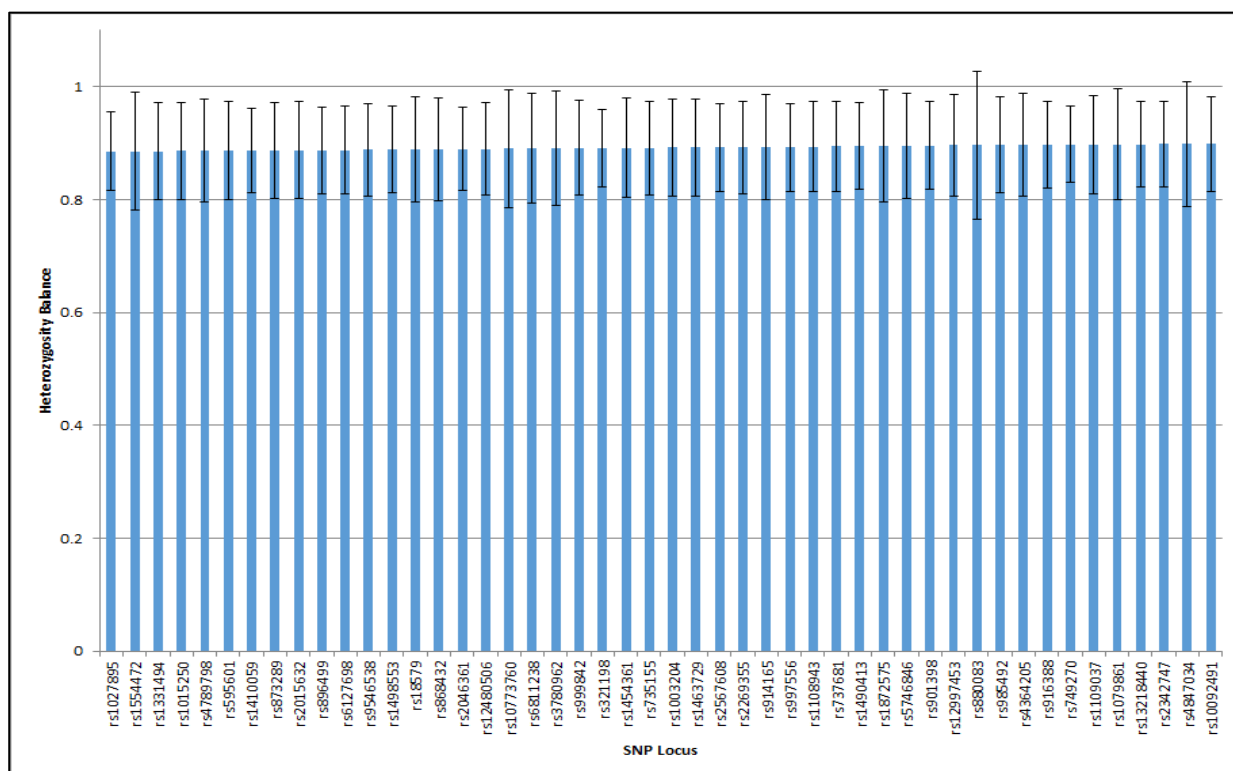
Chapter 4, Supplemental Figure 6. Heterozygosity balance – Autosomal SNPs (part 1). ACRs for heterozygous alleles, based on coverage, are shown for a subset of the 240 autosomal SNPs.



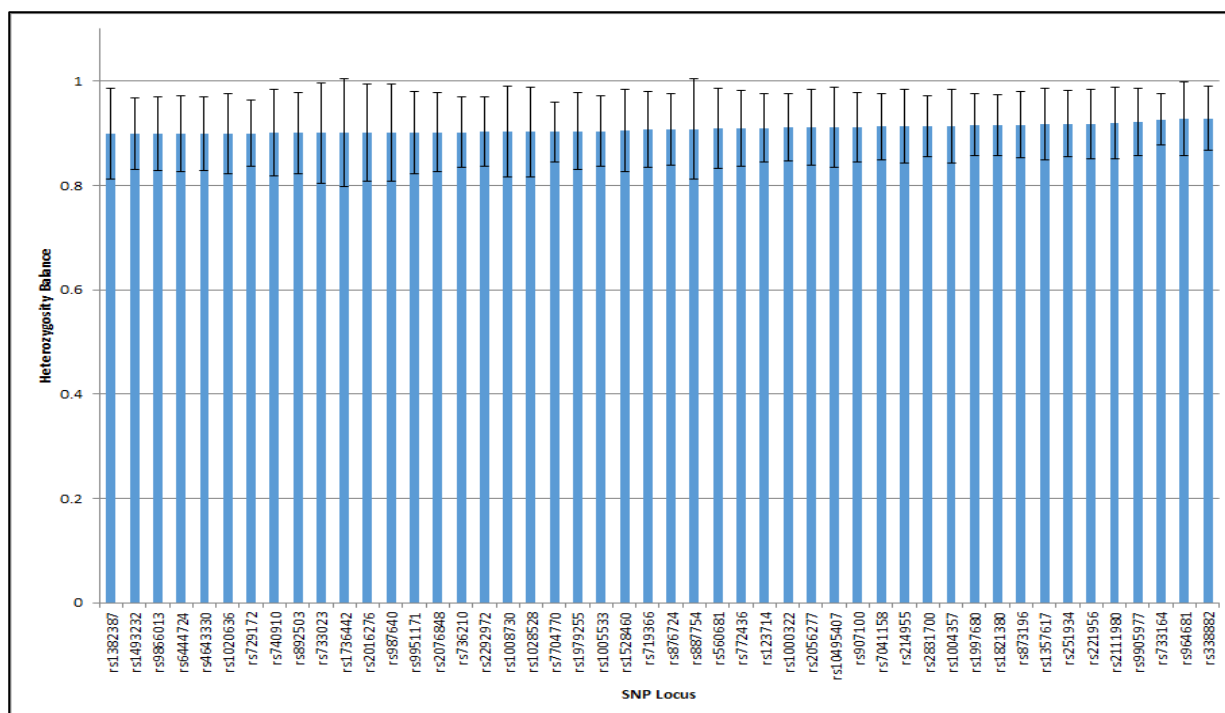
Chapter 4, Supplemental Figure 7. Heterozygosity balance – Autosomal SNPs (part 2). ACRs for heterozygous alleles, based on coverage, are shown for a subset of the 240 autosomal SNPs.



Chapter 4, Supplemental Figure 8. Heterozygosity balance – Autosomal SNPs (part 3). ACRs for heterozygous alleles, based on coverage, are shown for a subset of the 240 autosomal SNPs.



Chapter 4, Supplemental Figure 9. Heterozygosity balance – Autosomal SNPs (part 4). ACRs for heterozygous alleles, based on coverage, are shown for a subset of the 240 autosomal SNPs.



Chapter 4, Supplemental Figure 10. Heterozygosity balance – Autosomal SNPs (part 5). ACRs for heterozygous alleles, based on coverage, are shown for a subset of the 240 autosomal SNPs.

		Allele																																		
		2.2	4	5	6	7	8	8.3	9	9.3	10	11	11.3	12	12.1	12.2	13	13.2	13.3	14	14.2	15	15.2	15.3	16	16.2	16.3	17	17.1	17.2	17.3	18	18.2	18.3	18.4	19
S T R L O C I	CSF1PO	-	-	-	-	1	-	-	-	-	19	24	-	34	-	-	8	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	
	D10S1248	-	-	-	-	-	-	-	-	-	-	-	-	9	-	-	32	-	-	17	-	16	-	-	8	-	-	4	-	-	-	-	-	-		
	D12S391	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	16	-	-	-	-	-	8	-	-	-	14	-	-	22	
	D13S317	-	-	-	-	-	24	-	11	-	9	17	-	21	-	-	4	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-		
	D14S1434	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	37	-	-	44	-	-	-	-	1	-	-	-	-	-	-	-	-	-		
	D16S539	-	-	-	-	-	-	-	22	-	3	26	-	21	-	-	12	-	-	2	-	-	-	-	-	-	-	-	-	-	-	-	-	-		
	D17S1301	-	-	-	-	-	-	-	-	-	2	13	-	50	-	-	13	-	-	6	-	2	-	-	-	-	-	-	-	-	-	-	-	-		
	D18S51	-	-	-	-	-	-	-	-	-	-	-	-	5	-	-	12	-	-	9	-	18	-	-	16	-	-	6	-	-	3	-	-	5		
	D19S433	-	-	-	-	-	-	-	-	-	-	-	-	5	-	-	21	9	-	22	6	6	14	-	-	2	-	-	-	1	-	-	-	-		
	D1S1627	-	-	-	-	-	-	-	-	-	4	-	-	5	-	-	56	-	-	19	-	2	-	-	-	-	-	-	-	-	-	-	-	-		
	D1S1656	-	-	-	-	-	-	-	-	-	-	4	-	4	-	-	11	-	-	7	-	24	-	1	18	-	-	8	-	-	5	1	-	2	-	
	D21S11	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-		
	D22S1045	-	-	-	-	-	-	-	-	-	-	-	22	-	-	-	-	-	-	4	-	26	-	-	19	-	-	-	-	-	-	2	-	-		
	D2S1338	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	1	-	-	7	-	-	-	10	-	-	14	
	D2S1776	-	-	-	-	-	-	-	8	-	5	33	-	32	-	-	6	-	-	2	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	
	D2S441	-	-	-	-	-	-	-	-	-	21	25	4	16	-	-	1	-	-	18	-	1	-	-	-	-	-	-	-	-	-	-	-	-	-	
	D3S1358	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	5	-	29	-	-	31	-	-	19	-	-	-	1	-	-	1	
	D4S2408	-	-	-	-	-	21	-	25	-	26	13	-	1	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	
	D5S2500	-	-	-	-	-	-	-	1	-	3	25	-	19	-	-	1	-	-	7	-	22	-	-	6	-	-	2	-	-	-	-	-	-	-	
	D5S818	-	-	-	-	2	-	-	2	-	12	27	-	30	-	-	11	-	-	2	-	-														

[illegible]

Chapter 4, Supplemental Table 2. Autosomal STR allele counts – Asians. Counts are listed for each allele. (-) indicates an allele that was not detected at a particular locus, and therefore has no count associated with it.

		Allele																																		
		2.2	4	5	6	7	8	8.3	9	9.3	10	11	11.3	12	12.1	12.2	13	13.2	13.3	14	14.2	15	15.2	15.3	16	16.2	16.3	17	17.1	17.2	17.3	18	18.2	18.3	18.4	19
S T L O C I	CSF1PO	-	-	-	-	-	-	-	4	-	33	16	-	31	-	-	13	-	-	1	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	
	D10S1248	-	-	-	-	-	-	-	-	-	-	-	-	3	-	-	22	-	-	38	-	19	-	-	11	-	-	5	-	-	-	-	-	-	-	
	D12S391	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	1	-	-	5	-	-	-	1	23	-	-	27	
	D13S317	-	-	-	-	-	13	-	26	-	9	17	-	17	-	-	9	-	-	6	-	1	-	-	-	-	-	-	-	-	-	-	-	-	-	
	D14S1434	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	33	-	-	56	-	5	-	-	2	-	-	-	-	-	-	-	-	-	-	
	D16S539	-	-	-	-	-	1	-	9	-	14	34	-	22	-	-	17	-	-	1	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	
	D17S1301	-	-	-	-	-	-	-	-	-	2	32	-	35	-	-	24	-	-	5	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	
	D18S51	-	-	-	-	-	-	-	1	-	-	-	-	11	-	-	16	-	-	16	-	9	-	-	7	-	-	19	-	-	7	-	-	-	6	
	D19S433	-	-	-	-	-	-	-	-	-	-	-	-	2	-	1	16	5	-	30	7	22	7	-	2	1	-	-	-	1	-	-	-	-	-	
	D1S1627	-	-	-	-	-	-	-	-	-	23	10	-	3	-	-	41	-	-	20	-	1	-	-	-	-	-	-	-	-	-	-	-	-	-	
	D1S1656	-	-	-	-	-	-	-	-	-	-	3	-	3	-	-	21	-	-	7	-	9	-	1	24	-	2	4	-	-	16	1	-	6	-	
	D21S11	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	
	D22S1045	-	-	-	-	-	-	-	-	-	-	9	-	-	-	-	-	-	-	3	-	41	-	-	40	-	-	-	-	-	-	-	-	-	-	
	D2S1338	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	2	-	-	22	-	-	4	-	-	-	26	
	D2S1776	-	-	-	-	-	2	-	8	-	15	23	-	41	-	-	9	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	D2S441	-	-	-	-	-	-	-	-	-	38	20	5	7	-	-	2	-	-	23	-	3	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	D3S1358	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	1	-	-	2	-	39	-	-	26	-	-	16	-	-	14	-	-	-	-	
	D4S2408	-	-	-	-	-	18	-	55	-	15	7	-	3	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	D5S2500	-	-	-	-	-	-	-	-	-	13	37	-	13	-	-	6	-	-	3	-	18	-	-	7	-	-	-	-	-	-	1	-	-	-	-
	D5S818	-	-	-	-	7	-	-	3	-	-	1	45	-	30	-	12	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
D6S1017	-	-	-	-	2	15	-	8	-	44	7	-	20	-	-	2	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	
D6S474	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	2	-	-	27	-	13	-	-	21	-	-	21	-	-	-	13	-	-	-	1	
D7S820	-	-	-	-	-	8	-	5	-	28	34	-	19	-	-	4	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	
D8S1179	-	-	-	-	-	1	-	1	-	6	5	-	17	-	-	30	-	-	24	-	13	-	-	1	-	-	-	-	-	-	-	-	-	-	-	
FGA	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	2	-	-	8		
PENTAO	1	-	-	-	-	1	2	-	14	-	21	16	-	27	-	-	11	-	5	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	
PENTAE	-	-	3	-	10	2	-	-	-	9	7	-	16	1	-	8	-	-	5	-	9	-	-	9	-	-	5	-	-	-	4	-	-	-	2	
SE33	-	-	-	-	-	-	-	-	2	-	-	-	-	-	-	1	-	-	1	-	7	-	-	3	1	-	9	-	-	11	-	-	-	-	11	
TH01	-	-	-	25	29	4	-	14	26	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	
TPOX	-	-	-	-	-	44	-	7	-	6	30	-	10	-	-	1	-	-	-	-	6	-	-	-	-	-	-	-	-	-	-	-	-	-	-	
VWA	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	6	-	6	-	-	33	-	-	32	-	-	-	17	-	-	-	2	

[illegible]

Chapter 4, Supplemental Table 4. Autosomal STR allele counts – Hispanics. Counts are listed for each allele. (-) indicates an allele that was not detected at a particular locus, and therefore has no count associated with it.

		Allele																																				
		6	7	8	9	10	11	12	12.3	13	13.3	14	14.3	15	15.2	15.3	16	16.2	17	17.2	18	19	19.1	20	20.1	21	21.1	22	22.1	23	23.1	24	24.2	25	26	27	27.2	28
S T 	DXS10011	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	2	
	DXS10074	-	-	-	-	-	-	-	-	-	-	2	-	2	-	-	7	1	18	-	16	5	-	1	-	-	-	-	-	-	-	-	-	-	-	-	-	
	DXS101	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	3	-	6	-	10	-	14	8	4	-	4	
	DXS10101	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-		
	DXS10134	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-		
	DXS10135	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	1	-	1	8	-	4	-	7	-	4	-	2	-	4	-	4	3	2	-	4
	DXS6789	-	-	-	-	-	-	-	-	-	-	1	-	6	-	-	18	-	1	-	-	2	-	6	-	9	-	5	-	2	-	-	-	-	-	-	-	
	DXS6795	-	-	-	1	10	14	3	-	21	-	3	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	
	DXS6800	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	38	-	1	-	1	6	-	-	2	-	2	-	-	-	-	-	-	-	-	-	
	DXS6801	-	-	-	-	5	34	8	-	3	-	2	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	
	DXS6807	-	-	-	-	-	23	-	-	-	-	21	-	7	-	-	1	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	DXS6809	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	DXS6854	-	1	-	-	-	11	22	9	-	3	-	5	-	1	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	DXS7132	-	-	-	-	-	-	1	3	-	16	-	17	-	11	-	-	4	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	DXS7133	-	-	-	36	7	7	2	-	-	-	-	-	-	-	-	-	-	-	2	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
DXS7423	-	-	-	-	-	-	-	-	-	-	-	23	-	24	-	-	4	-	1	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	
DXS7424	-	-	-	-	-	1	1	-	4	-	11	-	16	-	-	13	-	6	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	
DXS8377	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	
DXS8378	-	-	-	3	30	10	8	-	1	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	
DXS981	-	-	-	-	-	-	-	3	10	12	13	6	5	-	1	2	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	
DXS9895	-	-	-	-	-	-	1	-	16	-	19	-	9	-	-	6	-	1	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	
DXS9902	-	-	-	-	2	20	19	-	10	-	1	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	
GATA165812	-	-	-	15	24	10	3	-	-	-	-	-	-	-	-	15	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	
GATA172005	1	-	14	7	20	8	2	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	
GATA31E08	-	5	1	12	14	15	4	-	1	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	
HPRTB	-	-	-	-	-	6	11	-	21	-	12	-	1	-	-	1	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	

		Allele																																																					
		28.2	29	29.2	30	30.2	31	31.2	32	32.2	33	33.2	34	34.2	35	35.2	36	36.2	37	37.2	38	39	39.3	40	41	42	43	44	45	46	47	48	49	50	51	52	53	54																	
S T L O C I	DXS10011	-	-	-	2	3	-	2	2	1	2	2	3	4	1	2	1	-	3	-	5	6	-	3	3	2	2	1	-	-	-	-	-	-	-	-	-	-	-																
	DXS10074	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-																	
	DXS101	-	1	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-																	
	DXS10101	-	5	1	6	4	13	6	6	2	7	-	2	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-																	
	DXS10134	-	-	-	-	-	-	-	3	-	4	-	7	-	10	-	12	-	11	-	4	1	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-																	
	DXS10135	-	4	-	1	-	1	-	-	-	1	-	1	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-																	
	DXS6789	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-																	
	DXS6795	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-																	
	DXS6800	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-																	
	DXS6801	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-																	
	DXS6807	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-																	
	DXS6809	-	1	-	1	-	11	-	6	-	12	-	7	-	6	-	2	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-																
	DXS6854	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-																
	DXS7132	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-																
	DXS7133	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-																
	DXS7423	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-																
	DXS7424	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-																
	DXS8377	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	3	-	5	3	5	8	12	4	-	5	3	3	1	-																
	DXS8378	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-																
	DXS981	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-																
DXS9895	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-																	
DXS9902	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-																	
GATA165812	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-																	
GATA172005	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-																	
GATA31E08	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-																	
HPRTB	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-																	

Chapter 4, Supplemental Table 6. X-chromosome STR allele counts – Asian Females. Counts are listed for each allele. (-) indicates an allele that was not detected at a particular locus, and therefore has no count associated with it.

		Allele																																				
		6	7	8	9	10	11	12	12.3	13	13.3	14	14.3	15	15.2	15.3	16	16.2	17	17.2	18	19	19.1	20	20.1	21	21.1	22	22.1	23	23.1	24	24.2	25	26	27	27.2	28
S T R	DXS10011	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	1	-	1	
	DXS10074	-	3	4	-	-	-	-	-	1	-	1	-	5	-	-	9	-	13	-	11	3	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	
	DXS101	-	-	-	-	-	-	-	-	-	-	-	-	1	-	-	1	-	-	-	8	3	-	-	-	-	-	2	-	3	-	9	-	9	2	7	-	4
	DXS10101	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	2	1	
	DXS10134	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-		
	DXS10135	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	2	-	1	1	-	3	-	6	-	3	-	5	-	5	-	4	8	4	-	2
	DXS6789	-	-	-	-	-	-	-	-	-	-	-	-	-	2	-	-	-	-	-	-	6	-	14	-	18	-	4	-	5	-	1	-	-	-	-	-	
	DXS6795	-	-	-	-	15	1	25	1	-	8	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	
	DXS6800	-	-	-	-	-	-	-	-	-	2	-	-	-	-	-	-	17	-	4	-	3	16	-	1	-	5	-	2	-	-	-	-	-	-	-	-	
	DXS6801	-	-	-	-	4	28	16	-	2	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	
L O C I	DXS6807	-	-	-	-	-	22	2	-	-	-	12	-	13	-	-	1	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	
	DXS6809	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	1	-	1
	DXS6854	-	-	-	-	-	8	4	-	22	-	11	-	4	-	-	1	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	
	DXS7132	-	-	-	-	-	-	5	-	13	-	20	-	10	-	-	2	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	
	DXS7133	-	-	-	27	5	16	-	-	1	-	1	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	
	DXS7423	-	-	-	-	-	-	-	-	5	-	9	-	25	-	-	11	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	
	DXS7424	-	-	-	-	-	1	2	-	2	-	13	-	12	-	-	12	-	8	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	
	DXS8377	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	
	DXS8378	-	-	-	1	16	17	15	-	1	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	
	DXS981	-	-	-	-	-	-	1	1	13	7	9	6	-	1	1	1	-	-	-	11	7	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
DXS9895	-	-	-	-	-	-	-	-	14	-	10	-	14	-	-	10	1	1	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	
DXS9902	-	-	-	-	2	19	15	-	14	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-		
GATA165B12	-	-	-	17	22	6	3	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-		
GATA172D05	11	-	7	3	10	10	9	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-		
GATA31E08	-	12	1	5	12	15	4	-	1	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-		
HPRTB	-	-	-	-	-	11	10	-	19	-	7	-	2	-	-	1	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	

[illegible]

Chapter 4, Supplemental Table 7. X-chromosome STR allele counts – Caucasian Females. Counts are listed for each allele. (-) indicates an allele that was not detected at a particular locus, and therefore has no count associated with it.

[illegible][illegible]

Chapter 4, Supplemental Table 8. X-chromosome STR allele counts – Hispanic Females. Counts are listed for each allele. (-) indicates an allele that was not detected at a particular locus, and therefore has no count associated with it.

[illegible][illegible]

Chapter 4, Supplemental Table 9. X-chromosome STR allele counts – African American males. Counts are listed for each allele. (-) indicates an allele that was not detected at a particular locus, and therefore has no count associated with it.

		Allele																																			
		6	7	8	9	10	11	12	12.1	12.3	13	13.3	14	14.3	15	15.2	16	16.2	17	18	19	19.1	20	21	21.1	22	23	23.1	24	25	26	26.2	27	27.2	28		
S T R L o c i	DXS10011	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-		
	DXS10074	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	3	-	7	1	5	-	1	-	-	-	-	-	-	-	-	-	-	-	-		
	DXS101	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	1	2	-	4	3	5	-	2	-	-		
	DXS10101	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	1		
	DXS10134	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-		
	DXS10135	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	3	-	1	2	-	2	-	-	2	1	1	-	-	-	-		
	DXS6789	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	5	-	-	-	-	-	6	4	-	1	-	-	-	-	-	-	-	-	-		
	DXS6795	-	-	-	-	2	6	1	-	-	7	-	1	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-		
	DXS6800	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	13	-	-	-	-	-	2	-	2	-	-	-	-	-	-	-	-	-	-	
	DXS6801	-	-	-	-	2	9	4	-	-	1	-	1	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	
	DXS6807	-	-	-	-	-	8	-	-	-	-	-	5	-	4	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	
	DXS6809	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	
	DXS6854	-	-	-	-	5	7	1	-	-	3	-	1	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	
	DXS7132	-	-	-	-	-	-	3	-	-	6	-	2	-	5	-	-	-	-	1	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	
	DXS7133	-	-	-	14	3	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	
	DXS7423	-	-	-	-	-	-	-	-	-	-	-	6	-	11	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	
	DXS7424	-	-	-	-	-	-	-	-	-	2	-	4	-	6	-	5	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	
	DXS8377	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	
	DXS8378	-	-	-	-	7	7	2	-	-	1	-	-	-	-	-	-	-	-	2	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	
	DXS981	-	-	-	-	-	-	1	-	-	2	3	8	1	2	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	DXS9895	-	-	-	-	-	-	-	-	-	7	-	4	-	4	-	2	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	
	DXS9902	-	-	-	-	-	9	5	-	-	3	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	
	GATA165B12	-	-	-	3	7	5	2	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	
	GATA172D05	1	-	2	-	7	6	1	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	
	GATA31E08	-	2	-	4	1	7	2	-	-	1	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	
	HPRTB	-	-	-	-	-	2	3	-	-	9	-	1	-	2	-	-	-	-	3	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	

		Allele																																		
		28.2	29	29.2	30	30.2	31	31.2	32	32.2	33	33.2	34	35	36	37	37.2	38	39	40	41	42	42.3	43	44	45	46	47	48	49	50	51	52	53	54	
S T R L o c i	DXS10011	-	-	-	2	1	1	-	1	-	-	-	4	1	-	-	-	-	3	2	1	-	-	-	-	-	-	-	-	-	-	-	-	-	-	
	DXS10074	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-		
	DXS101	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-		
	DXS10101	-	1	1	1	1	4	1	3	1	1	2	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-		
	DXS10134	-	-	-	-	-	-	-	1	-	-	-	-	3	4	4	-	2	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-		
	DXS10135	-	3	-	-	-	-	-	-	-	-	-	-	2	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-		
	DXS6789	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-		
	DXS6795	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-		
	DXS6800	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-		
	DXS6801	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-		
	DXS6807	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-		
	DXS6809	-	-	-	-	-	4	-	5	-	3	-	2	1	1	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	
	DXS6854	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-		
	DXS7132	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-		
	DXS7133	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-		
	DXS7423	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-		
	DXS7424	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-		
	DXS8377	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	3	1	1	1	3	3	1	-	2	1	1
	DXS8378	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	
	DXS981	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	
	DXS9895	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	
	DXS9902	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	
	GATA165B12	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	
	GATA172D05	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	
	GATA31E08	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	
	HPRTB	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	

Chapter 4, Supplemental Table 10. X-chromosome STR allele counts – Asian males. Counts are listed for each allele. (-) indicates an allele that was not detected at a particular locus, and therefore has no count associated with it.

[illegible][illegible]

Chapter 4, Supplemental Table 11. X-chromosome STR allele counts – Caucasian males. Counts are listed for each allele. (-) indicates an allele that was not detected at a particular locus, and therefore has no count associated with it.

		Allele																																	
		6	7	8	9	10	11	12	12.1	12.3	13	13.3	14	14.3	15	15.2	16	16.2	17	18	19	19.1	20	21	21.1	22	23	23.1	24	25	26	26.2	27	27.2	28
S T R L o c i	DXS10011	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	1	-	1	
	DXS10074	-	-	2	-	-	-	-	-	-	-	-	1	-	1	-	5	-	4	3	2	-	-	-	-	-	-	-	-	-	-	-	-	-	
	DXS101	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	1	1	-	-	-	-	-	1	-	4	2	1	-	3	-	2	
	DXS10101	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	
	DXS10134	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	
	DXS10135	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	1	1	1	-	3	1	4	1	-	-	-	1	-	1	-	2
	DXS6789	-	-	-	-	-	-	-	-	-	-	-	-	-	-	1	-	-	-	1	2	-	4	6	-	1	1	-	-	-	-	-	-	-	
	DXS6795	-	-	-	1	3	4	4	-	-	6	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	DXS6800	-	-	-	-	-	-	-	-	-	-	1	-	-	-	-	-	9	-	1	3	3	-	-	1	-	-	-	-	-	-	-	-	-	-
	DXS6801	-	-	-	-	6	10	-	-	-	2	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	DXS6807	-	-	-	-	-	8	-	-	-	4	-	4	-	2	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	DXS6809	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	DXS6854	-	-	-	-	1	-	2	-	-	13	-	1	-	1	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	DXS7132	-	-	-	-	-	-	3	-	-	2	-	6	-	5	-	2	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	DXS7133	-	-	-	11	3	2	2	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	DXS7423	-	-	-	-	-	-	-	-	-	-	-	4	-	9	-	3	-	2	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	DXS7424	-	-	-	-	-	1	-	-	-	2	-	3	-	5	-	4	-	1	1	1	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	DXS8377	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	DXS8378	-	-	-	-	6	6	6	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	DXS981	-	-	-	-	-	-	-	-	-	-	3	2	6	2	4	-	1	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	DXS9895	-	-	-	-	-	-	-	-	-	3	-	9	-	4	-	2	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	DXS9902	-	-	-	-	-	9	7	-	-	2	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	GATA165812	-	-	-	6	9	2	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	GATA172D05	2	-	1	2	8	3	2	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	GATA31E08	-	1	1	2	10	3	1	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	HPRTB	-	-	-	-	-	-	6	-	-	6	-	3	-	3	-	3	-	-	6	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-

		Allele																																	
		28.2	29	29.2	30	30.2	31	31.2	32	32.2	33	33.2	34	35	36	37	37.2	38	39	40	41	42	42.3	43	44	45	46	47	48	49	50	51	52	53	54
S T R L o c i	DXS10011	-	-	-	-	-	1	1	1	-	1	1	-	-	-	1	-	-	-	1	2	3	-	-	1	1	1	1	-	-	-	-	-	-	-
	DXS10074	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	DXS101	-	-	-	1	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	DXS10101	-	-	-	1	3	3	3	3	1	2	-	1	1	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	DXS10134	-	-	-	-	-	-	-	-	-	-	-	-	2	1	7	2	2	2	-	-	-	1	-	-	-	-	-	-	-	-	-	-	-	-
	DXS10135	-	1	-	-	-	-	-	1	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	DXS6789	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	DXS6795	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	DXS6800	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	DXS6801	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	DXS6807	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	DXS6809	-	-	-	-	2	-	2	-	5	-	4	2	1	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	DXS6854	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	DXS7132	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	DXS7133	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	DXS7423	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	DXS7424	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	DXS8377	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	4	-	1	-	1	2	6	2	-	1	1	-	-	-
	DXS8378	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	DXS981	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	DXS9895	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	DXS9902	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	GATA165812	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	GATA172D05	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	GATA31E08	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	HPRTB	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-

Chapter 4, Supplemental Table 12. X-chromosome STR allele counts – Hispanic males. Counts are listed for each allele. (-) indicates an allele that was not detected at a particular locus, and therefore has no count associated with it.

		Allele																																												
		8	9	10	11	12	13	13.2	14	15	16	17	17.2	18	19	20	21	22	23	24	25	26	27	28	28.2	29	30	31	32	33	34	35	36	37	38	39	40	41	42	43	44					
S T R L o c i	DYS19	-	-	-	-	-	-	-	7	7	4	4	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	
	DYS385a	-	-	-	7	1	1	-	1	5	1	5	-	2	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	
	DYS385b	-	-	-	-	1	1	-	6	3	3	1	-	7	1	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	
	DYS389	-	-	-	-	3	18	-	2	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	
	DYS389I	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	1	-	7	8	3	2	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	
	DYS390	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	12	2	5	3	1	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	
	DYS391	-	-	15	7	1	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-		
	DYS392	-	-	-	14	-	7	-	2	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-		
	DYS393	-	-	-	-	2	14	-	4	3	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-		
	DYS437	-	-	-	-	2	-	11	7	1	2	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-		
	DYS438	-	1	4	11	7	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-		
	DYS439	-	-	-	5	10	7	-	1	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	
	DYS448	-	-	-	-	-	-	-	-	-	-	-	-	-	5	5	5	1	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	
	DYS449	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	1	-	1	3	-	3	2	6	2	2	-	1	-	-	-	-	-	-	-	-	-	-	-	-	-	
	DYS456	-	-	-	-	-	-	-	4	12	5	2	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	
	DYS458	-	-	-	-	-	1	-	1	6	5	4	-	4	2	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	
	DYS460	-	-	11	11	1	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	
	DYS481	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	6	-	3	7	6	1	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	
	DYS505	-	-	1	6	5	7	-	2	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	
	DYS518	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	2	3	4	5	1	-	5	1	1	-	-	-		
	DYS522	-	-	9	9	4	-	-	1	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	
	DYS533	1	-	2	6	13	1	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	
	DYS549	-	-	1	7	7	7	-	-	1	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	
	DYS570	-	-	-	-	-	-	-	1	1	2	8	-	6	3	1	1	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	DYS576	-	-	-	-	-	2	-	-	1	8	7	-	3	2	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	DYS612	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	1	3	-	6	3	5	4	1	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	DYS635	-	-	-	-	-	-	-	-	-	-	1	-	-	-	-	8	5	7	1	-	1	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	DYS643	-	-	7	1	8	5	-	2	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	
	GATAH4	-	-	-	12	11	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	

Chapter 4, Supplemental Table 13. Y-chromosome STR allele counts – African Americans. Counts are listed for each allele. (-) indicates an allele that was not detected at a particular locus, and therefore has no count associated with it.

		Allele																																															
		8	9	10	11	12	13	13.2	14	15	16	17	17.2	18	19	20	21	22	23	24	25	26	27	28	28.2	29	30	31	32	33	34	35	36	37	38	39	40	41	42	43	44								
S T R L o c i	DYS19	-	-	-	-	-	-	-	4	9	4	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-				
	DYS385a	-	-	1	4	2	5	-	2	3	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-				
	DYS385b	-	-	-	2	1	1	-	1	-	3	1	-	1	3	2	1	1	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-			
	DYS389	-	-	-	-	8	7	-	2	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-			
	DYS389I	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	6	-	7	3	1	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-			
	DYS390	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	7	8	2	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-			
	DYS391	-	1	15	1	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-			
	DYS392	-	-	1	2	3	5	-	6	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-			
	DYS393	-	-	-	-	8	3	-	6	-	-	-	-	-	-	-	-	8	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-		
	DYS437	-	-	-	-	-	-	-	12	5	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-		
	DYS438	-	-	12	4	-	1	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-		
	DYS439	-	-	1	6	5	5	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-		
	DYS448	-	-	-	-	-	-	-	-	-	-	-	-	5	5	1	3	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	
	DYS449	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	1	-	2	-	3	1	1	2	1	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-		
	DYS456	-	-	-	-	-	-	-	2	11	4	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-		
	DYS458	-	-	-	-	-	-	-	-	4	2	2	-	6	3	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	
	DYS460	-	4	8	5	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-		
	DYS481	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	2	4	3	2	2	2	1	-	-	1	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	
	DYS505	-	-	-	9	4	2	-	2	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	
	DYS518	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	1	1	3	2	2	6	1	-	-	-	1	-	-	-	-	-			
	DYS522	-	-	2	8	4	2	-	1	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-		
	DYS533	-	-	3	9	5	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	
	DYS549	-	-	-	2	11	4	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	
	DYS570	-	-	-	-	-	-	-	-	5	6	-	2	2	2	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	
	DYS576	-	-	-	-	-	-	-	1	-	-	8	-	4	3	1	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	
	DYS612	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	3	-	-	2	3	3	3	3	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	
	DYS635	-	-	-	-	-	-	-	-	-	-	-	-	-	-	5	5	4	1	1	1	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	DYS643	2	2	3	7	3	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	
	GATAH4	-	-	2	9	6	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	

Chapter 4, Supplemental Table 14. Y-chromosome STR allele counts – Asians. Counts are listed for each allele. (-) indicates an allele that was not detected at a particular locus, and therefore has no count associated with it.

[illegible]

Chapter 4, Supplemental Table 15. Y-chromosome STR allele counts – Caucasians. Counts are listed for each allele. (-) indicates an allele that was not detected at a particular locus, and therefore has no count associated with it.

		Allele																																														
		8	9	10	11	12	13	13.2	14	15	16	17	17.2	18	19	20	21	22	23	24	25	26	27	28	28.2	29	30	31	32	33	34	35	36	37	38	39	40	41	42	43	44							
S T R L o c i	DYS19	-	-	-	-	5	-	10	1	1	1	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-		
	DYS385a	-	-	-	7	1	4	-	5	1	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-		
	DYS385b	-	-	-	-	-	1	-	7	4	1	2	-	1	2	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	
	DYS389	-	-	-	-	3	13	-	2	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	
	DYS389i	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	2	-	10	3	3	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	
	DYS390	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	2	4	11	1	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	
	DYS391	-	1	9	8	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-		
	DYS392	-	-	-	5	1	8	-	4	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-		
	DYS393	-	-	-	-	3	13	-	1	1	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	
	DYS437	-	-	-	-	-	-	-	10	6	2	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	
	DYS438	-	2	4	3	9	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	
	DYS439	-	-	1	8	6	3	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	
	DYS448	-	-	-	-	-	-	-	-	-	-	-	-	1	8	5	2	1	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	DYS449	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	1	-	1	2	1	8	1	3	1	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	DYS456	-	-	-	-	1	-	2	8	4	2	-	1	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	
	DYS458	-	-	-	-	-	-	-	2	6	5	1	3	1	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	
	DYS460	-	1	7	10	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	
	DYS481	-	-	-	-	-	-	-	-	-	-	-	-	-	-	1	6	1	7	2	1	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	DYS505	-	-	-	6	7	5	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	DYS518	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	4	3	6	3	-	1	1	-	-	-	-	-	-		
	DYS522	-	-	8	7	2	1	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	
	DYS533	-	1	-	6	9	2	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	
	DYS549	-	-	-	1	15	1	-	1	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	
	DYS570	-	-	-	-	-	-	-	-	3	9	-	2	3	-	-	1	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	DYS576	-	-	-	-	-	-	-	-	2	5	-	7	3	-	1	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	
	DYS612	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	1	-	2	4	2	6	2	-	1	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	DYS635	-	-	-	-	-	-	-	-	-	-	-	-	-	-	4	4	9	-	1	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	DYS643	1	4	7	3	3	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	
	GATAH4	-	-	1	10	6	1	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	

Chapter 4, Supplemental Table 16. Y-chromosome STR allele counts – Hispanics. Counts are listed for each allele. (-) indicates an allele that was not detected at a particular locus, and therefore has no count associated with it.

		Allele			
		A	C	G	T
S N P L o c u s	L298	23	-	-	-
	M479	-	23	-	-
	MSY2.2	-	-	22	-
	P202	23	-	-	-
	rs13447352	23	-	-	-
	rs13447443	23	-	-	-
	rs16980360	12	-	9	-
	rs16981290	-	23	-	-
	rs17250535	-	-	-	22
	rs17269816	-	23	-	-
	rs17306671	1	-	-	22
	rs20320	-	-	23	-
	rs2032595	-	1	-	22
	rs2032597	22	1	-	-
	rs2032599	-	1	-	22
	rs2032602	-	-	-	23
	rs2032624	16	7	-	-
	rs2032631	7	-	16	-

		Allele			
		A	C	G	T
S N P L o c u s	rs2032636	-	-	23	-
	rs2032652	-	15	-	8
	rs2032658	16	-	7	-
	rs2032673	-	-	-	23
	rs2032674	-	-	-	23
	rs2033003	16	7	-	-
	rs2534636	-	23	-	-
	rs34486382	-	-	21	-
	rs35284970	-	23	-	-
	rs3848982	-	9	-	14
	rs3900	-	16	7	-
	rs3911	23	-	-	-
	rs4141886	8	-	14	-
	rs8179021	-	23	-	-
	rs9341278	-	-	23	-
	rs9786139	14	-	8	-
	rs9786184	7	16	-	-

Chapter 4, Supplemental Table 21. Y-chromosome SNP allele counts – African Americans. Counts are listed for each allele. (-) indicates an allele that was not detected at a particular locus, and therefore has no count associated with it.

		Allele			
		A	C	G	T
S N P L o c u s	L298	17	-	-	-
	M479	-	17	-	-
	MSY2.2	-	-	17	-
	P202	17	-	-	-
	rs13447352	17	-	-	-
	rs13447443	13	-	4	-
	rs16980360	-	-	17	-
	rs16981290	12	5	-	-
	rs17250535	-	-	-	17
	rs17269816	-	10	-	7
	rs17306671	-	-	-	17
	rs20320	-	-	17	-
	rs2032595	-	-	-	17
	rs2032597	17	-	-	-
	rs2032599	-	-	-	17
	rs2032602	-	1	-	16
	rs2032624	17	-	-	-
	rs2032631	-	-	17	-

		Allele			
		A	C	G	T
S N P L o c u s	rs2032636	-	-	17	-
	rs2032652	-	4	-	13
	rs2032658	17	-	-	-
	rs2032673	-	-	-	17
	rs2032674	-	13	-	4
	rs2033003	4	13	-	-
	rs2534636	-	17	-	-
	rs34486382	-	-	17	-
	rs35284970	-	14	-	3
	rs3848982	-	16	-	1
	rs3900	-	4	13	-
	rs3911	17	-	-	-
	rs4141886	16	-	1	-
	rs8179021	-	17	-	-
	rs9341278	1	-	16	-
	rs9786139	4	-	13	-
	rs9786184	-	17	-	-

Chapter 4, Supplemental Table 22. Y-chromosome SNP allele counts – Asians. Counts are listed for each allele. (-) indicates an allele that was not detected at a particular locus, and therefore has no count associated with it.

		Allele			
		A	C	G	T
S N P L o c u s	L298	24	-	-	-
	M479	-	24	-	-
	MSY2.2	-	-	24	-
	P202	24	-	-	-
	rs13447352	23	1	-	-
	rs13447443	24	-	-	-
	rs16980360	2	-	22	-
	rs16981290	-	24	-	-
	rs17250535	1	-	-	23
	rs17269816	-	24	-	-
	rs17306671	6	-	-	18
	rs20320	-	-	24	-
	rs2032595	-	-	-	24
	rs2032597	19	5	-	-
	rs2032599	-	-	-	24
	rs2032602	-	-	-	24
	rs2032624	8	16	-	-
	rs2032631	16	-	8	-

		Allele			
		A	C	G	T
S N P L o c u s	rs2032636	-	-	24	-
	rs2032652	-	2	-	22
	rs2032658	8	-	16	-
	rs2032673	-	-	-	24
	rs2032674	-	-	-	24
	rs2033003	8	16	-	-
	rs2534636	-	23	-	1
	rs34486382	-	-	24	-
	rs35284970	-	24	-	-
	rs3848982	-	22	-	2
	rs3900	-	8	16	-
	rs3911	24	-	-	-
	rs4141886	22	-	2	-
	rs8179021	-	24	-	-
	rs9341278	-	-	24	-
	rs9786139	2	-	22	-
	rs9786184	15	9	-	-

Chapter 4, Supplemental Table 23. Y-chromosome SNP allele counts – Caucasians. Counts are listed for each allele. (-) indicates an allele that was not detected at a particular locus, and therefore has no count associated with it.

		Allele			
		A	C	G	T
S N P L o c u s	L298	18	-	-	-
	M479	-	18	-	-
	MSY2.2	-	-	18	-
	P202	18	-	-	-
	rs13447352	15	3	-	-
	rs13447443	18	-	-	-
	rs16980360	1	-	17	-
	rs16981290	-	18	-	-
	rs17250535	-	-	-	18
	rs17269816	-	18	-	-
	rs17306671	4	-	-	14
	rs20320	-	-	18	-
	rs2032595	-	-	-	18
	rs2032597	17	1	-	-
	rs2032599	-	-	-	18
	rs2032602	-	-	-	18
	rs2032624	10	8	-	-
	rs2032631	12	-	6	-

		Allele			
		A	C	G	T
S N P L o c u s	rs2032636	-	-	17	1
	rs2032652	-	1	-	17
	rs2032658	10	-	8	-
	rs2032673	-	-	-	18
	rs2032674	-	-	-	18
	rs2033003	6	12	-	-
	rs2534636	-	18	-	-
	rs34486382	-	-	18	-
	rs35284970	-	18	-	-
	rs3848982	-	17	-	1
	rs3900	-	6	12	-
	rs3911	18	-	-	-
	rs4141886	17	-	1	-
	rs8179021	-	14	-	4
	rs9341278	-	-	18	-
	rs9786139	2	-	16	-
	rs9786184	8	10	-	-

Chapter 4, Supplemental Table 24. Y-chromosome SNP allele counts – Hispanics. Counts are listed for each allele. (-) indicates an allele that was not detected at a particular locus, and therefore has no count associated with it.

[illegible]

Chapter 4, Supplemental Table 25. Autosomal STR allele frequencies – African Americans. Frequencies are listed for each allele. (-) indicates an allele that was not detected at a particular locus, and therefore has no frequency associated with it.

[illegible][illegible]

Chapter 4, Supplemental Table 26. Autosomal STR allele frequencies – Asians. Frequencies are listed for each allele. (-) indicates an allele that was not detected at a particular locus, and therefore has no frequency associated with it.

		Allele																																			
		2.2	4	5	6	7	8	8.3	9	9.3	10	11	11.3	12	12.1	12.2	13	13.2	13.3	14	14.2	15	15.2	15.3	16	16.2	16.3	17	17.1	17.2	17.3	18	18.2	18.3	18.4	19	
S T R L O C I	CSF1PO	-	-	-	-	-	-	-	0.020408	-	0.2551	0.26531	-	0.37755	-	-	0.071429	-	-	-	-	0.010204	-	-	-	-	-	-	-	-	-	-	-	-	-	-	
	D10S1248	-	-	-	-	-	-	-	-	-	-	0.030612	-	0.020408	-	-	0.31633	-	-	0.33673	-	0.22449	-	-	0.061224	-	-	0.010204	-	-	-	-	-	-	-	-	
	D12S391	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0.061224	-	-	-	-	-	0.040816	0.13265	-	-	0.010204	-	0.10204
	D13S317	-	-	-	-	-	-	0.11224	-	0.05102	-	0.091837	0.27551	-	0.27551	-	-	0.15306	-	-	0.040816	-	-	-	-	-	-	-	0.081633	-	-	-	-	-	-		
	D14S1434	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0.5	-	-	-	-	0.46739	-	-	-	-	-	-	-	-	-	-	-	-		
	D16S539	-	-	-	-	-	-	0.010204	-	0.13265	-	0.05102	0.22449	-	0.33673	-	0.23469	-	-	0.010204	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-		
	D17S1301	-	-	-	-	-	-	-	0.010204	-	0.030612	0.27551	-	0.45918	-	0.19388	-	-	0.020408	-	0.010204	-	-	-	-	-	-	-	-	-	-	-	-	-	-		
	D18S51	-	-	-	-	-	-	-	-	-	-	-	-	-	0.14286	-	-	0.13265	-	0.15306	-	0.13265	-	-	0.10204	-	-	-	0.14286	-	-	-	0.081633	-	-	0.091837	
	D19S433	-	-	-	-	-	-	-	-	-	-	-	-	-	0.14286	-	0.17347	-	0.37755	0.040816	0.18367	0.030612	-	0.030612	0.010204	-	-	-	-	-	-	-	-	-	-		
	D151627	-	-	-	-	-	-	-	-	-	0.22449	0.13265	-	0.010204	-	-	0.31633	-	-	0.30612	-	-	-	-	0.010204	-	-	-	-	-	-	-	-	-	-	-	
	D151656	-	-	-	-	-	-	-	-	-	-	-	-	0.040816	-	0.091837	-	0.091837	-	-	0.091837	-	-	0.20408	-	0.010204	0.13265	-	0.071429	0.05102	0.010204	-	0.13265	0.010204	-	0.040816	
	D21S11	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-		
	D22S1045	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0.010204	-	-	0.05102	-	0.38776	-	0.37755	-	-	0.061224	-	-	-	-	-	-	-		
	D251338	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0.05102	-	-	0.22449	-	-	-	0.091837	-	-	-	0.091837	
	D251776	-	-	-	-	-	-	0.081633	-	0.091837	-	0.11224	0.27551	-	0.35714	-	0.081633	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-		
	DZ5441	-	-	-	-	-	-	-	-	-	-	0.20408	0.16333	0.061224	0.020408	-	0.040816	-	-	0.32653	-	0.030612	-	-	-	-	-	-	-	-	-	-	-	-	-		
	D3S1358	-	-	-</																																	

[illegible]

Chapter 4, Supplemental Table 27. Autosomal STR allele frequencies – Caucasians. Frequencies are listed for each allele. (-) indicates an allele that was not detected at a particular locus, and therefore has no frequency associated with it.

		Allele																																		
		2.2	4	5	6	7	8	8.3	9	9.3	10	11	11.3	12	12.1	12.2	13	13.2	13.3	14	14.2	15	15.2	15.3	16	16.2	16.3	17	17.1	17.2	17.3	18	18.2	18.3	18.4	19
S T R L O C I	CSF1PO	-	-	-	-	-	-	-	0.040816	-	0.33673	0.16327	-	0.31633	-	-	0.13265	-	-	0.010204	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	
	D10S1248	-	-	-	-	-	-	-	-	-	-	-	-	0.030612	-	-	0.22449	-	-	0.38776	-	0.19388	-	-	0.11224	-	-	0.05102	-	-	-	-	-	-	-	
	D12S391	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0.010204	-	-	0.05102	-	-	-	0.010204	0.23469	-	-	0.27551
	D13S317	-	-	-	-	-	0.13265	-	0.26531	-	0.091837	0.17347	-	0.17347	-	-	0.091837	-	-	0.061224	-	0.010204	-	-	-	-	-	-	-	-	-	-	-	-	-	
	D14S1434	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0.34375	-	-	0.58333	-	0.052083	-	-	0.020833	-	-	-	-	-	-	-	-	-	-	
	D16S539	-	-	-	-	-	0.010204	-	0.091837	-	0.14286	0.34694	-	0.22449	-	-	0.17347	-	-	0.010204	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	
	D17S1301	-	-	-	-	-	-	-	-	-	0.020408	0.32653	-	0.35714	-	-	0.2449	-	-	0.05102	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	
	D18S51	-	-	-	-	-	-	-	0.010204	-	-	-	-	-	0.11224	-	0.16327	-	0.16327	0.091837	-	-	0.071429	-	-	-	-	0.19388	-	-	-	0.071429	-	-	0.061224	
	D19S433	-	-	-	-	-	-	-	-	-	-	-	-	-	0.020833	-	0.010417	0.16667	0.052083	0.3125	0.072917	0.072917	-	-	0.020833	0.010417	-	-	0.010417	-	-	-	-	-	-	
	D1S1627	-	-	-	-	-	-	-	-	-	0.23469	0.10204	-	0.030612	-	-	0.41837	-	-	0.20408	-	0.010204	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	D1S1656	-	-	-	-	-	-	-	-	-	-	-	0.030612	-	0.030612	-	0.21429	-	-	0.071429	-	0.091837	-	0.010204	0.2449	-	0.020408	0.040816	-	-	0.16327	0.010204	-	0.061224	-	-
	D21S11	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	
	D22S1045	-	-	-	-	-	-	-	-	-	-	-	0.091837	-	-	-	-	-	-	-	0.030612	-	0.41837	-	-	0.40816	-	-	0.05102	-	-	-	-	-	-	
	D2S1338	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0.020408	-	-	0.22449	-	-	-	0.040816	-	0.26531	
	D2S1776	-	-	-	-	-	-	0.020408	-	0.081633	-	0.15306	0.23469	-	0.41837	-	-	0.091837	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	
	D2S441	-	-	-	-	-	-	-	-	-	-	0.38776	0.20408	0.05102	0.071429	-	-	0.020408	-	-	0.23469	-	0.030612	-	-	-	-	-	-	-	-	-	-	-	-	-
	D3S1358	-	-	-	-</																															

		Allele																																		
		19.2	19.3	20	20.2	21	21.2	22	22.2	23	23.2	24	24.2	25	25.2	26	26.2	27	27.2	28	28.2	28.3	29	29.2	30	30.2	31	31.1	31.2	32	32.2	33	33.2	34.2	35	
S T R L O C I	CSF1PO	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	
	D10S1248	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-		
	D12S391	-	0.010204	0.16327	-	0.11224	-	0.061224	-	0.030612	-	-	-	0.030612	-	0.010204	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	
	D13S317	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-		
	D14S1434	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-		
	D16S539	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-		
	D17S1301	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-		
	D18S51	-	-	0.020408	-	0.020408	-	0.010204	-	-	-	-	-	-	0.010204	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	
	D19S433	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0.020833	-	-	-	-	-	-	-	-	-	-	-	-	-	
	D151627	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-		
	D151656	-	0.010204	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	
	D21S11	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0.071429	-	-	0.2551	-	0.2449	0.020408	0.05102	-	0.091837	0.010204	0.19388	0.010204	0.020408	0.020408	0.010204	
	D22S1045	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	
	D251338	-	-	0.061224	-	0.020408	-	0.081633	-	0.19388	-	0.061224	-	0.030612	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	
	D251776	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	
	D25441	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	
	D3S1358	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	
	D4S2408	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	
	D5S2500	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	
	D5S818	-	-	-	-																															

Chapter 4, Supplemental Table 28. Autosomal STR allele frequencies – Hispanics. Frequencies are listed for each allele. (-) indicates an allele that was not detected at a particular locus, and therefore has no frequency associated with it.

Chapter 4, Supplemental Table 32. X-chromosome STR allele frequencies – Hispanic females. Frequencies are listed for each allele. (-) indicates an allele that was not detected at a particular locus, and therefore has no frequency associated with it.

		Allele																																	
		28.2	29	29.2	30	30.2	31	31.2	32	32.2	33	33.2	34	35	36	37	37.2	38	39	40	41	42	42.3	43	44	45	46	47	48	49	50	51	52	53	54
S T R L o c i	DXS10011	-	0.090909	0.090909	0.045455	0.045455	-	0.18182	0.090909	0.090909	0.045455	0.045455	0.045455	-	0.090909	-	-	-	0.045455	0.090909	-	-	42.3	43	44	45	46	47	48	49	50	51	52	53	54
	DXS10074	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	DXS101	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	
	DXS10101	-	0.17391	0.13043	0.21739	0.043478	0.13043	0.043478	0.13043	0.043478	0.043478	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	DXS10134	-	-	-	-	-	-	-	0.05	-	0.15	-	0.15	0.2	0.15	0.25	-	0.05	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	DXS10135	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	
	DXS6789	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	
	DXS6795	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	
	DXS6800	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	
	DXS6801	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	
	DXS6807	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	
	DXS6809	-	0.05	-	0.05	-	0.15	-	0.15	-	0.25	-	0.05	0.2	0.05	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	DXS6854	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	DXS7132	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	DXS7133	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	DXS7423	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	DXS7424	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	DXS8377	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0.045455	-	-	-	0.13636	0.13636	0.090909	0.045455	0.090909	0.18182	0.13636	0.045455	0.045455	0.045455
DXS8378	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0.043478	-	-	-
DXS981	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	
DXS985	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	
DXS9902	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	
GATA165812	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	
GATA172D05	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	
GATA31E08	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	
HPRTB	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	

Chapter 4, Supplemental Table 33. X-chromosome STR allele frequencies – African American males. Frequencies are listed for each allele. (-) indicates an allele that was not detected at a particular locus, and therefore has no frequency associated with it.

		Allele																				
		8	9	10	11	12	13	13.2	14	15	16	17	17.2	18	19	20	21	22	23	24	25	
S T R L O C i	DYS19	-	-	-	-	-	-	-	0.31818	0.31818	0.18182	0.18182	-	-	-	-	-	-	-	-	-	
	DYS385a	-	-	-	0.30435	0.043478	0.043478	-	0.043478	0.21739	0.043478	0.21739	-	0.086957	-	-	-	-	-	-	-	
	DYS385b	-	-	-	-	0.043478	0.043478	-	0.26087	0.13043	0.13043	0.043478	-	0.30435	0.043478	-	-	-	-	-	-	
	DYS389I	-	-	-	-	0.13043	0.78261	-	0.086957	-	-	-	-	0.30435	-	-	-	-	-	-	-	
	DYS389II	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	
	DYS390	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0.52174	0.086957	0.21739	0.13043	0.043478	
	DYS391	-	-	0.65217	0.30435	0.043478	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	
	DYS392	-	-	-	0.6087	-	0.30435	-	0.086957	-	-	-	-	-	-	-	-	-	-	-	-	
	DYS393	-	-	-	-	0.086957	0.6087	-	0.17391	0.13043	-	-	-	-	-	-	-	-	-	-	-	
	DYS437	-	-	-	-	-	0.086957	-	0.47826	0.30435	0.043478	0.086957	-	-	-	-	-	-	-	-	-	
	DYS438	-	0.043478	0.17391	0.47826	0.30435	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	
	DYS439	-	-	-	0.21739	0.43478	0.30435	-	0.043478	-	-	-	-	-	-	-	-	-	-	-	-	
	DYS448	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0.3125	0.3125	0.3125	0.0625	-	-	
	DYS449	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	
	DYS456	-	-	-	-	-	-	-	-	0.17391	0.52174	0.21739	0.086957	-	-	-	-	-	-	-	0.047619	
	DYS458	-	-	-	-	-	0.043478	-	0.043478	0.26087	0.21739	0.17391	-	0.17391	0.086957	-	-	-	-	-	-	
	DYS460	-	-	0.47826	0.47826	0.043478	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	
	DYS481	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0.26087	-	0.13043	0.30435
	DYS505	-	-	0.047619	0.28571	0.2381	0.33333	-	0.095238	-	-	-	-	-	-	-	-	-	-	-	-	
	DYS518	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	
	DYS522	-	-	0.3913	0.3913	0.17391	-	-	0.043478	-	-	-	-	-	-	-	-	-	-	-	-	
	DYS533	0.043478	-	0.086957	0.26087	0.56522	0.043478	-	-	-	-	-	-	-	-	-	-	-	-	-	-	
	DYS549	-	-	0.043478	0.30435	0.30435	0.30435	-	-	0.043478	-	-	-	-	-	-	-	-	-	-	-	
	DYS570	-	-	-	-	-	-	-	-	0.043478	0.086957	0.34783	-	0.26087	0.13043	0.043478	0.043478	-	-	-	-	
	DYS576	-	-	-	-	-	0.086957	-	-	0.043478	0.34783	0.30435	-	0.13043	0.086957	-	-	-	-	-	-	
DYS612	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-		
DYS635	-	-	-	-	-	-	-	-	-	-	-	0.043478	-	-	-	-	0.34783	0.21739	0.30435	0.043478	-	
DYS643	-	-	0.30435	0.043478	0.34783	0.21739	-	0.086957	-	-	-	-	-	-	-	-	-	-	-	-		
GATAH4	-	-	-	0.52174	0.47826	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-		

[illegible]

Chapter 4, Supplemental Table 37. Y-chromosome STR allele frequencies – African Americans. Frequencies are listed for each allele. (-) indicates an allele that was not detected at a particular locus, and therefore has no frequency associated with it.

		Allele																			
		8	9	10	11	12	13	13.2	14	15	16	17	17.2	18	19	20	21	22	23	24	25
S T R L O C I	DYS19	-	-	-	-	-	-	-	0.23529	0.52941	0.23529	-	-	-	-	-	-	-	-	-	-
	DYS385a	-	-	0.058824	0.23529	0.11765	0.29412	-	0.11765	0.17647	-	-	-	-	-	-	-	-	-	-	-
	DYS385b	-	-	-	0.11765	0.058824	0.058824	-	0.058824	-	0.17647	0.058824	-	0.058824	0.17647	0.11765	0.058824	0.058824	-	-	-
	DYS389I	-	-	-	-	0.47059	0.41176	-	0.11765	-	-	-	-	-	-	-	-	-	-	-	-
	DYS389II	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	DYS390	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0.41176	0.47059	0.11765
	DYS391	-	0.058824	0.88235	0.058824	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	DYS392	-	-	0.058824	0.11765	0.17647	0.29412	-	0.35294	-	-	-	-	-	-	-	-	-	-	-	-
	DYS393	-	-	-	-	0.47059	0.17647	-	0.35294	-	-	-	-	-	-	-	-	-	-	-	-
	DYS437	-	-	-	-	-	-	-	0.70588	0.29412	-	-	-	-	-	-	-	-	-	-	-
	DYS438	-	-	0.70588	0.23529	-	0.058824	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	DYS439	-	-	0.058824	0.35294	0.29412	0.29412	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	DYS448	-	-	-	-	-	-	-	-	-	-	-	-	0.35714	0.35714	0.071429	0.21429	-	-	-	-
	DYS449	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	DYS456	-	-	-	-	-	-	-	0.11765	0.64706	0.23529	-	-	-	-	-	-	-	-	-	-
	DYS458	-	-	-	-	-	-	-	-	0.23529	0.11765	0.11765	-	0.35294	0.17647	-	-	-	-	-	-
	DYS460	-	0.23529	0.47059	0.29412	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	DYS481	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0.11765	0.23529	0.17647	0.11765
	DYS505	-	-	-	0.52941	0.23529	0.11765	-	0.11765	-	-	-	-	-	-	-	-	-	-	-	-
	DYS518	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	DYS522	-	-	0.11765	0.47059	0.23529	0.11765	-	0.058824	-	-	-	-	-	-	-	-	-	-	-	-
	DYS533	-	-	0.17647	0.52941	0.29412	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	DYS549	-	-	-	0.11765	0.64706	0.23529	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	DYS570	-	-	-	-	-	-	-	-	-	0.29412	0.35294	-	0.11765	0.11765	0.11765	-	-	-	-	-
	DYS576	-	-	-	-	-	-	-	0.058824	-	-	0.47059	-	0.23529	0.17647	0.058824	-	-	-	-	-
	DYS612	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	DYS635	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0.29412	0.29412	0.23529	0.058824	0.058824	0.058824
	DYS643	0.11765	0.11765	0.17647	0.41176	0.17647	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	GATAH4	-	-	0.11765	0.52941	0.35294	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-

		Allele																			
		26	27	28	28.2	29	30	31	32	33	34	35	36	37	38	39	40	41	42	43	44
S T R L O C I	DYS19	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	DYS385a	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	DYS385b	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	DYS389I	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	DYS389II	-	-	0.35294	-	0.41176	0.17647	0.058824	-	-	-	-	-	-	-	-	-	-	-	-	-
	DYS390	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	DYS391	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	DYS392	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	DYS393	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	DYS437	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	DYS438	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	DYS439	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	DYS448	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	DYS449	0.090909	-	0.18182	-	0.27273	0.090909	0.090909	0.18182	0.090909	-	-	-	-	-	-	-	-	-	-	-
	DYS456	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	DYS458	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	DYS460	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	DYS481	0.11765	0.11765	0.058824	-	-	0.058824	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	DYS505	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	DYS518	-	-	-	-	-	-	-	-	-	0.058824	0.058824	0.17647	0.11765	0.11765	0.35294	0.058824	-	-	0.058824	-
	DYS522	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	DYS533	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	DYS549	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	DYS570	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	DYS576	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	DYS612	-	0.17647	-	-	0.11765	0.17647	0.17647	0.17647	0.17647	-	-	-	-	-	-	-	-	-	-	-
	DYS635	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	DYS643	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	GATAH4	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-

Chapter 4, Supplemental Table 38. Y-chromosome STR allele frequencies – Asians. Frequencies are listed for each allele. (-) indicates an allele that was not detected at a particular locus, and therefore has no frequency associated with it.

		Allele																			
		8	9	10	11	12	13	13.2	14	15	16	17	17.2	18	19	20	21	22	23	24	25
S T R L O C I	DYS19	-	-	-	-	-	-	-	0.625	0.25	0.125	-	-	-	-	-	-	-	-	-	-
	DYS385a	-	-	0.041667	0.58333	0.16667	0.041667	-	0.083333	0.041667	0.041667	-	-	-	-	-	-	-	-	-	-
	DYS385b	-	-	-	-	-	-	0.041667	0.29167	0.375	-	0.16667	-	-	-	-	-	-	-	-	-
	DYS389I	-	0.083333	-	-	0.20833	0.58333	-	0.083333	0.041667	-	-	-	-	-	-	-	-	-	-	-
	DYS389II	-	-	-	-	-	-	-	-	0.043478	-	-	-	-	-	-	-	-	-	-	-
	DYS390	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0.16667	0.29167	0.33333	0.16667
	DYS391	-	-	0.45833	0.45833	0.083333	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	DYS392	-	-	-	0.375	-	0.54167	-	0.083333	-	-	-	-	-	-	-	-	-	-	-	-
	DYS393	-	-	-	-	0.125	0.79167	-	0.083333	-	-	-	-	-	-	-	-	-	-	-	-
	DYS437	-	-	-	-	-	-	-	0.20833	0.58333	0.20833	-	-	-	-	-	-	-	-	-	-
	DYS438	-	0.041667	0.16667	0.20833	0.54167	0.041667	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	DYS439	-	-	0.041667	0.41667	0.45833	0.083333	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	DYS448	-	-	-	-	-	-	-	-	-	-	-	-	0.058824	0.58824	0.29412	0.058824	-	-	-	-
	DYS449	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	DYS456	-	-	-	-	-	-	-	0.041667	0.54167	0.29167	0.125	-	-	-	-	-	-	-	-	-
	DYS458	-	-	-	-	-	-	-	-	0.125	0.16667	0.54167	-	0.125	0.041667	-	-	-	-	-	-
	DYS460	-	0.083333	0.33333	0.54167	0.041667	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	DYS481	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0.041667	0.54167	-	0.041667	0.25
	DYS505	-	-	-	0.16667	0.83333	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	DYS518	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	DYS522	-	-	0.33333	0.45833	0.20833	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	DYS533	0.041667	-	-	0.25	0.5	0.125	-	0.083333	-	-	-	-	-	-	-	-	-	-	-	-
	DYS549	-	-	0.041667	0.041667	0.45833	0.41667	-	0.041667	-	-	-	-	-	-	-	-	-	-	-	-
	DYS570	-	-	-	-	-	-	-	0.083333	-	0.16667	0.33333	-	0.25	0.041667	0.125	-	-	-	-	-
	DYS576	-	-	-	-	-	-	-	-	0.041667	0.375	0.083333	-	0.375	0.125	-	-	-	-	-	-
	DYS612	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	DYS635	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0.083333	0.125	0.58333	0.041667	0.16667
	DYS643	-	0.041667	0.625	0.125	0.083333	0.125	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	GATAH4	-	-	0.041667	0.58333	0.33333	0.041667	-	-	-	-	-	-	-	-	-	-	-	-	-	-

		Allele																			
		26	27	28	28.2	29	30	31	32	33	34	35	36	37	38	39	40	41	42	43	44
S T R L O C I	DYS19	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	DYS385a	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	DYS385b	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	DYS389I	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	DYS389II	-	-	0.086957	-	0.56522	0.21739	0.043478	0.043478	-	-	-	-	-	-	-	-	-	-	-	-
	DYS390	0.041667	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	DYS391	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	DYS392	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	DYS393	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	DYS437	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	DYS438	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	DYS439	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	DYS448	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	DYS449	-	-	0.16667	-	0.45833	0.20833	0.083333	-	0.041667	0.041667	-	-	-	-	-	-	-	-	-	-
	DYS456	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	DYS458	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	DYS460	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	DYS481	0.125	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	DYS505	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	DYS518	-	-	-	-	-	-	-	-	-	-	-	0.086957	0.21739	0.17391	0.30435	0.086957	0.13043	-	-	-
	DYS522	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	DYS533	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	DYS549	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	DYS570	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	DYS576	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	DYS612	-	0.041667	0.083333	-	0.125	0.25	0.16667	0.29167	0.041667	-	-	-	-	-	-	-	-	-	-	-
	DYS635	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	DYS643	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	GATAH4	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-

Chapter 4, Supplemental Table 39. Y-chromosome STR allele frequencies – Caucasians. Frequencies are listed for each allele. (-) indicates an allele that was not detected at a particular locus, and therefore has no frequency associated with it.

		Allele																			
		8	9	10	11	12	13	13.2	14	15	16	17	17.2	18	19	20	21	22	23	24	25
S T R L O C I	DYS19	-	-	-	-	-	0.27778	-	0.55556	0.055556	0.055556	0.055556	-	-	-	-	-	-	-	-	-
	DYS385a	-	-	-	0.38889	0.055556	0.22222	-	0.27778	0.055556	-	-	-	-	-	-	-	-	-	-	-
	DYS385b	-	-	-	-	-	0.055556	-	0.38889	0.22222	0.055556	0.11111	-	0.055556	0.11111	-	-	-	-	-	-
	DYS389I	-	-	-	-	0.16667	0.72222	-	0.11111	-	-	-	-	-	-	-	-	-	-	-	-
	DYS389II	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	DYS390	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0.11111	0.22222	0.61111	0.055556
	DYS391	-	0.055556	0.5	0.44444	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	DYS392	-	-	-	0.27778	0.055556	0.44444	-	0.22222	-	-	-	-	-	-	-	-	-	-	-	-
	DYS393	-	-	-	-	0.16667	0.72222	-	0.055556	0.055556	-	-	-	-	-	-	-	-	-	-	-
	DYS437	-	-	-	-	-	-	-	0.55556	0.33333	0.11111	-	-	-	-	-	-	-	-	-	-
	DYS438	-	0.11111	0.22222	0.16667	0.5	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	DYS439	-	-	0.055556	0.44444	0.33333	0.16667	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	DYS448	-	-	-	-	-	-	-	-	-	-	-	-	0.058824	0.47059	0.29412	0.11765	0.058824	-	-	-
	DYS449	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0.055556
	DYS456	-	-	-	-	-	0.055556	-	0.11111	0.44444	0.22222	0.11111	-	0.055556	-	-	-	-	-	-	-
	DYS458	-	-	-	-	-	-	-	-	0.11111	0.33333	0.27778	0.055556	0.16667	0.055556	-	-	-	-	-	-
	DYS460	-	0.055556	0.38889	0.55556	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	DYS481	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0.055556	0.33333	0.055556	0.38889	0.11111
	DYS505	-	-	-	0.33333	0.38889	0.27778	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	DYS518	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	DYS522	-	-	0.44444	0.38889	0.11111	0.055556	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	DYS533	-	0.055556	-	0.33333	0.5	0.11111	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	DYS549	-	-	-	0.055556	0.83333	0.055556	-	0.055556	-	-	-	-	-	-	-	-	-	-	-	-
	DYS570	-	-	-	-	-	-	-	-	-	0.16667	0.5	-	0.11111	0.16667	-	-	-	0.055556	-	-
	DYS576	-	-	-	-	-	-	-	-	-	0.11111	0.27778	-	0.38889	0.16667	-	-	0.055556	-	-	-
	DYS612	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	DYS635	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0.22222	0.22222	0.5	-	0.055556
	DYS643	0.055556	0.22222	0.38889	0.16667	0.16667	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	GATAH4	-	-	0.055556	0.55556	0.33333	0.055556	-	-	-	-	-	-	-	-	-	-	-	-	-	-

		Allele																			
		26	27	28	28.2	29	30	31	32	33	34	35	36	37	38	39	40	41	42	43	44
S T R L O C I	DYS19	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	DYS385a	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	DYS385b	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	DYS389I	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	DYS389II	-	-	0.11111	-	0.55556	0.16667	0.16667	-	-	-	-	-	-	-	-	-	-	-	-	-
	DYS390	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	DYS391	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	DYS392	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	DYS393	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	DYS437	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	DYS438	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	DYS439	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	DYS448	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	DYS449	-	0.055556	0.11111	0.055556	0.44444	0.055556	0.16667	0.055556	-	-	-	-	-	-	-	-	-	-	-	-
	DYS456	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	DYS458	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	DYS460	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	DYS481	0.055556	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	DYS505	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	DYS518	-	-	-	-	-	-	-	-	-	-	-	0.22222	0.16667	0.33333	0.16667	-	0.055556	0.055556	-	-
	DYS522	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	DYS533	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	DYS549	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	DYS570	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	DYS576	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	DYS612	-	-	0.055556	-	0.11111	0.22222	0.11111	0.33333	0.11111	-	0.055556	-	-	-	-	-	-	-	-	-
	DYS635	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	DYS643	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	GATAH4	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-

Chapter 4, Supplemental Table 40. Y-chromosome STR allele frequencies – Hispanics. Frequencies are listed for each allele. (-) indicates an allele that was not detected at a particular locus, and therefore has no frequency associated with it.

		Allele			
		A	C	G	T
S N P L O C U S	L298	1	-	-	-
	M479	-	1	-	-
	MSY2.2	-	-	1	-
	P202	1	-	-	-
	rs13447352	1	-	-	-
	rs13447443	1	-	-	-
	rs16980360	0.6087	-	0.3913	-
	rs16981290	-	1	-	-
	rs17250535	-	-	-	1
	rs17269816	-	1	-	-
	rs17306671	0.043478	-	-	0.95652
	rs20320	-	-	1	-
	rs2032595	-	0.043478	-	0.95652
	rs2032597	0.95652	0.043478	-	-
	rs2032599	-	0.043478	-	0.95652
	rs2032602	-	-	-	1
	rs2032624	0.69565	0.30435	-	-
	rs2032631	0.30435	-	0.69565	-

		Allele			
		A	C	G	T
S N P L O C U S	rs2032636	-	-	1	-
	rs2032652	-	0.65217	-	0.34783
	rs2032658	0.69565	-	0.30435	-
	rs2032673	-	-	-	1
	rs2032674	-	-	-	1
	rs2033003	0.69565	0.30435	-	-
	rs2534636	-	1	-	-
	rs34486382	-	-	1	-
	rs35284970	-	1	-	-
	rs3848982	-	0.3913	-	0.6087
	rs3900	-	0.69565	0.30435	-
	rs3911	1	-	-	-
	rs4141886	0.34783	-	0.65217	-
	rs8179021	-	1	-	-
	rs9341278	-	-	1	-
	rs9786139	0.65217	-	0.34783	-
	rs9786184	0.30435	0.69565	-	-

Chapter 4, Supplemental Table 45. Y-chromosome SNP allele frequencies – African Americans. Frequencies are listed for each allele. (-) indicates an allele that was not detected at a particular locus, and therefore has no frequency associated with it.

		Allele			
		A	C	G	T
S N P L O C U S	L298	1	-	-	-
	M479	-	1	-	-
	MSY2.2	-	-	1	-
	P202	1	-	-	-
	rs13447352	1	-	-	-
	rs13447443	0.76471	-	0.23529	-
	rs16980360	-	-	1	-
	rs16981290	0.70588	0.29412	-	-
	rs17250535	-	-	-	1
	rs17269816	-	0.58824	-	0.41176
	rs17306671	-	-	-	1
	rs20320	-	-	1	-
	rs2032595	-	-	-	1
	rs2032597	1	-	-	-
	rs2032599	-	-	-	1
	rs2032602	-	0.058824	-	0.94118
	rs2032624	1	-	-	-
	rs2032631	-	-	1	-

		Allele			
		A	C	G	T
S N P L O C U S	rs2032636	-	-	1	-
	rs2032652	-	0.23529	-	0.76471
	rs2032658	1	-	-	-
	rs2032673	-	-	-	1
	rs2032674	-	0.76471	-	0.23529
	rs2033003	0.23529	0.76471	-	-
	rs2534636	-	1	-	-
	rs34486382	-	-	1	-
	rs35284970	-	0.82353	-	0.17647
	rs3848982	-	0.94118	-	0.058824
	rs3900	-	0.23529	0.76471	-
	rs3911	1	-	-	-
	rs4141886	0.94118	-	0.058824	-
	rs8179021	-	1	-	-
	rs9341278	0.058824	-	0.94118	-
	rs9786139	0.23529	-	0.76471	-
	rs9786184	-	1	-	-

Chapter 4, Supplemental Table 46. Y-chromosome SNP allele frequencies – Asians. Frequencies are listed for each allele. (-) indicates an allele that was not detected at a particular locus, and therefore has no frequency associated with it.

		Allele			
		A	C	G	T
S N P L o c u s	L298	1	-	-	-
	M479	-	1	-	-
	MSY2.2	-	-	1	-
	P202	1	-	-	-
	rs13447352	0.95833	0.041667	-	-
	rs13447443	1	-	-	-
	rs16980360	0.083333	-	0.91667	-
	rs16981290	-	1	-	-
	rs17250535	0.041667	-	-	0.95833
	rs17269816	-	1	-	-
	rs17306671	0.25	-	-	0.75
	rs20320	-	-	1	-
	rs2032595	-	-	-	1
	rs2032597	0.79167	0.20833	-	-
	rs2032599	-	-	-	1
	rs2032602	-	-	-	1
	rs2032624	0.33333	0.66667	-	-
	rs2032631	0.66667	-	0.33333	-

		Allele			
		A	C	G	T
S N P L o c u s	rs2032636	-	-	1	-
	rs2032652	-	0.083333	-	0.91667
	rs2032658	0.33333	-	0.66667	-
	rs2032673	-	-	-	1
	rs2032674	-	-	-	1
	rs2033003	0.33333	0.66667	-	-
	rs2534636	-	0.95833	-	0.041667
	rs34486382	-	-	1	-
	rs35284970	-	1	-	-
	rs3848982	-	0.91667	-	0.083333
	rs3900	-	0.33333	0.66667	-
	rs3911	1	-	-	-
	rs4141886	0.91667	-	0.083333	-
	rs8179021	-	1	-	-
	rs9341278	-	-	1	-
	rs9786139	0.083333	-	0.91667	-
	rs9786184	0.625	0.375	-	-

Chapter 4, Supplemental Table 47. Y-chromosome SNP allele frequencies – Caucasians. Frequencies are listed for each allele. (-) indicates an allele that was not detected at a particular locus, and therefore has no frequency associated with it.

		Allele			
		A	C	G	T
S N P L o c u s	L298	1	-	-	-
	M479	-	1	-	-
	MSY2.2	-	-	1	-
	P202	1	-	-	-
	rs13447352	0.83333	0.16667	-	-
	rs13447443	1	-	-	-
	rs16980360	0.055556	-	0.94444	-
	rs16981290	-	1	-	-
	rs17250535	-	-	-	1
	rs17269816	-	1	-	0
	rs17306671	0.22222	-	-	0.77778
	rs20320	-	-	1	-
	rs2032595	-	-	-	1
	rs2032597	0.94444	0.055556	-	-
	rs2032599	-	-	-	1
	rs2032602	-	-	-	1
	rs2032624	0.55556	0.44444	-	-
	rs2032631	0.66667	-	0.33333	-

		Allele			
		A	C	G	T
S N P L o c u s	rs2032636	-	-	0.94444	0.055556
	rs2032652	-	0.055556	-	0.94444
	rs2032658	0.55556	-	0.44444	-
	rs2032673	-	-	-	1
	rs2032674	-	-	-	1
	rs2033003	0.33333	0.66667	-	-
	rs2534636	-	1	-	-
	rs34486382	-	-	1	-
	rs35284970	-	1	-	-
	rs3848982	-	0.94444	-	0.055556
	rs3900	-	0.33333	0.66667	-
	rs3911	1	-	-	-
	rs4141886	0.94444	-	0.055556	-
	rs8179021	-	0.77778	-	0.22222
	rs9341278	-	-	1	-
	rs9786139	0.11111	-	0.88889	-
	rs9786184	0.44444	0.55556	-	-

Chapter 4, Supplemental Table 48. Y-chromosome SNP allele frequencies – Hispanics. Frequencies are listed for each allele. (-) indicates an allele that was not detected at a particular locus, and therefore has no frequency associated with it.

		STR Alleles																																			
		19.2	19.3	20	20.2	21	21.2	22	22.2	23	23.2	24	24.2	25	25.2	26	26.2	27	27.2	28	28.2	28.3	29	29.2	30	30.2	31	31.1	31.2	32	32.2	33	33.2	34.2	35	All	
S T R L O C I	CSF1PO	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0.005806	
	D10S1248	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0.014967	
	D1Z391	-	-0.00396	-0.00749	-	0.011389	-	0.002086	-	0.021634	-	0.012704	-	-0.0041	-	-0.00687	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0.013777
	D1S317	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0.03801	
	D1S1434	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0.022868	
	D16S539	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0.01131	
	D17S1301	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0.019304	
	D18S51	-	-	-	-0.01195	-	0.01398	-	-0.00214	-	-	-	-	-	-0.00044	-	-	-	0.001493	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0.009828
	D19S433	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0.002046	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0.020801
	D1S1627	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0.038714	
	D1S1656	-	-0.00868	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0.02009
	D21S11	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0.022214	-	0.056357	0.001493	-	0.023795	-	-0.00832	-0.00551	0.003812	-0.00044	-0.00323	-0.00214	0.017239	-0.00396	-0.00195	0.001773	0.018839	0.014273
	D2S1045	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0.035192
	D2S1338	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0.022735
	D2S1776	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0.009452
	D2S441	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0.023551
	D3S1358	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0.013884
	D4S2408	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0.058301
	D5S2500	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0.01082
	D5S818	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0.021663
D6S1017	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0.051961	
D6S474	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0.014006	
D7S820	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0.013681	
D8S1179	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0.02185	
FGA	-	-	0.013407	-	-0.00276	0.001493	0.004024	-	-0.00431	0.001493	-0.00477	0.001493	-0.00169	-	0.002298	-	-0.0041	-	0.020222	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0.000181
PENTAD	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0.019268	
PENTAE	-	-	0.02132	-	0.032835	-	0.004357	-	0.001493	-	-0.00044	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0.018571
SE33	-0.00044	-	0.013872	-0.00396	-0.00821	-0.00291	-0.0036	-0.0069	-	0.059687	-	0.000062	0.001493	0.019867	-0.00044	-0.00097	0.001493	0.000257	-0.00044	-0.00305	-	-0.00396	0.032143	-0.00044	0.012693	-0.00044	-	-0.00291	-	-0.00396	-	-	-	-	-	0.004658	
TH01	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0.051128
TPOX	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0.024303	
VWA	-	-	-0.00346	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0.019144	

Chapter 4, Supplemental Table 49. F_{ST} – Autosomal STRs. F_{ST} values are listed for each allele, as well as for all alleles combined. (-) indicates an allele that was not detected at a particular locus, and therefore has no F_{ST} value associated with it.

		SNP Alleles				
		A	C	G	T	All
S N P L O C i	L298	***	-	-	-	***
	M479	-	***	-	-	***
	MSY2.2	-	-	***	-	***
	P202	***	-	-	-	***
	rs13447352	0.07312	0.07312	-	-	0.07312
	rs13447443	0.207911	-	0.207911	-	0.207911
	rs16980360	0.388878	-	0.388878	-	0.388878
	rs16981290	0.704277	0.704277	-	-	0.704277
	rs17250535	-0.010789	-	-	-0.010789	-0.010789
	rs17269816	-	0.397521	-	0.397521	0.397521
	rs17306671	0.085537	-	-	0.085537	0.085537
	rs20320	-	-	***	-	***
	rs2032595	-	-0.007409	-	-0.007409	-0.007409
	rs2032597	0.064804	0.064804	-	-	0.064804
	rs2032599	-	-0.007409	-	-0.007409	-0.007409
	rs2032602	-	0.01396	-	0.01396	0.01396
	rs2032624	0.261993	0.261993	-	-	0.261993
	rs2032631	0.323337	-	0.323337	-	0.323337

		SNP Alleles				
		A	C	G	T	All
S N P L O C i	rs2032636	-	-	0.009428	0.009428	0.009428
	rs2032652	-	0.348031	-	0.348031	0.348031
	rs2032658	0.261993	-	0.261993	-	0.261993
	rs2032673	-	-	-	***	***
	rs2032674	-	0.764285	-	0.764285	0.764285
	rs2033003	0.127642	0.127642	-	-	0.127642
	rs2534636	-	-0.009947	-	-0.009947	-0.009947
	rs34486382	-	-	***	-	***
	rs35284970	-	0.143751	-	0.143751	0.143751
	rs3848982	-	0.384552	-	0.384552	0.384552
	rs3900	-	0.127642	0.127642	-	0.127642
	rs3911	***	-	-	-	***
	rs4141886	0.412457	-	0.412457	-	0.412457
	rs8179021	-	0.190791	-	0.190791	0.190791
	rs9341278	0.01396	-	0.01396	-	0.01396
	rs9786139	0.29223	-	0.29223	-	0.29223
	rs9786184	0.230925	0.230925	-	-	0.230925

Chapter 4, Supplemental Table 54. F_{ST} – Y-chromosome SNPs. F_{ST} values are listed for each allele, as well as for all alleles combined. (-) indicates an allele that was not detected at a particular locus, and therefore has no F_{ST} value associated with it. (***) indicates a locus at which only one allele was detected, and therefore has no F_{ST} value associated with it.

Locus Pair	Distance	African American	Asian	Caucasian	Hispanic
CSF1PO/D5S2500	90,758,542 bp	0.883125 (0.831562)	0.167813 (0.153750)	0.220938 (0.103125)	0.784062 (0.485938)
CSF1PO/D5S818	26,344,537 bp	0.820937 (0.589375)	0.079687 (0.402500)	0.209375 (0.052812)	0.083437 (0.000937*)
D12S391/VWA	6,356,826 bp	0.669375 (0.501250)	0.841875 (0.897813)	0.948125 (1.000000)	0.493750 (0.245938)
D21S11/PENTAD	24,501,737 bp	0.725313 (1.000000)	0.361563 (0.940937)	0.725938 (1.000000)	0.856563 (0.685000)
D2S1338/D2S1776	49,234,304 bp	0.970313 (1.000000)	0.413125 (0.955937)	0.148438 (1.000000)	0.038750* (0.278438)
D2S441/TPOX	66,745,605 bp	0.453125 (0.453750)	0.521563 (0.265625)	0.243438 (0.233437)	0.869062 (0.703438)
D5S2500/D5S818	64,414,005 bp	0.649687 (0.564063)	0.834688 (0.980938)	0.190000 (0.047188*)	0.975938 (0.901250)
D6S474/SE33	23,892,286 bp	0.182500 (1.000000)	0.228125 (0.514062)	0.259375 (0.229375)	0.067500 (1.000000)

Chapter 4, Supplemental Table 55. Linkage disequilibrium – syntenic autosomal STR loci pairs. Linkage disequilibrium values are listed for pairs of loci that exist on the same arm of the same chromosome. The distance between the loci, in base pairs, is given. Linkage disequilibrium values in parentheses were calculated with the genotypes preserved, to eliminate the effects of Hardy-Weinberg equilibrium deviation. (*) represents a linkage disequilibrium value that is not significant after Bonferroni correction.

Locus Pairs	Distance	AFA	ASA	CAU	HIS
DYS19/DYS393	6,390,750 bp	0.182812	0.384375	0.185938	0.1375
DYS19/DYS449	1,303,893 bp	0.25875	0.89	0.01875*	0.953438
DYS19/DYS456	5,250,936 bp	0.947812	0.069375	0.521875	0.187188
DYS19/DYS458	1,654,038 bp	0.13125	0.220938	0.20625	0.789062
DYS19/DYS481	1,095,531 bp	0.003125*	0.619375	0.035937*	0.011562*
DYS19/DYS505	5,881,128 bp	0.1125	0.849063	0.570312	0.16
DYS19/DYS522	2,106,505 bp	0.39125	0.37375	0.206563	0.000313*
DYS19/DYS570	2,660,763 bp	0.959063	0.474062	0.151875	0.839688
DYS19/DYS576	2,468,576 bp	0.490938	0.622812	0.64375	0.214688
DYS393/DYS449	5,086,857 bp	0.259688	0.863437	0.215312	0.766563
DYS393/DYS456	1,139,814 bp	0.62375	0.770312	0.917813	0.038125*
DYS393/DYS458	4,736,712 bp	0.615313	0.030312*	0.809063	0.760312
DYS393/DYS481	5,295,219 bp	0.256875	0.884062	0.444375	0.320312
DYS393/DYS505	509,622 bp	0.732187	0.145	0.630625	0.202187
DYS393/DYS522	4,284,245 bp	0.653438	0.515625	0.119063	0.203437
DYS393/DYS570	3,729,987 bp	0.349687	0.000625*	0.01625*	0.307188
DYS393/DYS576	3,922,174 bp	0.565937	0.359063	0.246562	0.095
DYS449/DYS456	3,947,043 bp	0.328437	0.296563	0.903438	0.247188
DYS449/DYS458	350,145 bp	0.685937	1	0.155312	0.050625
DYS449/DYS481	208,362 bp	0.050313	0.391875	0.281875	0.475
DYS449/DYS505	4,577,235 bp	0.803438	0.78625	0.829063	0.814375
DYS449/DYS522	802,612 bp	0.540312	0.09875	0.115312	0.779062
DYS449/DYS570	1,356,870 bp	0.702812	0.289375	0.575313	0.88125
DYS449/DYS576	1,164,683 bp	0.75	0.8925	0.824375	0.73125
DYS456/DYS458	3,596,898 bp	0.377812	0.272188	0.034688*	0.379688
DYS456/DYS481	4,155,405 bp	0.938125	0.013125*	0.234375	0.331875
DYS456/DYS505	630,192 bp	0.699688	0.803438	0.272813	0.095
DYS456/DYS522	3,144,431 bp	0.910937	0.030625*	0.94	0.092813
DYS456/DYS570	2,590,173 bp	0.072813	0.641875	0.79	0.930625
DYS456/DYS576	2,782,360 bp	0.492812	0.03*	0.442812	0.259062
DYS458/DYS481	558,507 bp	0.051875	0.580937	0.179375	0.734688
DYS458/DYS505	4,227,090 bp	0.141563	0.735313	0.024063*	0.438125
DYS458/DYS522	452,467 bp	0.241875	0.992188	0.295312	0.13125
DYS458/DYS570	1,006,725 bp	0.530625	0.394062	0.034688*	0.2525
DYS458/DYS576	814,538 bp	0.875313	0.125	0.304375	0.00125*
DYS481/DYS505	4,785,597 bp	0.072813	0.8675	0.452188	0.624687
DYS481/DYS522	1,010,974 bp	0.00625*	0.20875	0.04625*	0.022187*
DYS481/DYS570	1,565,232 bp	0.535312	0.971875	0.29875	0.097187
DYS481/DYS576	1,373,045 bp	0.994062	0.023125*	0.53875	0.439375
DYS505/DYS522	3,774,623 bp	0.198437	0.108438	0.166563	0.032188*
DYS505/DYS570	3,220,365 bp	0.79	0.611563	0.227187	0.31375
DYS505/DYS576	3,412,552 bp	0.535	0.425	0.15375	0.03125*
DYS522/DYS570	554,258 bp	0.97625	0.280938	0.271562	0.444688
DYS522/DYS576	362,071 bp	0.899375	0.643125	0.511563	0.127812
DYS570/DYS576	192,187 bp	0.28375	0.372188	0.077188	0.067812
DYS385a/DYS385b	40,874 bp	0.001875*	0.775	0.472187	0.7275
DYS385a/DYS389I	6,230,380 bp	0.473438	0.108438	0.474687	0.685937
DYS385a/DYS389II	6,230,380 bp	0.080625	0.263437	0.122813	0.410313
DYS385a/DYS390	3,567,566 bp	0.000313*	0.00125*	0.06375	0.01125*
DYS385a/DYS391	6,739,684 bp	0.049688*	0.3975	0.083437	0.215
DYS385a/DYS392	1,791,397 bp	0.001875*	0.447813	0.0275*	0**
DYS385a/DYS437	6,375,486 bp	0.010313*	0.040313*	0.024375*	0.149375
DYS385a/DYS438	5,904,665 bp	0**	0.39125	0.000937*	0**
DYS385a/DYS439	6,327,262 bp	0.5175	0.8325	0.143437	0.13875

Locus Pairs	Distance	AFA	ASA	CAU	HIS
DYS385a/DYS448	3,522,514 bp	0.172813	0.657188	0.12	0.044687*
DYS385a/DYS460	208,342 bp	0.429688	0.801562	0.343125	0.098437
DYS385a/DYS518	3,522,611 bp	0.4175	0.8925	0.024375*	0.155938
DYS385a/DYS533	2,449,345 bp	0.7025	0.005*	0.179688	0.005938*
DYS385a/DYS549	677,628 bp	0.218438	0.109687	0.144687	0.46375
DYS385a/DYS612	5,089,901 bp	0.1025	0.429688	0.361875	0.520938
DYS385a/DYS635	6,462,933 bp	0.02*	0.080625	0.035312*	0**
DYS385a/DYS643	3,416,465 bp	0.0125*	0.002188*	0**	0.002188*
DYS385a/GATAH4	2,098,922 bp	0.04*	0.036562*	0.403438	0.333125
DYS385b/DYS389I	6,189,506 bp	0.347187	0.58	0.35	0.293438
DYS385b/DYS389II	6,189,506 bp	0.05	0.44625	0.055	0.168125
DYS385b/DYS390	3,526,692 bp	0.000625*	0.560312	0.440625	0.192188
DYS385b/DYS391	6,698,810 bp	0.075313	0.374375	0.389688	0.812187
DYS385b/DYS392	1,832,271 bp	0.00125*	0.744375	0.00125*	0.146875
DYS385b/DYS437	6,334,612 bp	0.007812*	0.622812	0.345313	0.665312
DYS385b/DYS438	5,863,791 bp	0**	0.527188	0.024688*	0.174687
DYS385b/DYS439	6,286,388 bp	0.083437	0.930312	0.078125	0.830625
DYS385b/DYS448	3,563,388 bp	0.072188	0.333125	0.534687	0.2475
DYS385b/DYS460	249,216 bp	0.360625	0.06875	0.481875	0.53375
DYS385b/DYS518	3,481,737 bp	0.485	0.584375	0.936875	0.0575
DYS385b/DYS533	2,408,471 bp	0.090313	0.873125	0.68875	0.890625
DYS385b/DYS549	718,502 bp	0.7575	0.837187	0.4625	0.695937
DYS385b/DYS612	5,049,027 bp	0.105313	0.635	0.972187	0.254063
DYS385b/DYS635	6,422,059 bp	0.122813	0.460313	0.850938	0.057813
DYS385b/DYS643	3,375,591 bp	0.038437*	0.74375	0.624375	0.281562
DYS385b/GATAH4	2,058,048 bp	0.089063	0.134063	0.246562	0.26625
DYS389I/DYS389II	bp	0.014687*	0.010937*	0.024375*	0.028125*
DYS389I/DYS390	2,662,814 bp	0.201563	0.359063	0.13	0.717187
DYS389I/DYS391	509,304 bp	1	0.230313	0.708125	0.692813
DYS389I/DYS392	8,021,777 bp	0.735625	0.639687	0.241875	0.623125
DYS389I/DYS437	145,106 bp	0.804375	0.247188	0.447188	0.0325*
DYS389I/DYS438	325,715 bp	0.047813*	0.015938*	0.061562	0.741563
DYS389I/DYS439	96,882 bp	0.019062*	0.570625	0.248438	0.943125
DYS389I/DYS448	9,752,894 bp	0.199063	0.756563	0.624375	0.247188
DYS389I/DYS460	6,438,722 bp	1	0.362812	0.650625	0.202187
DYS389I/DYS518	2,707,769 bp	0.1925	1	0.896563	0.907188
DYS389I/DYS533	3,781,035 bp	0.719375	0.172813	0.448437	0.481875
DYS389I/DYS549	6,908,008 bp	0.209687	0.136563	0.24875	0.661875
DYS389I/DYS612	1,140,479 bp	0.2525	0.89375	0.718125	0.0525
DYS389I/DYS635	232,553 bp	0.128125	0.039062*	0.254375	0.294375
DYS389I/DYS643	2,813,915 bp	0.578438	0.277813	0.103438	0.425938
DYS389I/GATAH4	4,131,458 bp	0.782188	0.12625	0.552188	0.519687
DYS389II/DYS390	2,662,814 bp	0.040625*	0.157812	0.936562	0.130938
DYS389II/DYS391	509,304 bp	0.000313*	1	0.86125	0.044062*
DYS389II/DYS392	8,021,777 bp	0.2075	0.154688	0.748437	0.06375
DYS389II/DYS437	145,106 bp	0.240625	0.664687	0.494375	0.0025*
DYS389II/DYS438	325,715 bp	0.005313*	0.322188	0.021562*	0.170937
DYS389II/DYS439	96,882 bp	0.08125	0.439375	0.09375	0.73875
DYS389II/DYS448	9,752,894 bp	0.083437	0.114375	1	0.03*
DYS389II/DYS460	6,438,722 bp	0.404687	0.875625	0.81375	0.500938
DYS389II/DYS518	2,707,769 bp	0.960313	0.43125	0.55625	0.615625
DYS389II/DYS533	3,781,035 bp	0.3975	0.57375	0.471562	0.708438
DYS389II/DYS549	6,908,008 bp	0.239063	0.580313	0.437188	0.5225
DYS389II/DYS612	1,140,479 bp	0.56625	0.925	0.19875	0.625

Chapter 4, Supplemental Table 59. Linkage disequilibrium – Y-chromosome STR loci pairs (part 1). Linkage disequilibrium values are listed for pairs of loci that exist on the same arm of the Y-chromosome. The distance between the loci, in base pairs, is given. (*) represents a linkage disequilibrium value that is not significant after Bonferroni correction, while (**) represents a linkage disequilibrium value that is significant after Bonferroni correction.

Locus Pairs	Distance	AFA	ASA	CAU	HIS
DYS389II/DYS635	232,553 bp	0.170625	0.391562	0.225	0.61875
DYS389II/DYS643	2,813,915 bp	0.168125	0.127812	0.086875	0.295625
DYS389II/GATAH4	4,131,458 bp	0.29125	0.39125	0.424375	0.1175
DYS390/DYS391	3,172,118 bp	0**	0.12125	0**	0.009687*
DYS390/DYS392	5,358,963 bp	0.000937*	0.499688	0.000313*	0.005625*
DYS390/DYS437	2,807,920 bp	0**	0.2375	0.17	0.135
DYS390/DYS438	2,337,099 bp	0**	0.892188	0.185625	0**
DYS390/DYS439	2,759,696 bp	0.068437	0.761563	0.160312	0.869687
DYS390/DYS448	7,090,080 bp	0.347187	0.895938	0.149688	0.041563*
DYS390/DYS460	3,775,908 bp	0.124063	0.06	0.57	0.140625
DYS390/DYS518	44,955 bp	0.57875	0.219062	0.68875	0.2925
DYS390/DYS533	1,118,221 bp	0.419687	0.344375	0.005313*	0.180312
DYS390/DYS549	4,245,194 bp	0.11625	0.080312	0.39375	1
DYS390/DYS612	1,522,335 bp	0.00625*	0.8925	0.615	0.268125
DYS390/DYS635	2,895,367 bp	0.004375*	0.59625	0.01875*	0.037187*
DYS390/DYS643	151,101 bp	0.0475*	0.232187	0.0775	0.02*
DYS390/GATAH4	1,468,644 bp	0.025937*	0.17	0.07625	0.268437
DYS391/DYS392	8,531,081 bp	0.090938	1	0.00125*	0.166875
DYS391/DYS437	364,198 bp	0.10125	1	0.097187	0.001563*
DYS391/DYS438	835,019 bp	0.050937	0.519062	0.051562	0.01375*
DYS391/DYS439	412,422 bp	1	0.35375	0.069062	0.07375
DYS391/DYS448	10,262,198 bp	0.720625	0.615625	0.010625*	0.014063*
DYS391/DYS460	6,948,026 bp	0.357187	0.7925	0.262813	0.2575
DYS391/DYS518	3,217,073 bp	0.960625	0.889062	0.534062	0.185312
DYS391/DYS533	4,290,339 bp	0.351875	0.73125	0.010937*	0.244688
DYS391/DYS549	7,417,312 bp	0.404062	0.600938	0.315312	0.03625*
DYS391/DYS612	1,649,783 bp	0.45125	0.257812	0.146563	0.364063
DYS391/DYS635	276,751 bp	0.224062	0.115312	0**	0.26875
DYS391/DYS643	3,323,219 bp	0.237813	0.84875	0.0225*	0.277813
DYS391/GATAH4	4,640,762 bp	1	0.341875	0.57875	0.249375
DYS392/DYS437	8,166,883 bp	0.000313*	0.794687	0.050313	0.108125
DYS392/DYS438	7,696,062 bp	0**	0.100625	0**	0**
DYS392/DYS439	8,118,659 bp	0.780937	0.553438	0.422187	0.027813*
DYS392/DYS448	1,731,117 bp	0.041875*	0.595938	0.000937*	0.46375
DYS392/DYS460	1,583,055 bp	0.366875	0.909375	0.114375	0.436875
DYS392/DYS518	5,314,008 bp	0.257812	0.38875	0.342813	0.34375
DYS392/DYS533	4,240,742 bp	0.176563	0.74875	0.115937	0.370312
DYS392/DYS549	1,113,769 bp	0.011875*	0.376563	0.024688*	0.625625
DYS392/DYS612	6,881,298 bp	0.513125	0.914375	0.355938	0.117188
DYS392/DYS635	8,254,330 bp	0.001563*	0.440937	0.038437*	0.012188*
DYS392/DYS643	5,207,862 bp	0.000937*	0.00625*	0.019062*	0.080937
DYS392/GATAH4	3,890,319 bp	0.173125	0.600313	0.00125*	0.170313
DYS437/DYS438	470,821 bp	0**	0.049063*	0.03*	0.025*
DYS437/DYS439	48,224 bp	0.481563	1	0.504375	0.761563
DYS437/DYS448	9,898,000 bp	0.003438*	0.252188	0.004063*	0.017188*
DYS437/DYS460	6,583,828 bp	0.465313	0.053437	0.897188	0.561875
DYS437/DYS518	2,852,875 bp	0.69125	0.024063*	0.04375*	0.32
DYS437/DYS533	3,926,141 bp	0.18875	0.476875	0.059375	0.0075*
DYS437/DYS549	7,053,114 bp	0.008125*	0.580313	0.065312	0.220938
DYS437/DYS612	1,285,585 bp	0.240313	0.57	0.1275	0.156562
DYS437/DYS635	87,447 bp	0.181562	0.5225	0.051562	0.0425*
DYS437/DYS643	2,959,021 bp	0.00125*	0.07125	0.023125*	0.089063
DYS437/GATAH4	4,276,564 bp	0.134375	1	0.186562	0.406562
DYS438/DYS439	422,597 bp	0.2175	0.683125	0.057813	0.23

Locus Pairs	Distance	AFA	ASA	CAU	HIS
DYS438/DYS448	9,427,179 bp	0.043125*	0.404687	0.038125*	0.029375*
DYS438/DYS460	6,113,007 bp	0.08125	0.279687	0.15875	0.43125
DYS438/DYS518	2,382,054 bp	0.454375	0.397188	0.431875	0.173125
DYS438/DYS533	3,455,320 bp	0.19	0.802813	0.175625	0.114062
DYS438/DYS549	6,582,293 bp	0.059062	1	0.142813	0.2875
DYS438/DYS612	814,764 bp	0.12125	0.969688	0.119063	0.04375*
DYS438/DYS635	558,268 bp	0.0025*	0.6675	0.140313	0.000313*
DYS438/DYS643	2,488,200 bp	0**	0.196875	0.017188*	0.005938*
DYS438/GATAH4	3,805,743 bp	0.181875	0.079375	0.029375*	0.476562
DYS439/DYS448	9,849,776 bp	0.935937	0.8175	0.780625	0.721562
DYS439/DYS460	6,535,604 bp	0.599688	0.770312	0.766563	0.247812
DYS439/DYS518	2,804,651 bp	0.158438	0.85625	0.079062	0.52625
DYS439/DYS533	3,877,917 bp	0.511875	0.360625	0.92625	0.371875
DYS439/DYS549	7,004,890 bp	0.940312	1	0.454688	0.02375*
DYS439/DYS612	1,237,361 bp	0.031875*	0.8525	0.633125	0.219375
DYS439/DYS635	135,671 bp	0.225	0.212188	0.309688	0.21375
DYS439/DYS643	2,910,797 bp	0.160312	0.855	0.685937	0.841562
DYS439/GATAH4	4,228,340 bp	0.209375	0.766875	0.332188	0.332813
DYS448/DYS460	3,314,172 bp	0.727187	1	0.037812*	0.074375
DYS448/DYS518	7,045,125 bp	0.830625	0.689688	0.212812	0.142187
DYS448/DYS533	5,971,859 bp	0.124063	0.188125	0.835625	0.39625
DYS448/DYS549	2,844,886 bp	0.892813	0.717187	0.007188*	0.92375
DYS448/DYS612	8,612,415 bp	0.13625	0.150625	0.180938	0.770625
DYS448/DYS635	9,985,447 bp	0.913125	0.980625	0.578438	0.023438*
DYS448/DYS643	6,938,979 bp	0.065625	0.125938	0.117188	0.0175*
DYS448/GATAH4	5,621,436 bp	0.650625	0.162812	0.175937	0.274687
DYS460/DYS518	3,730,953 bp	0.755625	0.040313*	0.649062	0.795625
DYS460/DYS533	2,657,687 bp	0.14	0.49	0.81375	0.379063
DYS460/DYS549	469,286 bp	0.5375	0.105938	0.014687*	1
DYS460/DYS612	5,298,243 bp	0.851875	0.255	0.16125	0.289375
DYS460/DYS635	6,671,275 bp	0.262813	0.614062	0.889375	0.201875
DYS460/DYS643	3,624,807 bp	0.004375*	0.91875	0.41	0.594375
DYS460/GATAH4	2,307,264 bp	0.514687	0.744375	0.30375	0.190625
DYS518/DYS533	1,073,266 bp	0.122188	0.122813	0.980313	0.87375
DYS518/DYS549	4,200,239 bp	0.985938	0.495	0.072813	0.379063
DYS518/DYS612	1,567,290 bp	0.97125	0.972812	0.069687	0.185625
DYS518/DYS635	2,940,322 bp	0.699375	0.785937	0.11625	0.063125
DYS518/DYS643	106,146 bp	0.657813	0.975625	0.190625	0.238125
DYS518/GATAH4	1,423,689 bp	0.128125	0.981875	0.420625	0.459062
DYS533/DYS549	3,126,973 bp	0.85	0.145937	0.962187	0.35375
DYS533/DYS612	2,640,556 bp	0.70125	0.351875	0.9775	0.740938
DYS533/DYS635	4,013,588 bp	0.736563	0.633125	0.115937	0.045312*
DYS533/DYS643	967,120 bp	0.035*	0.0975	0.324688	0.164375
DYS533/GATAH4	350,423 bp	0.850313	0.492188	0.503125	0.968437
DYS549/DYS612	5,767,529 bp	0.192812	0.90875	0.285	0.293438
DYS549/DYS635	7,140,561 bp	0.003438*	0.206563	0.166563	0.375312
DYS549/DYS643	4,094,093 bp	0.083437	0.32875	0.095937	0.696875
DYS549/GATAH4	2,776,550 bp	0.167813	0.210938	0.005938*	0.535625
DYS612/DYS635	1,373,032 bp	0.25625	0.981563	0.263437	0.129375
DYS612/DYS643	1,673,436 bp	0.085625	1	0.214062	0.350938
DYS612/GATAH4	2,990,979 bp	0.068125	0.209063	0.827812	0.667188
DYS635/DYS643	3,046,468 bp	0.167813	0.193125	0.00875*	0.000625*
DYS635/GATAH4	4,364,011 bp	0.220938	0.474375	0.323437	0.267813
DYS643/GATAH4	1,317,543 bp	0.175937	0.257188	0.454062	0.717812

Chapter 4, Supplemental Table 60. Linkage disequilibrium – Y-chromosome STR loci pairs (part 2). Linkage disequilibrium values are listed for pairs of loci that exist on the same arm of the Y-chromosome. The distance between the loci, in base pairs, is given. (*) represents a linkage disequilibrium value that is not significant after Bonferroni correction, while (**) represents a linkage disequilibrium value that is significant after Bonferroni correction.

Locus Pairs	Distance	AFA	ASA	CAU	HIS
L298/rs16981290	898,721 bp	1	1	1	1
L298/rs2534636	5,810,113 bp	1	1	1	1
L298/rs35284970	5,732,435 bp	1	1	1	1
L298/rs9786139	1,713,770 bp	1	1	1	1
L298/rs9786184	5,579,465 bp	1	1	1	1
M479/MSY2.2	5,819,167 bp	1	1	1	1
M479/P202	6,833,643 bp	1	1	1	1
M479/rs13447352	1,915,187 bp	1	1	1	1
M479/rs13447443	1,904,635 bp	1	1	1	1
M479/rs16980360	5,025,340 bp	1	1	1	1
M479/rs17250535	2,638,535 bp	1	1	1	1
M479/rs17269816	3,780,895 bp	1	1	1	1
M479/rs17306671	6,803,332 bp	1	1	1	1
M479/rs20320	5,936,503 bp	1	1	1	1
M479/rs2032595	6,020,675 bp	1	1	1	1
M479/rs2032597	5,986,874 bp	1	1	1	1
M479/rs2032599	5,983,112 bp	1	1	1	1
M479/rs2032602	5,880,386 bp	1	1	1	1
M479/rs2032624	5,808,242 bp	1	1	1	1
M479/rs2032631	1,033,121 bp	1	1	1	1
M479/rs2032636	5,807,137 bp	1	1	1	1
M479/rs2032652	1,082,647 bp	1	1	1	1
M479/rs2032658	5,252,683 bp	1	1	1	1
M479/rs2032673	1,059,392 bp	1	1	1	1
M479/rs2032674	5,362,741 bp	1	1	1	1
M479/rs2033003	2,716,258 bp	1	1	1	1
M479/rs34486382	5,353,231 bp	1	1	1	1
M479/rs3848982	882,542 bp	1	1	1	1
M479/rs3900	895,591 bp	1	1	1	1
M479/rs3911	898,788 bp	1	1	1	1
M479/rs4141886	6,636,799 bp	1	1	1	1
M479/rs8179021	5,816,084 bp	1	1	1	1
M479/rs9341278	5,364,942 bp	1	1	1	1
MSY2.2/P202	1,014,476 bp	1	1	1	1
MSY2.2/rs13447352	7,734,354 bp	1	1	1	1
MSY2.2/rs13447443	7,723,802 bp	1	1	1	1
MSY2.2/rs16980360	793,827 bp	1	1	1	1
MSY2.2/rs17250535	8,457,702 bp	1	1	1	1
MSY2.2/rs17269816	2,038,272 bp	1	1	1	1
MSY2.2/rs17306671	984,165 bp	1	1	1	1
MSY2.2/rs20320	117,336 bp	1	1	1	1
MSY2.2/rs2032595	201,508 bp	1	1	1	1
MSY2.2/rs2032597	167,707 bp	1	1	1	1
MSY2.2/rs2032599	163,945 bp	1	1	1	1
MSY2.2/rs2032602	61,219 bp	1	1	1	1
MSY2.2/rs2032624	10,925 bp	1	1	1	1
MSY2.2/rs2032631	6,852,288 bp	1	1	1	1
MSY2.2/rs2032636	12,030 bp	1	1	1	1
MSY2.2/rs2032652	6,901,814 bp	1	1	1	1
MSY2.2/rs2032658	566,484 bp	1	1	1	1
MSY2.2/rs2032673	6,878,559 bp	1	1	1	1
MSY2.2/rs2032674	456,426 bp	1	1	1	1
MSY2.2/rs2033003	8,535,425 bp	1	1	1	1

Locus Pairs	Distance	AFA	ASA	CAU	HIS
MSY2.2/rs34486382	465,936 bp	1	1	1	1
MSY2.2/rs3848982	6,701,709 bp	0**	1	1	1
MSY2.2/rs3900	6,714,758 bp	1	1	1	1
MSY2.2/rs3911	6,717,955 bp	1	1	1	1
MSY2.2/rs4141886	817,632 bp	1	1	1	1
MSY2.2/rs8179021	3,083 bp	1	1	1	1
MSY2.2/rs9341278	454,225 bp	1	1	1	1
P202/rs13447352	8,748,830 bp	1	1	1	1
P202/rs13447443	8,738,278 bp	1	1	1	1
P202/rs16980360	1,808,303 bp	1	1	1	1
P202/rs17250535	9,472,178 bp	1	1	1	1
P202/rs17269816	3,052,748 bp	1	1	1	1
P202/rs17306671	30,311 bp	1	1	1	1
P202/rs20320	897,140 bp	1	1	1	1
P202/rs2032595	812,968 bp	1	1	1	1
P202/rs2032597	846,769 bp	1	1	1	1
P202/rs2032599	850,531 bp	1	1	1	1
P202/rs2032602	953,257 bp	1	1	1	1
P202/rs2032624	1,025,401 bp	1	1	1	1
P202/rs2032631	7,866,764 bp	1	1	1	1
P202/rs2032636	1,026,506 bp	1	1	1	1
P202/rs2032652	7,916,290 bp	1	1	1	1
P202/rs2032658	1,580,960 bp	1	1	1	1
P202/rs2032673	7,893,035 bp	1	1	1	1
P202/rs2032674	1,470,902 bp	1	1	1	1
P202/rs2033003	9,549,901 bp	1	1	1	1
P202/rs34486382	1,480,412 bp	1	1	1	1
P202/rs3848982	7,716,185 bp	1	1	1	1
P202/rs3900	7,729,234 bp	1	1	1	1
P202/rs3911	7,732,431 bp	1	1	1	1
P202/rs4141886	196,844 bp	1	1	1	1
P202/rs8179021	1,017,559 bp	1	1	1	1
P202/rs9341278	1,468,701 bp	1	1	1	1
rs13447352/rs13447443	10,552 bp	1	1	1	1
rs13447352/rs16980360	6,940,527 bp	1	1	1	1
rs13447352/rs17250535	723,348 bp	1	1	1	1
rs13447352/rs17269816	5,696,082 bp	1	1	1	1
rs13447352/rs17306671	8,718,519 bp	1	1	0.253125	0.005313*
rs13447352/rs20320	7,851,690 bp	1	1	1	1
rs13447352/rs2032595	7,935,862 bp	1	1	1	1
rs13447352/rs2032597	7,902,061 bp	1	1	1	1
rs13447352/rs2032599	7,898,299 bp	1	1	1	1
rs13447352/rs2032602	7,795,573 bp	1	1	1	1
rs13447352/rs2032624	7,723,429 bp	1	1	0.333125	0.199063
rs13447352/rs2032631	882,066 bp	1	1	0.3425	0.02625*
rs13447352/rs2032636	7,722,324 bp	1	1	1	1
rs13447352/rs2032652	832,540 bp	1	1	1	1
rs13447352/rs2032658	7,167,870 bp	1	1	0.32875	0.209687
rs13447352/rs2032673	855,795 bp	1	1	1	1
rs13447352/rs2032674	7,277,928 bp	1	1	1	1
rs13447352/rs2033003	801,071 bp	1	1	0.326562	0.020313*
rs13447352/rs34486382	7,268,418 bp	1	1	1	1
rs13447352/rs3848982	1,032,645 bp	1	1	1	1

Chapter 4, Supplemental Table 69. Linkage disequilibrium – Y-chromosome SNP loci pairs (part 1). Linkage disequilibrium values are listed for pairs of loci that exist on the same arm of the Y-chromosome. The distance between the loci, in base pairs, is given. (*) represents a linkage disequilibrium value that is not significant after Bonferroni correction, while (**) represents a linkage disequilibrium value that is significant after Bonferroni correction.

Locus Pairs	Distance	AFA	ASA	CAU	HIS
rs13447352/rs3900	1,019,596 bp	1	1	0.347187	0.020938*
rs13447352/rs3911	1,016,399 bp	1	1	1	1
rs13447352/rs4141886	8,551,986 bp	1	1	1	1
rs13447352/rs8179021	7,731,271 bp	1	1	1	1
rs13447352/rs9341278	7,280,129 bp	1	1	1	1
rs13447443/rs16980360	6,929,975 bp	1	1	1	1
rs13447443/rs17250535	733,900 bp	1	1	1	1
rs13447443/rs17269816	5,685,530 bp	1	0.109063	1	1
rs13447443/rs17306671	8,707,967 bp	1	1	1	1
rs13447443/rs20320	7,841,138 bp	1	1	1	1
rs13447443/rs2032595	7,925,310 bp	1	1	1	1
rs13447443/rs2032597	7,891,509 bp	1	1	1	1
rs13447443/rs2032599	7,887,747 bp	1	1	1	1
rs13447443/rs2032602	7,785,021 bp	1	1	1	1
rs13447443/rs2032624	7,712,877 bp	1	1	1	1
rs13447443/rs2032631	871,514 bp	1	1	1	1
rs13447443/rs2032636	7,711,772 bp	1	1	1	1
rs13447443/rs2032652	821,988 bp	1	0.50875	1	1
rs13447443/rs2032658	7,157,318 bp	1	1	1	1
rs13447443/rs2032673	845,243 bp	1	1	1	1
rs13447443/rs2032674	7,267,376 bp	1	0.52125	1	1
rs13447443/rs2033003	811,623 bp	1	0.5225	1	1
rs13447443/rs34486382	7,257,866 bp	1	1	1	1
rs13447443/rs3848982	1,022,093 bp	1	1	1	1
rs13447443/rs3900	1,009,044 bp	1	0.510938	1	1
rs13447443/rs3911	1,005,847 bp	1	1	1	1
rs13447443/rs4141886	8,541,434 bp	1	1	1	1
rs13447443/rs8179021	7,720,719 bp	1	1	1	1
rs13447443/rs9341278	7,269,577 bp	1	1	1	1
rs16980360/rs17250535	7,663,875 bp	1	1	1	1
rs16980360/rs17269816	1,244,445 bp	1	1	1	1
rs16980360/rs17306671	1,777,992 bp	0.4325	1	1	1
rs16980360/rs20320	911,163 bp	1	1	1	1
rs16980360/rs2032595	995,335 bp	0.429375	1	1	1
rs16980360/rs2032597	961,534 bp	0.43	1	1	1
rs16980360/rs2032599	957,772 bp	0.432188	1	1	1
rs16980360/rs2032602	855,046 bp	1	1	1	1
rs16980360/rs2032624	782,902 bp	0.000313*	1	0.1	1
rs16980360/rs2032631	6,058,461 bp	0**	1	0.104375	0.322188
rs16980360/rs2032636	781,797 bp	1	1	1	1
rs16980360/rs2032652	6,107,987 bp	0**	1	0.004375*	0.059687
rs16980360/rs2032658	227,343 bp	0.000937*	1	0.09625	1
rs16980360/rs2032673	6,084,732 bp	1	1	1	1
rs16980360/rs2032674	337,401 bp	1	1	1	1
rs16980360/rs2033003	7,741,598 bp	0**	1	0.105	0.342187
rs16980360/rs34486382	327,891 bp	1	1	1	1
rs16980360/rs3848982	5,907,882 bp	0**	1	0.0025*	0.051875
rs16980360/rs3900	5,920,931 bp	0.000313*	1	0.104063	0.353125
rs16980360/rs3911	5,924,128 bp	1	1	1	1
rs16980360/rs4141886	1,611,459 bp	0**	1	0.00375*	0.059375
rs16980360/rs8179021	790,744 bp	1	1	1	1
rs16980360/rs9341278	339,602 bp	1	1	1	1
rs16981290/rs2534636	4,911,392 bp	1	1	1	1

Locus Pairs	Distance	AFA	ASA	CAU	HIS
rs16981290/rs35284970	4,833,714 bp	1	0.01625*	1	1
rs16981290/rs9786139	815,049 bp	1	0.001875*	1	1
rs16981290/rs9786184	4,680,744 bp	1	1	1	1
rs17250535/rs17269816	6,419,430 bp	1	1	1	1
rs17250535/rs17306671	9,441,867 bp	1	1	1	1
rs17250535/rs20320	8,575,038 bp	1	1	1	1
rs17250535/rs2032595	8,659,210 bp	1	1	1	1
rs17250535/rs2032597	8,625,409 bp	1	1	1	1
rs17250535/rs2032599	8,621,647 bp	1	1	1	1
rs17250535/rs2032602	8,518,921 bp	1	1	1	1
rs17250535/rs2032624	8,446,777 bp	1	1	1	1
rs17250535/rs2032631	1,605,414 bp	1	1	1	1
rs17250535/rs2032636	8,445,672 bp	1	1	1	1
rs17250535/rs2032652	1,555,888 bp	1	1	1	1
rs17250535/rs2032658	7,891,218 bp	1	1	1	1
rs17250535/rs2032673	1,579,143 bp	1	1	1	1
rs17250535/rs2032674	8,001,276 bp	1	1	1	1
rs17250535/rs2033003	77,723 bp	1	1	1	1
rs17250535/rs34486382	7,991,766 bp	1	1	1	1
rs17250535/rs3848982	1,755,993 bp	1	1	1	1
rs17250535/rs3900	1,742,944 bp	1	1	1	1
rs17250535/rs3911	1,739,747 bp	1	1	1	1
rs17250535/rs4141886	9,275,334 bp	1	1	1	1
rs17250535/rs8179021	8,454,619 bp	1	1	1	1
rs17250535/rs9341278	8,003,477 bp	1	1	1	1
rs17269816/rs17306671	3,022,437 bp	1	1	1	1
rs17269816/rs20320	2,155,608 bp	1	1	1	1
rs17269816/rs2032595	2,239,780 bp	1	1	1	1
rs17269816/rs2032597	2,205,979 bp	1	1	1	1
rs17269816/rs2032599	2,202,217 bp	1	1	1	1
rs17269816/rs2032602	2,099,491 bp	1	1	1	1
rs17269816/rs2032624	2,027,347 bp	1	1	1	1
rs17269816/rs2032631	4,814,016 bp	1	1	1	1
rs17269816/rs2032636	2,026,242 bp	1	1	1	1
rs17269816/rs2032652	4,863,542 bp	1	0.105625	1	1
rs17269816/rs2032658	1,471,788 bp	1	1	1	1
rs17269816/rs2032673	4,840,287 bp	1	1	1	1
rs17269816/rs2032674	1,581,846 bp	1	0.099062	1	1
rs17269816/rs2033003	6,497,153 bp	1	0.1075	1	1
rs17269816/rs34486382	1,572,336 bp	1	1	1	1
rs17269816/rs3848982	4,663,437 bp	1	1	1	1
rs17269816/rs3900	4,676,486 bp	1	0.104688	1	1
rs17269816/rs3911	4,679,683 bp	1	1	1	1
rs17269816/rs4141886	2,855,904 bp	1	1	1	1
rs17269816/rs8179021	2,035,189 bp	1	1	1	1
rs17269816/rs9341278	1,584,047 bp	1	1	1	1
rs17306671/rs20320	866,829 bp	1	1	1	1
rs17306671/rs2032595	782,657 bp	1	1	1	1
rs17306671/rs2032597	816,458 bp	0.043437*	1	0.000313*	0.225312
rs17306671/rs2032599	820,220 bp	1	1	1	1
rs17306671/rs2032602	922,946 bp	1	1	1	1
rs17306671/rs2032624	995,090 bp	1	1	0.000313*	0.094687
rs17306671/rs2032631	7,836,453 bp	1	1	0.000313*	0.006875*

Chapter 4, Supplemental Table 70. Linkage disequilibrium – Y-chromosome SNP loci pairs (part 2). Linkage disequilibrium values are listed for pairs of loci that exist on the same arm of the Y-chromosome. The distance between the loci, in base pairs, is given. (*) represents a linkage disequilibrium value that is not significant after Bonferroni correction, while (**) represents a linkage disequilibrium value that is significant after Bonferroni correction.

Locus Pairs	Distance	AFA	ASA	CAU	HIS
rs17306671/rs2032636	996,195 bp	1	1	1	1
rs17306671/rs2032652	7,885,979 bp	0.345	1	1	1
rs17306671/rs2032658	1,550,649 bp	1	1	0**	0.090313
rs17306671/rs2032673	7,862,724 bp	1	1	1	1
rs17306671/rs2032674	1,440,591 bp	1	1	1	1
rs17306671/rs2033003	9,519,590 bp	1	1	0**	0.005938*
rs17306671/rs34486382	1,450,101 bp	1	1	1	1
rs17306671/rs3848982	7,685,874 bp	0.393125	1	1	1
rs17306671/rs3900	7,698,923 bp	1	1	0.000625*	0.007188*
rs17306671/rs3911	7,702,120 bp	1	1	1	1
rs17306671/rs4141886	166,533 bp	0.367812	1	1	1
rs17306671/rs8179021	987,248 bp	1	1	1	0.527813
rs17306671/rs9341278	1,438,390 bp	1	1	1	1
rs20320/rs2032595	84,172 bp	1	1	1	1
rs20320/rs2032597	50,371 bp	1	1	1	1
rs20320/rs2032599	46,609 bp	1	1	1	1
rs20320/rs2032602	56,117 bp	1	1	1	1
rs20320/rs2032624	128,261 bp	1	1	1	1
rs20320/rs2032631	6,969,624 bp	1	1	1	1
rs20320/rs2032636	129,366 bp	1	1	1	1
rs20320/rs2032652	7,019,150 bp	1	1	1	1
rs20320/rs2032658	683,820 bp	1	1	1	1
rs20320/rs2032673	6,995,895 bp	1	1	1	1
rs20320/rs2032674	573,762 bp	1	1	1	1
rs20320/rs2033003	8,652,761 bp	1	1	1	1
rs20320/rs34486382	583,272 bp	1	1	1	1
rs20320/rs3848982	6,819,045 bp	1	1	1	1
rs20320/rs3900	6,832,094 bp	1	1	1	1
rs20320/rs3911	6,835,291 bp	1	1	1	1
rs20320/rs4141886	700,296 bp	1	1	1	1
rs20320/rs8179021	120,419 bp	1	1	1	1
rs20320/rs9341278	571,561 bp	1	1	1	1
rs2032595/rs2032597	33,801 bp	1	1	1	1
rs2032595/rs2032599	37,563 bp	0.045625*	1	1	1
rs2032595/rs2032602	140,289 bp	1	1	1	1
rs2032595/rs2032624	212,433 bp	1	1	1	1
rs2032595/rs2032631	7,053,796 bp	1	1	1	1
rs2032595/rs2032636	213,538 bp	1	1	1	1
rs2032595/rs2032652	7,103,322 bp	1	1	1	1
rs2032595/rs2032658	767,992 bp	1	1	1	1
rs2032595/rs2032673	7,080,067 bp	1	1	1	1
rs2032595/rs2032674	657,934 bp	1	1	1	1
rs2032595/rs2033003	8,736,933 bp	1	1	1	1
rs2032595/rs34486382	667,444 bp	1	1	1	1
rs2032595/rs3848982	6,903,217 bp	0.381563	1	1	1
rs2032595/rs3900	6,916,266 bp	1	1	1	1
rs2032595/rs3911	6,919,463 bp	1	1	1	1
rs2032595/rs4141886	616,124 bp	1	1	1	1
rs2032595/rs8179021	204,591 bp	1	1	1	1
rs2032595/rs9341278	655,733 bp	1	1	1	1
rs2032597/rs2032599	3,762 bp	1	1	1	1
rs2032597/rs2032602	106,488 bp	1	1	1	1
rs2032597/rs2032624	178,632 bp	1	1	0.001875*	1

Locus Pairs	Distance	AFA	ASA	CAU	HIS
rs2032597/rs2032631	7,019,995 bp	1	1	0.000937*	0.33875
rs2032597/rs2032636	179,737 bp	1	1	1	1
rs2032597/rs2032652	7,069,521 bp	0.350938	1	1	1
rs2032597/rs2032658	734,191 bp	1	1	0.002188*	1
rs2032597/rs2032673	7,046,266 bp	1	1	1	1
rs2032597/rs2032674	624,133 bp	1	1	1	1
rs2032597/rs2033003	8,703,132 bp	1	1	0.000937*	0.339062
rs2032597/rs34486382	633,643 bp	1	1	1	1
rs2032597/rs3848982	6,869,416 bp	0.387813	1	1	1
rs2032597/rs3900	6,882,465 bp	1	1	0.0025*	0.328125
rs2032597/rs3911	6,885,662 bp	1	1	1	1
rs2032597/rs4141886	649,925 bp	0.365625	1	1	1
rs2032597/rs8179021	170,790 bp	1	1	1	1
rs2032597/rs9341278	621,932 bp	1	1	1	1
rs2032599/rs2032602	102,726 bp	1	1	1	1
rs2032599/rs2032624	174,870 bp	1	1	1	1
rs2032599/rs2032631	7,016,233 bp	1	1	1	1
rs2032599/rs2032636	175,975 bp	1	1	1	1
rs2032599/rs2032652	7,065,759 bp	1	1	1	1
rs2032599/rs2032658	730,429 bp	1	1	1	1
rs2032599/rs2032673	7,042,504 bp	1	1	1	1
rs2032599/rs2032674	620,371 bp	1	1	1	1
rs2032599/rs2033003	8,699,370 bp	1	1	1	1
rs2032599/rs34486382	629,881 bp	1	1	1	1
rs2032599/rs3848982	6,865,654 bp	0.39875	1	1	1
rs2032599/rs3900	6,878,703 bp	1	1	1	1
rs2032599/rs3911	6,881,900 bp	1	1	1	1
rs2032599/rs4141886	653,687 bp	1	1	1	1
rs2032599/rs8179021	167,028 bp	1	1	1	1
rs2032599/rs9341278	618,170 bp	1	1	1	1
rs2032602/rs2032624	72,144 bp	1	1	1	1
rs2032602/rs2032631	6,913,507 bp	1	1	1	1
rs2032602/rs2032636	73,249 bp	1	1	1	1
rs2032602/rs2032652	6,963,033 bp	1	0.246562	1	1
rs2032602/rs2032658	627,703 bp	1	1	1	1
rs2032602/rs2032673	6,939,778 bp	1	1	1	1
rs2032602/rs2032674	517,645 bp	1	0.243125	1	1
rs2032602/rs2033003	8,596,644 bp	1	0.222812	1	1
rs2032602/rs34486382	527,155 bp	1	1	1	1
rs2032602/rs3848982	6,762,928 bp	1	0.05125	1	1
rs2032602/rs3900	6,775,977 bp	1	0.221875	1	1
rs2032602/rs3911	6,779,174 bp	1	1	1	1
rs2032602/rs4141886	756,413 bp	1	0.060312	1	1
rs2032602/rs8179021	64,302 bp	1	1	1	1
rs2032602/rs9341278	515,444 bp	1	1	1	1
rs2032624/rs2032631	6,841,363 bp	0**	1	0**	0.011875*
rs2032624/rs2032636	1,105 bp	1	1	1	1
rs2032624/rs2032652	6,890,889 bp	0**	1	0.102188	1
rs2032624/rs2032658	555,559 bp	0**	1	0**	0**
rs2032624/rs2032673	6,867,634 bp	1	1	1	1
rs2032624/rs2032674	445,501 bp	1	1	1	1
rs2032624/rs2033003	8,524,500 bp	0**	1	0**	0.01125*
rs2032624/rs34486382	455,011 bp	1	1	1	1

Chapter 4, Supplemental Table 71. Linkage disequilibrium – Y-chromosome SNP loci pairs (part 3). Linkage disequilibrium values are listed for pairs of loci that exist on the same arm of the Y-chromosome. The distance between the loci, in base pairs, is given. (*) represents a linkage disequilibrium value that is not significant after Bonferroni correction, while (**) represents a linkage disequilibrium value that is significant after Bonferroni correction.

Locus Pairs	Distance	AFA	ASA	CAU	HIS
rs2032624/rs3848982	6,690,784 bp	0**	1	0.104063	1
rs2032624/rs3900	6,703,833 bp	0**	1	0**	0.010937*
rs2032624/rs3911	6,707,030 bp	1	1	1	1
rs2032624/rs4141886	828,557 bp	0**	1	0.100937	1
rs2032624/rs8179021	7,842 bp	1	1	1	0.087813
rs2032624/rs9341278	443,300 bp	1	1	1	1
rs2032631/rs2032636	6,840,258 bp	1	1	1	0.350313
rs2032631/rs2032652	49,526 bp	0**	1	0.10875	0.334375
rs2032631/rs2032658	6,285,804 bp	0**	1	0.000313*	0.01375*
rs2032631/rs2032673	26,271 bp	1	1	1	1
rs2032631/rs2032674	6,395,862 bp	1	1	1	1
rs2032631/rs2033003	1,683,137 bp	0**	1	0**	0**
rs2032631/rs34486382	6,386,352 bp	1	1	1	1
rs2032631/rs3848982	150,579 bp	0**	1	0.10125	0.33375
rs2032631/rs3900	137,530 bp	0**	1	0**	0**
rs2032631/rs3911	134,333 bp	1	1	1	1
rs2032631/rs4141886	7,669,920 bp	0.000313*	1	0.113437	0.33375
rs2032631/rs8179021	6,849,205 bp	1	1	1	0.237813
rs2032631/rs9341278	6,398,063 bp	1	1	1	1
rs2032636/rs2032652	6,889,784 bp	1	1	1	1
rs2032636/rs2032658	554,454 bp	1	1	1	1
rs2032636/rs2032673	6,866,529 bp	1	1	1	1
rs2032636/rs2032674	444,396 bp	1	1	1	1
rs2032636/rs2033003	8,523,395 bp	1	1	1	0.3425
rs2032636/rs34486382	453,906 bp	1	1	1	1
rs2032636/rs3848982	6,689,679 bp	1	1	1	1
rs2032636/rs3900	6,702,728 bp	1	1	1	0.346562
rs2032636/rs3911	6,705,925 bp	1	1	1	1
rs2032636/rs4141886	829,662 bp	1	1	1	1
rs2032636/rs8179021	8,947 bp	1	1	1	1
rs2032636/rs9341278	442,195 bp	1	1	1	1
rs2032652/rs2032658	6,335,330 bp	0**	1	0.097187	1
rs2032652/rs2032673	23,255 bp	1	1	1	1
rs2032652/rs2032674	6,445,388 bp	1	0**	1	1
rs2032652/rs2033003	1,633,611 bp	0**	0.00125*	0.0975	0.331875
rs2032652/rs34486382	6,435,878 bp	1	1	1	1
rs2032652/rs3848982	200,105 bp	0**	0.238125	0.001563*	0.056875
rs2032652/rs3900	187,056 bp	0**	0**	0.106563	0.3275
rs2032652/rs3911	183,859 bp	1	1	1	1
rs2032652/rs4141886	7,719,446 bp	0**	0.22875	0.00375*	0.050625
rs2032652/rs8179021	6,898,731 bp	1	1	1	1
rs2032652/rs9341278	6,447,589 bp	1	1	1	1
rs2032658/rs2032673	6,312,075 bp	1	1	1	1
rs2032658/rs2032674	110,058 bp	1	1	1	1
rs2032658/rs2033003	7,968,941 bp	0**	1	0**	0.016875*
rs2032658/rs34486382	100,548 bp	1	1	1	1
rs2032658/rs3848982	6,135,225 bp	0**	1	0.101562	1
rs2032658/rs3900	6,148,274 bp	0**	1	0**	0.015312*
rs2032658/rs3911	6,151,471 bp	1	1	1	1
rs2032658/rs4141886	1,384,116 bp	0**	1	0.107188	1
rs2032658/rs8179021	563,401 bp	1	1	1	0.103125
rs2032658/rs9341278	112,259 bp	1	1	1	1

Locus Pairs	Distance	AFA	ASA	CAU	HIS
rs2032673/rs2032674	6,422,133 bp	1	1	1	1
rs2032673/rs2033003	1,656,866 bp	1	1	1	1
rs2032673/rs34486382	6,412,623 bp	1	1	1	1
rs2032673/rs3848982	176,850 bp	1	1	1	1
rs2032673/rs3900	163,801 bp	1	1	1	1
rs2032673/rs3911	160,604 bp	1	1	1	1
rs2032673/rs4141886	7,696,191 bp	1	1	1	1
rs2032673/rs8179021	6,875,476 bp	1	1	1	1
rs2032673/rs9341278	6,424,334 bp	1	1	1	1
rs2032674/rs2033003	8,078,999 bp	1	0.000625*	1	1
rs2032674/rs34486382	9,510 bp	1	1	1	1
rs2032674/rs3848982	6,245,283 bp	1	0.225938	1	1
rs2032674/rs3900	6,258,332 bp	1	0.000313*	1	1
rs2032674/rs3911	6,261,529 bp	1	1	1	1
rs2032674/rs4141886	1,274,058 bp	1	0.236563	1	1
rs2032674/rs8179021	453,343 bp	1	1	1	1
rs2032674/rs9341278	2,201 bp	1	1	1	1
rs2033003/rs34486382	8,069,489 bp	1	1	1	1
rs2033003/rs3848982	1,833,716 bp	0**	0.243438	0.103125	0.344687
rs2033003/rs3900	1,820,667 bp	0**	0.000625*	0**	0**
rs2033003/rs3911	1,817,470 bp	1	1	1	1
rs2033003/rs4141886	9,353,057 bp	0**	0.239063	0.106875	0.340938
rs2033003/rs8179021	8,532,342 bp	1	1	1	0.233437
rs2033003/rs9341278	8,081,200 bp	1	1	1	1
rs2534636/rs35284970	77,678 bp	1	1	1	1
rs2534636/rs9786139	4,096,343 bp	1	1	1	1
rs2534636/rs9786184	230,648 bp	1	1	0.367812	1
rs34486382/rs3848982	6,235,773 bp	1	1	1	1
rs34486382/rs3900	6,248,822 bp	1	1	1	1
rs34486382/rs3911	6,252,019 bp	1	1	1	1
rs34486382/rs4141886	1,283,568 bp	1	1	1	1
rs34486382/rs8179021	462,853 bp	1	1	1	1
rs34486382/rs9341278	11,711 bp	1	1	1	1
rs35284970/rs9786139	4,018,665 bp	1	0.007812*	1	1
rs35284970/rs9786184	152,970 bp	1	1	1	1
rs3848982/rs3900	13,049 bp	0**	0.245312	0.092813	0.337187
rs3848982/rs3911	16,246 bp	1	1	1	1
rs3848982/rs4141886	7,519,341 bp	0**	0.05625	0.004063*	0.052187
rs3848982/rs8179021	6,698,626 bp	1	1	1	1
rs3848982/rs9341278	6,247,484 bp	1	1	1	1
rs3900/rs3911	3,197 bp	1	1	1	1
rs3900/rs4141886	7,532,390 bp	0**	0.225625	0.1025	0.337187
rs3900/rs8179021	6,711,675 bp	1	1	1	0.24375
rs3900/rs9341278	6,260,533 bp	1	1	1	1
rs3911/rs4141886	7,535,587 bp	1	1	1	1
rs3911/rs8179021	6,714,872 bp	1	1	1	1
rs3911/rs9341278	6,263,730 bp	1	1	1	1
rs4141886/rs8179021	820,715 bp	1	1	1	1
rs4141886/rs9341278	1,271,857 bp	1	1	1	1
rs8179021/rs9341278	451,142 bp	1	1	1	1
rs9786139/rs9786184	3,865,695 bp	0**	1	0.129688	0.474687

Chapter 4, Supplemental Table 72. Linkage disequilibrium – Y-chromosome SNP loci pairs (part 4). Linkage disequilibrium values are listed for pairs of loci that exist on the same arm of the Y-chromosome. The distance between the loci, in base pairs, is given. (*) represents a linkage disequilibrium value that is not significant after Bonferroni correction, while (**) represents a linkage disequilibrium value that is significant after Bonferroni correction.

REFERENCES

DEVELOPMENT OF A COMPREHENSIVE MASSIVELY PARALLEL SEQUENCING PANEL OF SINGLE NUCLEOTIDE POLYMORPHISM AND SHORT TANDEM REPEAT MARKERS FOR HUMAN IDENTIFICATION

Adessi, C., Matton, G., Ayala, G., Turcatti, G., Mermoud, J.J., Mayer, P., and Kawashima, E. 2000. Solid phase DNA amplification: characterisation of primer attachment and amplification mechanisms. *Nucleic Acids Res.* 28(20):E87.

Adey, A., Morrison, H.G., Asan, Xun, X., Kitzman, J.O., Turner, E.H., Stackhouse, B., MacKenzie, A.P., Caruccio, N.C., Zhang, X., and Shendure, J. 2010. Rapid, low-input, low-bias construction of shotgun fragment libraries by high-density *in vitro* transposition. *Genome Biol.* 11:R119.

Agilent Technologies® HaloPlex™ Specifications: <http://www.genomics.agilent.com/GenericB.aspx?pagetype=Custom&subpagetype=Custom&pageid=3081>.

AGREP: <http://laurikari.net/tre/download/>.

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., Ploski, R., Dobosz, T., Henke, L., Henke, J., Furtado, M.R., and Kayser, M. 2010. Mutability of Y-chromosomal microsatellites: rates, characteristics, molecular bases, and forensic implications. *Am. J. Hum. Genet.* 87:341-353.

Becker D, Bender K, Edelmann J, Götz F, Henke L, Hering S, Hohoff C, Hoppe K, Klintschar M, Muche M, Rolf B, Szibor R, Weirich V, Jung M, and Brabetz W. 2007. New alleles and mutational events at 14 STR loci from different German populations. *Forens. Sci. Int. Genet.* 1(3-4):232-237.

Berglund, E.C., Kiialainen, A., and Syvänen, A.C. 2011. Next-generation sequencing technologies and applications for human genetic history and forensics. *Investig. Genet.* 2:1–15.

Bornman, D.M., Hester, M.E., Schuetter, J.M., Kasoji, M.D., Minard-Smith, A., Barden, C.A., Nelson, S.C., Godbold, G.D., Baker, C.H., Yang, B., Walther, J.E., Tornes, I.E., Yan, P.S., Rodriguez, B., Bundschuh, R., Dickens, M.L., Young, B.A., and Faith, S.A. 2012. Short-read, high-throughput sequencing technology for STR genotyping. *Biotech. Rapid Dispatches* 1–6.

Børsting, C., Sanchez, J.J., Morling, N. 2005. SNP typing on the NanoChip electronic microarray. *Methods Mol. Biol.* 297:155– 168.

Brenner, S. *et al.* 2000. Gene expression analysis by massively parallel signature sequencing (MPSS) on microbead arrays. *Nat. Biotechnol.* 18:630-634.

Brenner, S., Williams, S.R., Vermaas, E.H., Storck, T., Moon, K., McCollum, C., Mao, J.I., Luo, S., Kirchner, J.J., Eletre, S., DuBridge, R.B., Burcham, T., and Albrecht, G. 2000. In vitro cloning of complex mixtures of DNA on microbeads: physical separation of differentially expressed cDNAs. *Proc. Natl. Acad. Sci. USA* 97:1665-1670.

Budowle, B. 2004. SNP typing strategies. *Forens. Sci. Int.* 146(suppl):S139-S142.

Budowle, B. and Eisenberg, A.J. 2007. Forensic Genetics. In: Emery and Rimoin's Principles and Practice of Medical Genetics, fifth edition, Vol. 1, Rimoin, D.L., Connor, J.M., Pyeritz, R.E., and Korf, B.R. (eds.), Elsevier, Philadelphia, pp.501-517.

Budowle, B. and van Daal, A. 2008. Forensically relevant SNP classes. *Biotechniques* 44:603-610.

Budowle, B., Moretti, T.R., Niezgoda, S.J., and Brown, B.L. 1998. CODIS and PCR-based short tandem repeat loci: Law enforcement tools. In: Second European Symposium on Human Identification 1998, Promega Corporation, Madison, Wisconsin, pp. 73-88.

Budowle, B., Planz, J., Campbell, R., and Eisenberg, A. 2004. SNPs and microarray technology in forensic genetics: development and application to mitochondrial DNA. *Forens. Sci. Rev.* 16:22-36.

Budowle, B., Planz, J.V., Campbell, R., and Eisenberg, A.J. 2005. Molecular diagnostic applications in forensic science. In: *Molecular Diagnostics*, Patrinos, G. and Ansorge, W., (eds.), Elsevier, Amsterdam, pp. 267-280.

Budowle, B.: Familial searching: extending the investigative lead potential of DNA typing. *Profiles in DNA* 13(2), 2010, Available at: www.promega.com/profiles/1302/1302_07.html.

Butler, J.M., Decker, A.E., Vallone, P.M., and Kline, M.C. 2006. Allele frequencies for 27 Y-STR loci with U.S. Caucasian, African American, and Hispanic samples. *Forensic Sci. Int.* 156:250-260.

Carpenter, M.L. *et al.* 2013. Pulling out the 1%: whole-genome capture for the targeted enrichment of ancient DNA sequencing libraries. *Amer. J. Hum. Genet.* 93(5):852-864.

CASAVA v.1.8.2: http://support.illumina.com/downloads/casava_182.ilmn.

ChrX-STR.org 2.0: <http://www.chrx-str.org/>.

Churchill, J.D., Chang, J., Ge, J., Rajagopalan, N., Lagacé, R., Liao, W., King, J.L., and Budowle, B. (submitted). Blind study evaluation of the Ion PGM™ System for use in human identity DNA typing. *Croat. Med. J.*

CODIS-NDIS Statistics: <http://www.fbi.gov/about-us/lab/biometric-analysis/codis/ndis-statistics>.

Collins, P.J., Hennessy, L.K., Leibelt, C.S., Roby, R.K., Reeder, D.J., and Foxall, P.A. 2004. Developmental validation of a single-tube amplification of the 13 CODIS STR loci, D2S1338, D19S433, and amelogenin: the AmpFØSTR® Identifiler® PCR Amplification Kit. *J. Forensic Sci.* 49:1265–1277.

Complete Genomics FTP site: <ftp://ftp2.completegenomics.com>.

Craig, D.W., Pearson, J.V., Szelinger, S., Sekar, A., Redman, M., Corneveaux, J.J., Pawlowski, T.L., Laub, T., Nunn, G., Stephan, D.A., Homer, N., and Huentelman, M.J. 2008. Identification of genetic variants using bar-coded multiplexed sequencing. *Nat. Methods* 5:887–893.

D’Amato, M.E., Ehrenreich, L., Cloete, K., Benjeddou, M., and Davison, S. 2010. Characterization of the highly discriminatory loci DYS449, DYS481, DYS518, DYS612, DYS626, DYS644 and DYS710. *Forensic Sci. Int. Genet.* 4:104-110.

Davis, C., Ge, J., Sprecher, C., Chidambaram, A., Thompson, J., Ewing, M., Fulmer, P., Rabbach, D., Storts, D., and Budowle, B. 2013. Prototype PowerPlex® Y23 System: a concordance study. *Forensic Sci. Int. Genet.* 7:204-208.

Dixon, L.A., Murry, C.M., Archer, E.J., Dobbins, A.E., Koumi, P., and Gill, P. 2005. Validation of a 21-locus autosomal SNP multiplex for forensic identification purposes. *Forensic Sci. Int.* 154:62–77.

DNA Sequencing Costs: <https://www.genome.gov/sequencingcosts/>.

Edwards, A., Civitello, A., Hammond, H.A., and Caskey, C.T. 1991. DNA typing and genetic mapping with trimeric and tetrameric tandem repeats. *Am. J. Hum. Genet.* 49:746–756.

Edwards, A., Hammond, H.A., Jin, L., Caskey, C.T., and Chakraborty, R. 1992. Genetic variation at five trimeric and tetrameric repeat loci in four human population groups. *Genomics* 12:241–253.

Ellegren, H. 2004. Microsatellites: simple sequences with complex evolution. *Nat. Rev. Genet.* 5:435–445.

Excoffier, L. and Lischer, H.E.L. 2010. Arlequin suite ver 3.5: A new series of programs to perform population genetics analyses under Linux and Windows. *Mol. Ecol. Resour.* 10:564-567.

Flores, S., Sun, J., King, J., and Budowle, B. 2014. Internal validation of the GlobalFiler™ Express PCR Amplification Kit for the direct amplification of reference DNA samples on a high-throughput automated workflow. *Forensic Sci. Int. Genet.* 10:33-39.

Fordyce, S.L., Avila-Arcos, M.C., Rockenbauer, E., Børsting, C., Frank-Hansen, R., Petersen, F.T., Willerslev, E., Hansen, A.J., Morling, N., and Gilbert, M.T. 2011. High-throughput sequencing of core STR loci for forensic genetic investigations using the Roche Genome Sequencer FLX platform. *Biotechniques* 51:127–133.

Freire-Aradas, A., Fondevila, M., Kriegel, A.-K., Phillips, C., Gill, P., Prieto, L., Schneider, P.M., Carracedo, Á., and Lareu, M.V. 2012. A new SNP assay for identification of highly degraded human DNA. *Forensic Sci. Int. Genet.* 6:341–349.

GDA (Genetic Data Analysis): <http://www.eeb.uconn.edu/people/plewis/downloads/gda-1.1.win32.zip>

- Ge, J., Budowle, B., Aranda, X.G., Planz, J.V., Eisenberg, A.J., and Chakraborty, R. 2009. Mutation rates at Y chromosome short tandem repeats in Texas populations. *Forensic Sci. Int. Genet.* 3:179-184.
- Ge, J., Chakraborty, R., Eisenberg, A. and Budowle, B. 2011. Comparisons of the familial DNA database searching policies. *J. Forens. Sci.* 56(6):1448-1456.
- Ge, J., Eisenberg, A., and Budowle, B. 2012. Developing criteria and data to determine best options for expanding the core CODIS loci. *BMC Investigative Genetics* 3:1.
- Geppert, M., Edelmann, J., and Lessig, R. 2009. The Y-chromosomal STRs DYS481, DYS570, DYS576, and DYS643. *Leg. Med.* 11:S109-S110.
- Gill, P., Werrett, D.J., Budowle, B., and Guerrieri, R. 2004. An assessment of whether SNPs will replace STRs in national DNA databases – joint considerations of the DNA working group of the European Network of Forensic Science Institutes (ENFSI) and the Scientific Working Group on DNA Analysis Methods (SWGDM). *Sci. Justice.* 44:51–53.
- Gymrek, M., Golan, D., Rosset, S., and Erlich, Y. 2012. lobSTR: a short tandem repeat profiler for personal genomes. *Genome Res.* 22:1154–1162.
- Haplogroup Predictor: <http://www.hprg.com/hapest6/files/hapest6.htm?order=orig>.
- Hares, D.R. 2011. Expanding the CODIS core loci in the United States. *Forensic Sci. Int. Genet.* Doi:10.1016/j.fsigen.2011.04.012.
- Hares, D.R. 2015. Selection and implementation of expanded CODIS core loci in the United States. *Forensic Sci. Int. Genet.* 17:33-34.
- Hill, C.R., Butler, J.M., and Vallone, P.M. 2009. A 26plex autosomal STR assay to aid human identity testing. *J. Forens. Sci.* 54(5):1008-1015.
- Hill, C.R., Kline, M.C., Coble, M.D., and Butler, J.M. 2008. Characterization of 26 miniSTR loci for improved analysis of degraded DNA samples. *J. Forens. Sci.* 53(1):73-80.
- Holland, M.M., McQuillan, M.R., and O’Hanlon, K.A. 2011. Second generation sequencing allows for mtDNA mixture deconvolution and high resolution detection of heteroplasmy. *Croat. Med. J.* 52:299–313.
- Holt, K.E., Parkhill, J., Mazzoni, C.J., Roumagnac, P., Weill, F.X., Goodhead, I., Rance, R., Baker, S., Maskell, D.J., Wain, J., Dolecek, C., Achtman, M., and Dougan, G. 2008. High-throughput sequencing provides insights into genome variation and evolution in *Salmonella Typhi*. *Nat. Genet.* 40:987-993.

Holt, K.E., Parkhill, J., Mazzoni, C.J., Roumagnac, P., Weill, F.X., Goodhead, I., Rance, R., Baker, S., Maskell, D.J., Wain, J., Dolecek, C., Achtman, M., and Dougan, G. 2008. High-throughput sequencing provides insights into genome variation and evolution in *Salmonella Typhi*. *Nat. Genet.* 40:987-993.

Illumina® ForenSeq™ information: <http://www.illumina.com/products/forenseq-dna-signature-kit.html>

Illumina® GAIIX™ Specifications: http://www.illumina.com/documents/products/datasheets/datasheet_genome_analyzerIIx.pdf.

Illumina® MiSeq™ Specifications: http://www.illumina.com/documents/products/datasheets/datasheet_miseq.pdf.

Illumina® TruSeq™ Specifications: http://www.illumina.com/Documents/%5Cproducts%5Cdatsheets%5Cdatsheet_truseq_custom_enrichment_kit.pdf.

Kent, W.J., Sugnet, C.W., Furey, T.S., Roskin, K.M., Pringle, T.H., Zahler, A.M., and Haussler, D. 2002. The human genome browser at UCSC. *Genome Res.* 12(6):996–1006.

Koboldt, D.C., Chen, K., Wylie, T., Larson, D.E., McLellan, M.D., Mardis, E.R., Weinstock, G.M., Wilson, R.K., and Ding, L. 2009. VarScan: variant detection in massively parallel sequencing of individual and pooled samples. *Bioinformatics* 25:2283–2285.

Langmead, B., Trapnell, C., Pop, M., and Salzberg, S.L. 2009. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.* 10:R25.

Lazaruk, K., Walsh, P.S., Oaks, F., Gilbert, D., Rosenblum, B.B., Menchen, S., Scheibler, D., Wenz, H.M., Holt, C., and Wallin, J. 1998. Genotyping of forensic short tandem repeat (STR) systems based on sizing precision in a capillary electrophoresis instrument. *Electrophoresis* 19:86–93.

Margulies, M. *et al.* 2005. Genome sequencing in microfabricated high-density picolitre reactors. *Nature* 437:376-380.

Marshfield Clinic: <http://research.marshfieldclinic.org/genetics/GeneticResearch/sets/Set10PrimerSequences.htm>.

Martin, P.D., Schmitter, H., and Schneider, P.M. 2001. A brief history of the formation of DNA databases in forensic science within Europe. *Forens. Sci. Int.* 119(2):225-231.

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernysky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., and DePristo, M.A. 2010. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* 20:1297–1303.

Mengel-Jørgensen, J., Sanchez, J.J., Børsting, C., Kirpekar, F., and Morling, N. 2005. Typing of multiple single-nucleotide polymorphisms using ribonuclease cleavage of DNA/RNA chimeric single-base extension primers and detection by MALDI-TOF mass spectrometry. *Anal. Chem.* 77:5229–5235.

Mentype Chimera: http://www.biotype.de/fileadmin/user/MANUALS/190213_Manual_MentypeChimera_CE.pdf.

MiSeq™ Reporter: http://support.illumina.com/sequencing/sequencing_software/miseq_reporter/downloads.ilmn.

Mulero, J.J., Chang, C.W., Calandro, L.M., Green, R.L., Li, Y., Johnson, C.L., and Hennessy, L.K. 2006. Development and validation of the AmpFℓSTR® Yfiler™ PCR Amplification Kit: a male specific, single amplification 17 Y-STR multiplex system. *J. Forensic Sci.* 51:64-75.

Musgrave-Brown, E. *et al.* 2007. Forensic validation of the SNPforID 52-plex assay. *Forensic Sci. Int. Genet.* 1:186–190.

Myers, N.M., Ekins, J.E., Lin, A.A., Cavalli-Sforza, L.L., Woodward, S.R., and Underhill, P.A. 2007. Y-chromosome short tandem repeat DYS458.2 non-consensus alleles occur independently in both binary haplogroups J1-M267 and R1b3-M405. *Croat. Med. J.* 48:450-459.

Myers, N.M., Ritchie, K.H., Lin, A.A., Hughes, R.H., Woodward, S.R., and Underhill, P.A. 2009. Y-chromosome short tandem repeat intermediate variant alleles DYS392.2, DYS449.2, and DYS385.2 delineate new phylogenetic substructure in human Y-chromosome haplogroup tree. *Croat. Med. J.* 50:239-249.

NCBI: <http://www.ncbi.nlm.nih.gov/>.

Nextera™ Rapid Capture information: <http://www.illumina.com/products/nextera-rapid-capture-custom-enrichment-kit.html>.

Nielsen, R., Paul, J.S., Albrechtsen, A., and Song, Y.S. 2011. Genotype and SNP calling from next-generation sequencing data. *Nat. Rev. Genet.* 12:443–451.

NuGEN Encore™ 384 Multiplex System: <http://www.nugeninc.com/nugen/index.cfm/products/pl/library-preparation/encore-384-multiplex-system/>.

Olofsson, J., Andersen, M.M., Mogensen, H.S., Eriksen, P.S., and Morling, N. 2012. Sequence variants of allele 22 and 23 of DYS635 causing different stutter rates. *Forensic Sci. Int. Genet.* 6:e161-e162.

Oostdik, K. *et al.* 2014. Developmental validation of the PowerPlex® Fusion System for analysis of casework and reference samples: a 24-locus multiplex for new database standards. *Forensic Sci. Int. Genet.* 12:69-76.

Oostdik, K., French, J., Yet, D., Smalling, B., Nolde, C., Vallone, P.M., Butts, E.L., Hill, C.R., Kline, M.C., Rinta, T., Gerow, A.M., Allen, S.R., Huber, C.K., Teske, J., Krenke, B., Ensenberger, M., Fulmer, P., and Sprecher, C. 2013. Developmental validation of the PowerPlex® 18D system, a rapid STR multiplex for analysis of reference samples. *Forensic Sci. Int. Genet.* 7:129–135.

Ottaviani, E., Vernarecci, S., Asili, P., Agostino, A., and Montagna, P. 2014. Preliminary assessment of the prototype Yfiler® Plus kit in a population study of Northern Italian males. *Int. J. Legal Med.* Doi:10.1007/s00414-014-1078-x.

Pakstis, A.J., Speed, W.C., Fang, R., Hyland, F.C., Furtado, M.R., Kidd, J.R., and Kidd, K.K. 2010. SNPs for a universal individual identification panel. *Hum. Genet.* 127:315–324.

Park, P.J. 2009. ChIP-seq: advantages and challenges of a maturing technology. *Nat. Rev. Genet.* 10:669–680.

Penta D Facts Sheet (STRBase): http://www.cstl.nist.gov/strbase/str_Penta_D.htm.

Phillips, C., Fang, R., Ballard, D., Fondevila, M., Harrison, C., Hyland, F., Musgrave-Brown, E., Proff, C., Ramos-Luis, E., Sobrino, B., Carracedo, A., Furtado, M.R., Syndercombe Court, D., Schneider, P.M., and the SNPforID Consortium. 2007. Evaluation of the Genplex SNP typing system and a 49plex forensic marker panel. *Forensic Sci. Int. Genet.* 1: 180–185.

Pitterl, F., Schmidt, K., Huber, G., Zimmermann, B., Delport, R., Amory, S., Ludes, B., Oberacher, H., and Parson, W. 2010. Increasing the discrimination power of forensic STR testing by employing high-performance mass spectrometry, as illustrated in indigenous South African and Central Asian populations. *Int. J. Legal Med.* 124:551-558.

Planz, J.V., Sannes-Lowery, K.A., Duncan, D.D., Manalili, S., Budowle, B., Chakraborty, R., Hofstadler, S.A., and Hall, T.A. 2012. Automated analysis of sequence polymorphism in STR alleles by PCR and direct electrospray ionization mass spectrometry. *Forensic Sci. Int. Genet.* 6:594-606.

PPSS: <http://code.google.com/p/ppss/>.

Pumpernik, D., Oblak, B., and Borstnik, B. 2008. Replication slippage versus point mutation rates in short tandem repeats of the human genome. *Mol. Genet. Genomics* 279:53-61.

Qiagen® Investigator HDplex Kit:

https://www.google.com/url?sa=t&rct=j&q=&esrc=s&source=web&cd=3&cad=rja&uact=8&ved=0CC8QFjAC&url=http%3A%2F%2Fwww.qiagen.com%2Fresources%2Fdownload.aspx%3Fid%3D7d1661bd-a47b-4b19-a882-357a61b48c64%26lang%3Den&ei=f9tjVI7GJoiYyASKl4GQDw&usg=AFQjCNEXCa2eHD1yg5bzJl0AbkyB9k_4gg&sig2=uJ7RG8yki7pvbp_w4aVFkw&bvm=bv.79189006,d.aWw.

Quail, M.A., Kozarewa, I., Smith, F., Scally, A., Stephens, P.J., Durbin, R., Swerdlow, H., and Turner, D.J. 2008. A large genome center's improvements to the Illumina sequencing system. *Nat. Methods* 5:1005-1010.

Rothberg, J.M. *et al.* 2011. An integrated semiconductor device enabling non-optical genome sequencing. *Nature* 475(7356):348-352.

Sanchez, J.J., Phillips, C., Børsting, C., Balogh, K., Bogus, M., Fondevila, M., Harrison, C.D., Musgrave-Brown, E., Salas, A., Syndercombe-Court, D., Schneider, P.M., Carracedo, A., and Morling, N. 2009. A multiplex assay with 52 single nucleotide polymorphisms for human identification. *Electrophoresis* 27:1713–1724.

Sanger, F., Nicklen, S., and Coulson, A.R. 1977. DNA sequencing with chain-terminating inhibitors. *Proc Natl Acad Sci USA* 74(12):5463-5467.

Seo, S.B., King, J.L., Warshauer, D.H., Davis, C.P., Ge, J., and Budowle, B. 2013. Single nucleotide polymorphism typing with massively parallel sequencing for human identification. *Int. J. Legal Med.* 127:1079–1086.

Sorenson Molecular Genealogy Foundation Y Marker Details: http://www.smgf.org/ychromosome/marker_details.jsp.

SRM 2395 Human Y-Chromosome DNA Profiling Standard: <http://www.cstl.nist.gov/strbase/srm2395.htm>.

STRait Razor Unique Sequences Compiler (USC): http://web.unthsc.edu/info/200210/molecular_and_medical_genetics/887/research_and_development_laboratory/5.

STRBase: http://www.cstl.nist.gov/strbase/str_fact.htm.

STRBase Y-STR fact sheets: http://www.cstl.nist.gov/strbase/ystr_fact.htm.

Tajima, F. 1989. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* 123(3):585-595.

Templeton, J.E., Brotherton, P.M., Llamas, B., Soubrier, J., Haak, W., Cooper, A., and Austin, J.J. 2013. DNA capture and next-generation sequencing can recover whole mitochondrial genomes from highly degraded samples for human identification. *Investigative Genetics*, 4(1):1-13.

Thiede, C., Bornhauser, M., and Ehninger, G. 2004. Evaluation of STR informativity for chimerism testing – comparative analysis of 27 STR systems in 203 matched related donor recipient pairs. *Leukemia* 18: 248–254.

Tomas, C. *et al.* 2011. Autosomal SNP typing of forensic samples with the GenPlex™ HID System: results of a collaborative study. *Forensic Sci. Int. Genet.* 5:369–375.

Untergrasser, A., Cutcutache, I., Koressaar, T., Ye, J., Faircloth, B.C., Remm, M., and Rozen, S.G. 2012. Primer3—new capabilities and interfaces. *Nucleic Acids Res.* 40(15):e115.

Van Neste, C., Vandewoestyne, M., Van Criekinge, W., Deforce, D., and Van Nieuwerburgh, F. 2014. My-Forensic-Loci-queries (MyFLq) framework for analysis of forensic STR data generated by massive parallel sequencing. *Forensic Sci. Int. Genet.* 9:1–8.

Van Tassell, C.P., Smith, T.P., Matukumalli, L.K., Taylor, J.F., Schnabel, R.D., Lawley, C.T., Haudenschild, C.D., Moore, S.S., Warren, W.C., and Sonstegard, T.S. 2008. SNP discovery and allele frequency estimation by deep sequencing of reduced representation libraries. *Nat. Methods* 5:247-252.

Warshauer, D.H., Davis, C.P., Holt, C., Han, Y., Walichiewicz, P., Richardson, T., Stephens, K., Jager, A., King, J., and Budowle, B. 2015. Massively parallel sequencing of forensically relevant single nucleotide polymorphisms using TruSeq™ Forensic Amplicon. *Int. J. Legal Med.* 129(1):31-36.

Warshauer, D.H., King, J.L., and Budowle, B. 2015. STRait Razor v2.0: the improved STR Allele Identification Tool – Razor. *Forensic Sci. Int. Genet.* 14:182-186.

Warshauer, D.H., Lin, D., Hari, K., Jain, R., Davis, C., LaRue, B., King, J.L., and Budowle, B. 2013. STRait Razor: a length-based forensic STR allele-calling tool for use with second generation sequencing data. *Forensic Sci. Int. Genet.* 7:409–417.

Wheeler, D.A. *et al.* 2008. The complete genome of an individual by massively parallel DNA sequencing. *Nature* 452:872-876.

Wilson, M.R., DiZinno, J.A., Polanskey, D., Replogle, J., and Budowle, B. 1995. Validation of mitochondrial DNA sequencing for forensic casework analysis. *Int. Journal Leg. Med.* 108:68-74.

Wilson, M.R., Polanskey, D., Butler, J., DiZinno, J.A., Replogle, J., and Budowle, B. 1995. Extraction, PCR amplification, and sequencing of mitochondrial DNA from human hair shafts. *BioTechniques* 18:662-669.

Y-Chromosome Haplotype Reference Database: <http://www.yhrd.org/Research/Loci>.

Zeng, X., King, J., Hermanson, S., Patel, J., Storts, D.R., and Budowle, B. (submitted). An evaluation of the PowerSeq™ Auto system: a multiplex short tandem repeat marker kit compatible with massively parallel sequencing. *Forensic Sci. Int. Genet.*

Zeng, X., King, J.L., Stoljarova, M., Warshauer, D.H., LaRue, B.L., Sajantila, A., Patel, J., Storts, D.R., and Budowle, B. 2015. High sensitivity multiplex short tandem repeat loci analyses with massively parallel sequencing. *Forensic Sci. Int. Genet.* 16:38-47.