Cherian, Holly, Compiling an Autosomal and Y Chromosomal DNA Database for the Country of Sri Lanka. Master of Science (Forensic Genetics), August, 2009, 32 pp., 3 tables, 5 illustrations, references, 29 titles.

The purpose of this project was to compile an autosomal STR and Y chromosomal STR DNA database for Sri Lanka. Profiles from previous processing that are not interpretable will be redone. Additionally, on 10% of the samples a concordance study was conducted to check for reliability. The results proved to be concordant. Testing for Hardy Weinberg Equilibrium was conducted on the autosomal data. The database demonstrated to be in Hardy Weinberg Equilibrium, no linkage disequilibrium was observed, and there was no evidence of inbreeding present. A report was generated including tables of autosomal STR allele frequencies and summary of Y-STR haplotypes. Y-chromosomal haplotypes were reported and the population shows high genetic diversity with 168 haplotypes present.

COMPILING AN AUTOSOMAL AND Y CHROMOSOMAL DNA DATABASE FOR THE COUNTRY OF SRILANKA

INTERNSHIP PRACTICUM REPORT

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

MASTER OF SCIENCE

By

Holly Cherian, B.S

Fort Worth, Texas

August 2009

ACKNOWLEDGEMENTS

I would like to give my appreciation to my major professor Dr. Joseph Warren, who supported and guided me through every aspect of this project. In addition I would like to thank Dr. Arthur Eisenberg, Dr. John Planz and Dr. Bruce Budowle for the opportunity to complete my internship at the University of North Texas Health Science Center at Fort Worth. Also I would like to thank Patricia Gibson for her support and for allowing me to use the paternity lab for my internship. I am very grateful to Xavier G. Aranda for the encouragement and numerous hours of training in helping me with my internship. Linda Larose and Hector Seanz went out of there way many times to assist me with my project and I am appreciative for that. Most importantly I would like to thank God and my family, because without them I would not have been able to successfully complete my masters. My parents' and sister's prayers and love carried me through these past two years. Last but not least, I would like to thank Alexandra Newman who was a best friend and great help to me countless times not only academically but also emotionally.

TABLE OF CONTENTS

LIST OF FIGURES AND TABLES

FIGURES:

TABLES:

CHAPTER I

INTRODUCTION

The primary focus in this project is to compile an autosomal and Y chromosomal DNA database of natives from Sri Lanka at the University of North Texas Center for Human Identification (UNTCHI) in Fort Worth, Texas. The DNA is extracted from 2mm FTA blood card punches. Autosomal short tandem repeats (STRs) and Y chromosomal STRs will be typed and analyzed. Usable data that is generated from the 400 FTA blood card samples will be considered for the databases. Those profiles that were not interpretable will be requantified, amplified, and analyzed. This database will include 15 autosomal markers plus Amelogenin, and 17 Y chromosome markers. Haplotype frequencies for the Y chromosomal maker will be calculated using the counting method. Upper bound confidence intervals will also be determined (1). The autosomal marker database will undergo testing for Hardy Weinberg equilibrium (2). A results report will be generated for this population including tables of autosomal STR allele frequencies and summary of Y-STR haplotypes (3). Additionally, population analysis and biostatistical conclusions may also be drawn from the results obtained in comparison with publicly available databases. Once the data are shown to be reliable they can be used for kinship analyses and forensic casework.

The secondary focus in this project is to examine the data generated from previously extracted DNA. Once profiles have been interpreted, 40 samples for autosomal markers and 39 samples for Y chromosomal markers will be chosen for validation purposes. To validate

previously typed samples, they will be reprocessed by quantifying, amplifying and analyzing by capillary electrophoresis. After genetic analysis, they will be compared to previously obtained profiles. If the profiles are concordant, one may conclude that previously obtained data is reliable.

Background and Importance of DNA Databases:

The national DNA database used in the Unites States is called the Combined DNA Index System or CODIS. This database consists of DNA profiles that have been collected from convicted offenders, crime scene samples, relatives of missing persons, and unknown human remains. The benefit to these databases is that one may cross-search these databases to link a convicted offender to a crime scene, associate two different crime scenes to each other, or connect a relative to a unknown human remain. However, the foundation of the CODIS DNA database depends on a strong population database. A population database is used to determine the frequency of each allele from the different loci examined. This in turn can be used to statistically calculate the rarity of a DNA profile within the general population (4).

The importance of this project is to compile autosomal and Y-chromosomal databases of Sri Lankan samples. They will then be used for the calculation of statistics associated with forensic casework and missing persons cases; currently Sri Lanka does not have a DNA databases representative of its population. These databases will create strength in statistical results for relationship and forensic casework that is presented in court (5).

Currently, the United States uses three major populations including Caucasians, African Americans, and Hispanics for autosomal short tandem repeats (STRs), and five major populations for Y-STR analysis including African, Asian, Caucasian, Hispanic, and Native

American as references to compare statistical results in casework (1). While the integrity of autosomal STR DNA databases is not affected by the size of the database, Y-STRs follow Non-Mendelian inheritance patterns. Autosomal DNA follows Mendelian inheritance patterns. Mendel summarized the patterns of inheritance into two laws: law of segregation and law of independent assortment. The law of segregation states that during meiosis, both maternal and paternal chromosomes are passed down to the next generation. The law of independent assortment states that alleles of different genes assort independently of one another during meiosis. However, Y chromosomal DNA is patrilineraly inherited without recombination events during meiosis. It is inherited as a single unit or haplotype because there is only one copy of that chromosome to inherit. This means that their respective databases need to contain as much data as possible, by increasing the sample size. Developing a robust Y-STR haplotype reference DNA database of Sri Lankan samples will allow for the expansion of the common Y-STR databases not only for use in the United Stated but also in other parts of the world (1, 4, 5).

Genetic Marker Background and General Methodology:

Currently, forensics uses STR DNA markers to create a profile in human DNA identity testing. These markers consist of tandem repeat units of four base pairs long called a tetranucleotide. Due to their prevalence within the genome and lower stutter frequencies, tetranucleotides are the most commonly used marker in forensics. STR markers are useful because they work well with low quantity DNA templates and degraded DNA. They are also amendable to polymerase chain reaction (PCR) amplification, easily automated, and rapid analysis by fluorescent detection systems (2, 4). Additionally, when multiple STR locations are tested, it is highly discriminating due to the high polymorphism of each locus (2). Sample

throughput is further increased by the development of multiplex reaction kits, which allow for the amplification of 16 loci per PCR reaction (6, 7, 8).

Autosomal DNA is comprised of the DNA found in all 22 chromosomes, excluding the sex chromosomes. Half the autosomal DNA is inherited maternally and half is inherited paternally. During meiosis, the genetic information is shuffled by recombination, creating wide genetic diversity. In 1996, the Federal Bureau of Investigation (FBI) in the United States conducted a study involving 20 laboratories to select the core loci to be used in forensics. This was done to establish standardization among all laboratories in the United States (4). In 1997, the thirteen specific autosomal STR locations currently used in forensics were selected. These locations include: CSF1PO, FGA, TH01, TPOX, vWA, D3S1358, D5S818, D7S820, D8S1179, D13S317, D16S539, D18S51, and D21S11. These thirteen locations were selected due to unique characteristics, such as the number of alleles present, the type of repeat sequence, or the types of microvariants that have been observed. There are various commercial kits that include these thirteen loci and a number of additional loci (2). Some of these kits include Applied Biosystems AmpFℓSTR® kit and Promega's PowerPlex® kit (4, 7).

For the purposes of this experimental design, Applied Biosystems AmpFℓSTR® Identifiler™ PCR amplification kit will be used. This kit is a STR multiplex assay which amplifies 15 tetranucleotide repeat loci and the Amelogenin marker in a single PCR amplification. Within this kit, the 13 core CODIS loci are included. This kit uses a fluorescent system that includes five dyes in distinct spectral components, including: 6-FAM™, VIC™, NED™, PET™, and LIZ™ **(Figure 1)**. The fifth dye LIZ™ is used to label the GeneScan™ 500 Size Standard. Every fluorescent dye emits its maximum fluorescence at a different wavelength. During electrophoresis, the fluorescent signals are separated according to their wavelength and

then captured by a charge-coupled device (CCD) camera. Although all the dyes emit their fluorescence at various wavelengths, there is some overlap between the dyes (6, 7).

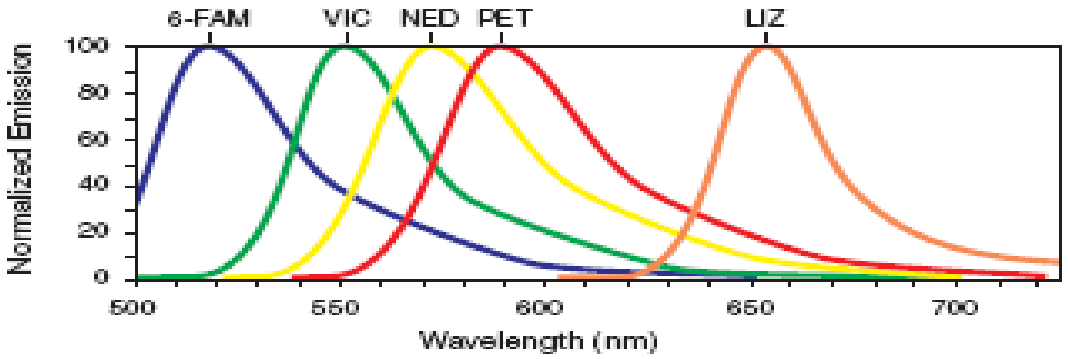


Figure 1-1 Emission spectra of the five dyes used in the AmpFℓSTR Identifiler PCR Amplification Kit

Figure 1: Emission Spectra of Five Dyes

The Y chromosome is found in males and contains the sex-determining region. The Y chromosome is inherited from the father to the son, and the genetic information does not change except for mutational events. This DNA is known as a paternal lineage marker. Because of the lack of recombination the individual STRs are inherited as a single unit. This is called a haplotype and behaves as single allele per individual. Y chromosomal DNA is beneficial in forensics because there are specific tests intended to examine only the male portion in the presence of high levels of female DNA, which is often seen in sexual assault cases. Y-STRs can also be useful in missing persons or mass disasters because this marker can expand the potential number of paternal family reference samples that can be obtained (1, 9). A Y-STR profile is shared by all members of the same paternal linage and cannot be used to distinguish a specific individual within that linage. This is because all individuals in a paternal linage share the same

allele combination, or haplotype, causing the statistical weight of discrimination to be decreased (10). There are numerous Y-STRs that have been identified (5). In 1997, Kayser and Pascali conducted a study in which eight loci were selected and integrated as the minimal haplotypes. Following this study, in 1993 the U.S. Scientific Working Group on DNA Analysis Methods (SWGDAM) recommended the use of this minimal haplotype plus two more additional loci. This was done to standardize the locations that are used in forensic casework at a national level. These loci include DYS19, DYS389I, DYS389II, DYS390, DYS391, DYS392, DYS393, DYS385a/b, DYS438, and DYS439. The loci DYS385a/b is included because it is known to be a highly polymorphic multi-copy locus. Furthermore, DYS385 is located on two different regions along the long arm of the Y chromosome. These two regions are approximately 40,000 base pairs apart, and can create two different alleles when amplified with a single set of primers (11). These loci are included in commercially available kits such as Applied Biosystems Y-Filer® kit and Promega's PowerPlex Y® kit (8,12).

For the purposes of this experimental design, Applied Biosystems Y-Filer® PCR Amplification kit will be used. This kit amplifies 17 Y STR Loci simultaneously. These loci include the European minimal haplotype, the recommended loci by the Scientific Working Group on DNA Analysis Methods (SWGDAM), and some additional highly polymorphic loci (11). Compared to Autosomal STRs, Y STRs exhibit a reduced power of discrimination due to a lack of recombination. To increase the power of discrimination, this kit was designed with 17 Y STR loci. The chemistry in which this kit performs is similar to that of the Applied Biosystems Identifiler® PCR Amplification kit (8,12).

All samples undergo specific tests to eventually generate profiles for analyses and comparison. After samples are obtained, the DNA is extracted from the cells. This process

involves the lysing of cells in order to release the DNA within the nucleus, which is then separated from other cellular material. Other cellular components such as proteins must be removed, because they can inhibit the ability to analyze the DNA. There are various methods for DNA extraction; however there are three common techniques that are used. Organic DNA extraction uses a series of chemicals to first break open the cell and the proteins that protect the DNA (13). Then phenol-chloroform is added to separate the proteins from the DNA. Another common technique that is faster is the Chelex® DNA extraction (14). A chelating-resin suspension is added to a sample, in which the cells become lysed and releases the DNA. Then a quick incubation at a high temperature will denature the DNA and disrupt the cell membrane and destroy other proteins that may cause inhibition. The last technique that is commonly used is an FTA™ paper DNA extraction (15, 16). This special paper is useful to store collected DNA blood samples because it contains a matrix which protects the sample from nuclease degradation and bacterial growth. When the sample comes into contact with the FTA paper, the cell immediately becomes lysed. The paper will then undergo a series of washes to release the DNA from the paper and remove heme and other inhibitors (15, 16).

Following DNA extraction, DNA quantitation takes place. This step ensures that DNA was in fact recovered from the DNA extraction and the DNA is human specific. Quantitation is essential, because the next step, PCR amplification, is optimized within a specific DNA concentration range. From the DNA quantification step, one is able to normalize the extracted DNA to the desired concentration. Too much template DNA within a reaction can cause split peaks, pull-up from dyes bleeding from one spectrum to another, stochastic effects such as unbalanced peak height ratios for heterozygote alleles and dye artifacts. On the contrary, too little template DNA can cause alleles to “drop-out” because amplification was unsuccessful, thus

sometimes leading to false homozygosity results (4, 18). Like extraction, there are several methods used for quantification, however today forensic laboratories commonly use real-time PCR (17, 18).

Following DNA quantification, PCR amplification will make many of copies of the template DNA at specific locations. This enzymatic process undergoes a series of heating and cooling parameters, which allows the target sequences to become amplified. In the first heating, the DNA becomes denatured; then specifically designed primers will identify the complimentary target portion of the DNA. This is followed by the primer extension, creating copies of the template DNA. During each cycle the rate at which the DNA is copied is exponential, creating multiple copies of the target region (6, 7, 8, 12). Lastly, the reactions are analyzed on a genetic analyzer capillary electrophoresis instrument. The process of capillary electrophoresis involves separating the amplified DNA fragments by size, using a polymer solution (19).

Validating and Methodology of Creating a Database:

There are several parameters used in order to create a population database. First, the laboratory decides the number of samples that will be tested for a specific ethnic group. Usually samples are obtained in the form of blood or buccal swabs. The samples obtained are typically convenience samples. An example of a convenience sample includes samples obtained from staff or blood banks from randomly chosen unrelated individuals. After the samples have been processed, there are specific tests that check for unrelatedness, which will be discussed later. The primary purpose in obtaining samples from unrelated individuals is to refine the precision of allele frequency estimates by increasing the number of independent genes sampled (20, 21, 22). Sometimes allele frequencies are categorized by a specific race or ethnicity, such as Native

American Apache Indian, and the allele frequency differences within a sample set are small and negligible. Ideally, DNA databases would include every individual in the world from every population. However, this is not feasible because of time and cost (20). Therefore, smaller databases are created that is representative of a population. This allows for a reliable prediction of allele and genotype frequencies in the entire population. Chakraborty conducted a study in 1992, addressing sample size requirements for a population to provide robust statistical results. He statistically concluded that 100 to 150 individuals per population could provide a sufficient sampling size that was 99% accurate when using allele frequencies for forensic purposes (3). After collection of samples, they are de-identified so that DNA typing results are not linked back to the donor. Samples then undergo extraction, PCR amplification, and genotyping at specific STR loci. Once the genotype data is summarized, allele frequencies are gathered. In nuclear DNA, each polymorphism is independently assorted, thus one may use the product rule to obtain a DNA profile frequency. The product rule is calculated by multiplying all the allele frequencies together from independently inherited loci (3). When allele frequencies are multiplied, the DNA profile becomes statistically very rare. To facilitate a reliable estimation of an allele frequency, more than one data point for that allele must be collected. This is to ensure that an allele that has been sampled adequately to be used dependably in statistical tests (20, 21, 22). According to the National Research Council (NRC II) report in 1996, an allele should be observed at least five times to be included within a database. This is recommended because if an allele is rare and is only expressed one time, the allele frequency can become very inaccurate. The minimum allele frequency is calculated as 5/2N, where N is the number of individuals that were sampled. The N is multiplied by 2 because autosomal chromosomes are inherited in pairs (23).

Once frequencies are obtained, specific statistical tests are conducted on the data to evaluate whether the database will be useful when applied to human identity testing. Certain software programs will test the independence of alleles within a locus and between multiple loci by checking Hardy-Weinberg equilibrium (HWE) proportions based on the observed allele frequencies. This predicts the stability of allele and genotype frequencies in consecutive generations because frequencies of alleles should not change over the course of several generations. Additional to checking for stability, HWE also examines any indication of excess homozygosity. This can usually be caused if Hardy Weinberg assumptions are violated. There are some basic assumptions when testing for Hardy Weinberg including the following: that there is random mating, there should be no natural selection, there should be no mutation that initiates new alleles, there should be no immigration or emigration which introduce new alleles, and the population in question should be large. Usually when inbreeding occurs genes that are not in random association are in linkage disequilibrium. Although the human population does not follow these assumptions perfectly, there is not significant deviation from it. For example, despite the fact we choose our mates, we do not select them because of a known DNA profile, therefore "random" mating occurs. When creating a database, specific test are conducted to determine if the population is within HWE. If the population follows Mendel's law of inheritance, then one can make a prediction on the accuracy of allele frequencies. Therefore, the frequencies of occurrence follow a predictable pattern of probability. In addition, one can statistically determine how powerful the loci are at individualizing a DNA profile; this is called the power of discrimination. How powerful the marker panel is at excluding particular combination of alleles is referred to as the power of exclusion (20, 21, 22).

Both expected genotype frequencies and observed genotype frequencies are compared, and if the values are similar, then it is assumed that alleles within a genetic locus are in equilibrium. There are several different types of HWE tests that are performed, such as the exact test and the likelihood ratio. There are two specific types of exact tests, namely the chi-square test and the Fisher test. This test indicates the closeness of fit between the expected and the observed genotypes. This is done by squaring the differences between the observed and expected genotypes in all the categories, and then dividing by the expected counts to give the greatest weight the largest proportional differences. Chi-square (χ^2) is calculated by the following formula:

$$\chi^2=\Sigma\ [(\text{observed-expected})^{\,2}/\text{expected}]$$

Similarly, the fisher exact test also indicates the closeness of fit between the expected and observed genotypes. The major difference is that the Fisher exact test is used for a smaller population size, where as the chi-square exact test is used for larger population sizes (20, 21). If the *p*-value is above 5% significance level, then the observed genotypes indicate no deviation from HWE. The likelihood ratio test is the comparison of two likelihood scenarios. The ratio is the hypothesized parameter (L_0) divided by the unconstrained value (L_1). Where the unconstrained value meets the parameters and assumptions of Hardy Weinberg, the likelihood ratio is defined as the following:

$$\lambda= L_0/ L_1$$

This ratio should be close to one when the hypothesis is true; otherwise it would be less than one (21, 22). Minor departures from HWE are usually not a major consideration because a perfect HWE cannot exist within a real human population. There are many reasons that contribute to this such as: the parents may be related therefore leading to inbreeding and a higher number of

expected homozygotes, population substructure, and selection (20). When testing for random match probability amongst multiple loci, the loci must be independent from one another indicating that recombination is occurring. Linkage equilibrium occurs when two DNA regions are transferred independently of another DNA segment during meiosis. Usually deviations from independence do not occur with the 13 core CODIS STRs because all of them are located on separate chromosomes except for CSF1PO and D5S818. Although CSF1PO and D5S818 are on the same chromosome, they are far enough apart that genetic recombination occurs. Parents often share some common ancestry that increases the number of homozygotes due to non-random mating. This population substructure is corrected for by using a theta (θ) correction factor (20). The NRC II report recommends using a θ of 0.01 for a general population and a θ of 0.03 for a small isolated population where inbreeding is more possible. Similarly to the θ, an Fst value can be calculated to measure population subdivision (23). Both θ and Fst measures population substructure by evaluating if mating within the subpopulation is random among populations (20, 23). In other words, these calculations assess the correlation of alleles within individuals and are related to inbreeding coefficients.

A Y-STR haplotype database is compiled by determining the haplotypes present within a population (24). These sequence frequencies are obtained by the counting method, where an analyst counts the number of observations of a profile, then divides that number by the size of the database to obtain the haplotype frequency (24, 25). Because the product rule cannot be used, the statistical power of the genetic test is vastly reduced. Additionally, the upper bound of the 95% confidence interval is accounted for in final frequency estimates. When using this method to compile a database, a large sample size is very important because of the Y chromosomal DNA paternal inheritance. A large sample size in a Y chromosomal database is vital because this in

turn causes the 95% confidence interval to have tighter bounds than if the sample size was small. In Y chromosomal DNA, the genetic information is inherited as a group and are linked, so their individual frequencies cannot be multiplied together to obtain the combined frequency (3, 5 24, 25). Therefore, a large sample size for haplotype frequencies is necessary to incorporate as many types of haplotypes as possible. Unless individuals are related, the haplotype for a person is not common amongst a population, thus the larger the sample size, the more various types of haplotypes is included within the database.

CHAPTER II

MATERIALS AND METHODS

DNA Samples and Extraction:

In order to compile this database, the University of North Texas Center for Human Identification has obtained 400 FTA blood card samples from the country of Sri Lanka. DNA was extracted using a Tecan Freedom EVO® 100 Robot with DNA IQ™ (Promega, Madison, WI) chemistry according to manufacturers recommended protocol. This extraction was conducted previously and raw data was obtained. Once the usable profiles were analyzed, the unusable profiles for both autosomal STRs and Y-chromosomal STRs were requantified, amplified, and underwent genetic analysis (15).

DNA Quantification:

The previously extracted samples first underwent quantification using the Quantifiler™ Human DNA Quantification Assay kit. This real-time PCR based DNA quantification kit is used to determine the quantity of amplifiable human or higher primate DNA within a sample. This in turn allows one to determine if there is an adequate amount of DNA for further testing, and how much sample needs to be used to optimize DNA amplification. The manufacturers recommended protocol was followed for quantification. The Applied Biosystems ABI PRISM® 7500 is the detection system which is used for the real-time PCR.

For each of the reactions, a master mix is added. This master mix contains Quantifiler human primer mix and Quantifiler PCR reaction mix. To each well 2µl of standard or sample is added to the appropriate well. Then 23µl of master mix is added to all the reactions, giving a total 25µl reaction volume.

Polymerase Chain Reaction Amplification:

PCR is a process used to amplify a specific region of DNA. This process is capable of generating multiple copies from a reasonably small amount of template DNA. For this project, the AmpFℓSTR® Identifiler™ PCR Amplification kit is used. Manufacturers recommended protocol was followed for autosomal STR amplification using 28 cycles. This particular kit uses primers that are labeled with fluorescent dyes for detection purposes during capillary electrophoresis. The table below illustrates the loci that are amplified and the corresponding dyes used for the AmpFℓSTR® Identifiler™ PCR Amplification kit.

Locus Designation	**Chromosome Location**	**Alleles Included in Identifiler Allelic Ladder**	**Dye Label**	**Control DNA 9947A**
D8S1179	8	8, 9 10, 11, 12, 13, 14, 15, 16, 17, 18, 19	6-FAM	13[a]
D21S11	21q11.2-q21	24, 24.2, 25, 26, 27, 28, 28.2, 29, 29.2, 30, 30.2, 31, 31.2, 32, 32.2, 33, 33.2, 34, 34.2, 35, 35.2, 36, 37, 38		30[b]
D7S820	7q11.21-22	6, 7, 8, 9, 10, 11, 12, 13, 14, 15		10, 11
CSF1PO	5q33.3-34	6, 7, 8, 9, 10, 11, 12, 13, 14, 15		10, 12
D3S1358	3p	12, 13, 14, 15, 16, 17, 18, 19	VIC	14, 15
TH01	11p15.5	4, 5, 6, 7, 8, 9, 9.3, 10, 11, 13.3		8, 9.3
D13S317	13q22-31	8, 9, 10, 11, 12, 13, 14, 15		11[c]
D16S539	16q24-qter	5, 8, 9, 10, 11, 12,13, 14, 15		11, 12
D2S1338	2q35-37.1	15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28		19, 23

D19S433	19q12-13.1	9, 10, 11, 12, 12.2, 13, 13.2, 14, 14.2, 15, 15.2, 16, 16.2, 17, 17.2	NED	14, 15
VWA	12p12-pter	11,12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24		17, 18
TPOX	2p23-2per	6, 7, 8, 9, 10, 11, 12, 13		8d
D18S51	18q21.3	7, 9, 10, 10.2, 11, 12, 13, 13.2, 14, 14.2, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27		15, 19
Amelogenin	X: p22.1-22.3 Y: p11.2	X, Y	PET	X
D5S818	5q21-31	7, 8, 9, 10, 11, 12, 13, 14, 15, 16		11e
FGA	4q28	17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 26.2, 27, 28, 29, 30, 30.2, 31.2, 32.2, 33.2, 42.2, 43.2, 44.2, 45.2, 46.2, 47.2, 48.2, 50.2, 51.2		23, 24

Figure 2: AmpFℓSTR® Identifiler™ Kit Loci and Dyes

Similarly the AmpFℓSTR® Y-filer™ PCR Amplification kit is used for Y-STR analysis. Manufacturers recommended protocol was followed for Y-STR amplification using 30 cycles. This kit also utilizes primers that are labeled with fluorescent dyes for detection purposes during capillary electrophoresis. The table below illustrates the loci that are amplified and the corresponding dyes used for the AmpFℓSTR® Yfiler™ PCR Amplification kit.

Locus Designation	Alleles Included in Yfiler Kit Allelic Ladder[a]	Dye Label	DNA 007 Genotype
DYS456	13—18	6-FAM™	15
DYS389I	10—15		13
DYS390	18—27		24
DYS389II	24—34		29
DYS458	14—20	VIC®	17
DYS19	10—19		15
DYS385 a/b	7—25		11,14
DYS393	8—16	NED™	13
DYS391	7—13		11
DYS439	8—15		12
DYS635	20—26		24
DYS392	7—18		13
Y GATA H4	8—13	PET®	13
DYS437	13—17		15
DYS438	8—13		12
DYS448	17—24		19

Figure 3: AmpFℓSTR® Yfiler™ Kit Loci and Dyes

To set up amplification, first data from quantification is evaluation. Necessary dilutions were made to the DNA extracts to obtain an optimal concentration ranging from 0.50ng/µl to 1.5ng/µl. Once proper dilutions have been made, a master mix that will be added to all the reactions is prepared. The master mix contains AmpFℓSTR® PCR reaction mix, AmpFℓSTR® Primer set, Amplitaq Gold® DNA polymerase. 2µl of sample or controls are added to the appropriate wells. The positive control that is used for AmpFℓSTR® Identifiler™ kit is 9947A.

The positive control that was used for Y-STR analysis in this project is 9948. Then 10µl of master mix is added to all the reactions, giving a total 12µl reaction volume (6, 7, 8, 9). The Applied Biosystems ABI PRISM® GeneAmp® PCR System 9700 was used for amplification.

Capillary Electrophoresis:

Capillary electrophoresis is a method in which DNA is separated by size in order to be analyzed. For the purposes of this project, the fragments were separated in POP-4™ polymer using the ABI PRISM® 3130xl Genetic Analyzer instrument was used for capillary electrophoresis and data collection. The analysis software program used for this project is Applied Biosystems GeneMapper® ID version 3.2, which has precise base sizing capabilities and designates appropriate allele calls.

To set up for capillary electrophoresis, a master mix is prepared containing HiDi Formamide and GeneScan™ 500 LIZ. The master mix for both Identifiler™ and Yfiler™ is the identical. 1µl of amplified STR product, controls and allelic ladder are added to the appropriate wells. 9µl of master mix is added to all the reactions, giving a total 10µl reaction volume (19).

Statistical Analysis:

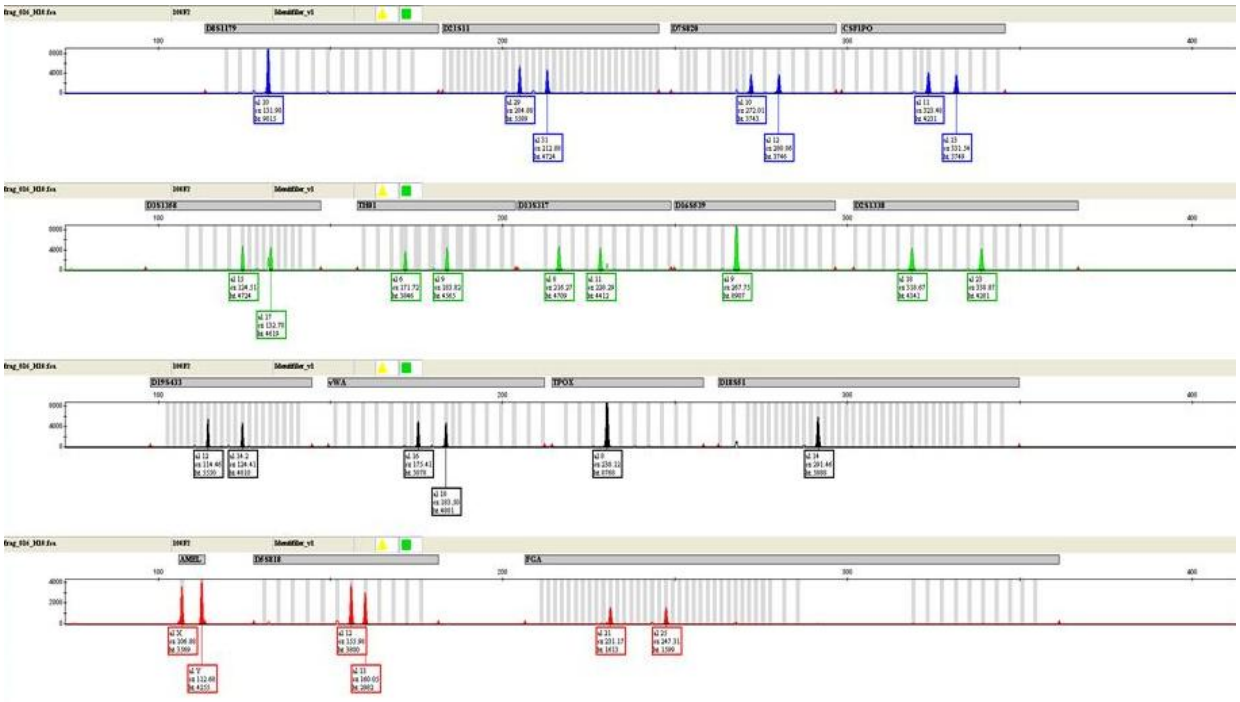
Once allele designations were made for all of the autosomal STR samples and Y-chromosomal STR samples, the frequencies for each locus were determined. For autosomal STRs the allele frequencies were determined using the counting method with the software program Genetic Data Analysis. Additional to allele frequencies, heterozygosity, expected heterozygosity, homozygosity, power of exclusion using the NRC II method, power of discrimination, mean power of exclusion using the Brenner method, determining Hardy

Weinberg equilibrium using χ^2 and Fisher method, Fis inbreeding coefficient, and checking the population for linkage was also determined. These calculations were determined using Genetic Data Analysis software (University of North Carolina), DNA Typing software (Ranajit Chakraborty) and Powerstats Version 12 software (Promega). Likewise, Y-chromosomal haplotype frequencies were determined using Arlequin Version 3.1 Software (20, 21, 22).

CHAPTER III

RESULTS AND DISCUSSION

The first phase of this project was to check concordance with previously determined profiles. Approximately 10% of the samples for both autosomal STRs and Y-chromosomal STRs were requantified, amplified and underwent capillary electrophoresis to check for concordance. All the samples that were ran showed no discrepancies with prior allele calls, and therefore it was concluded that all sample profiles determined previously are dependable and accurate. Below is an illustration of an example of an autosomal STR profile from the first run and the concordance run for sample number 100F2:



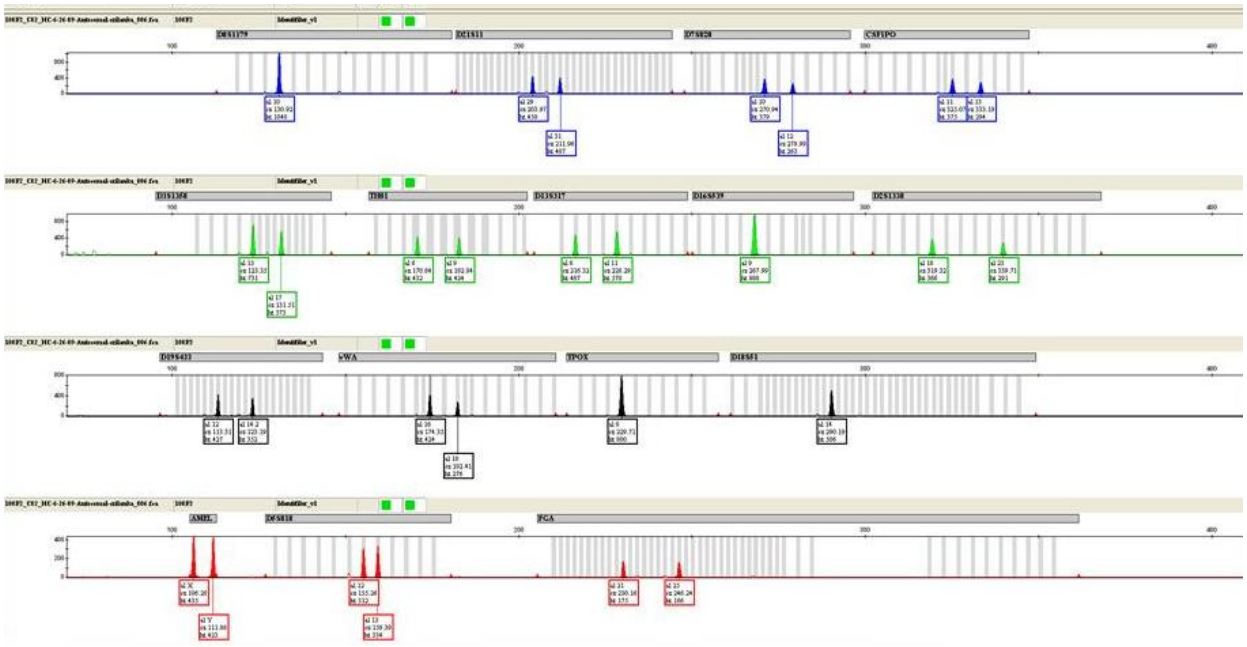
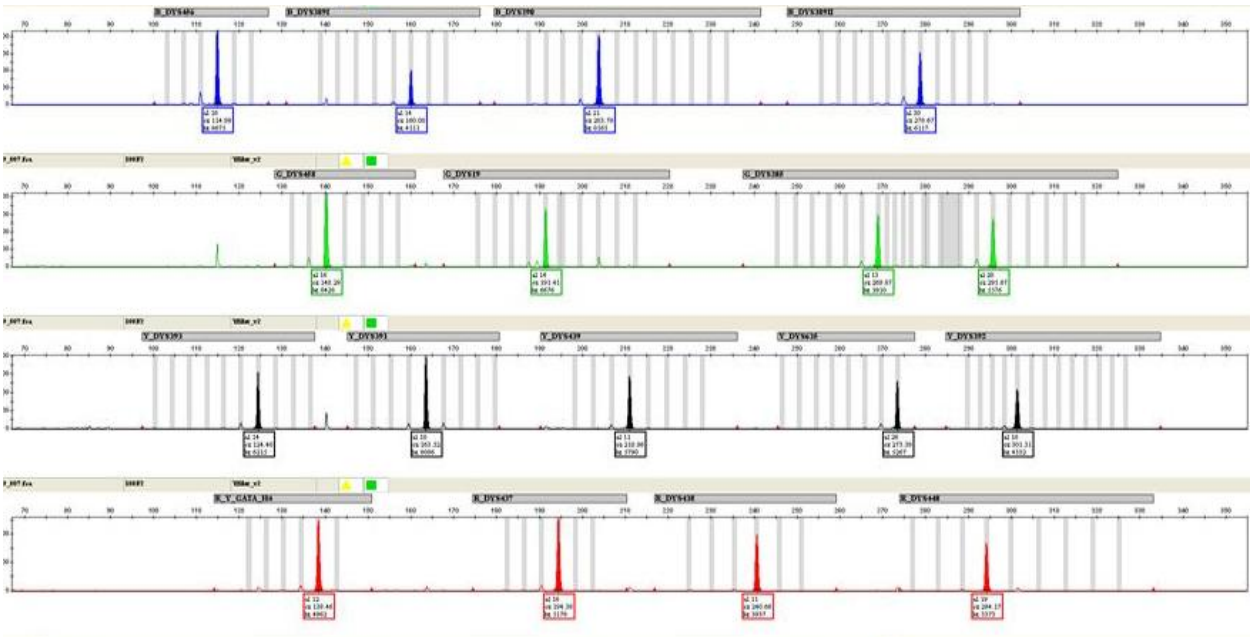


Figure 4: Autosomal STR Concordance Profiles

This demonstrates that the profiles obtained were exactly the same for both runs with no discrepancies. Below is an illustration of an example of a Y-chromosomal STR profile from the first run and the concordance run for sample number 100F2:



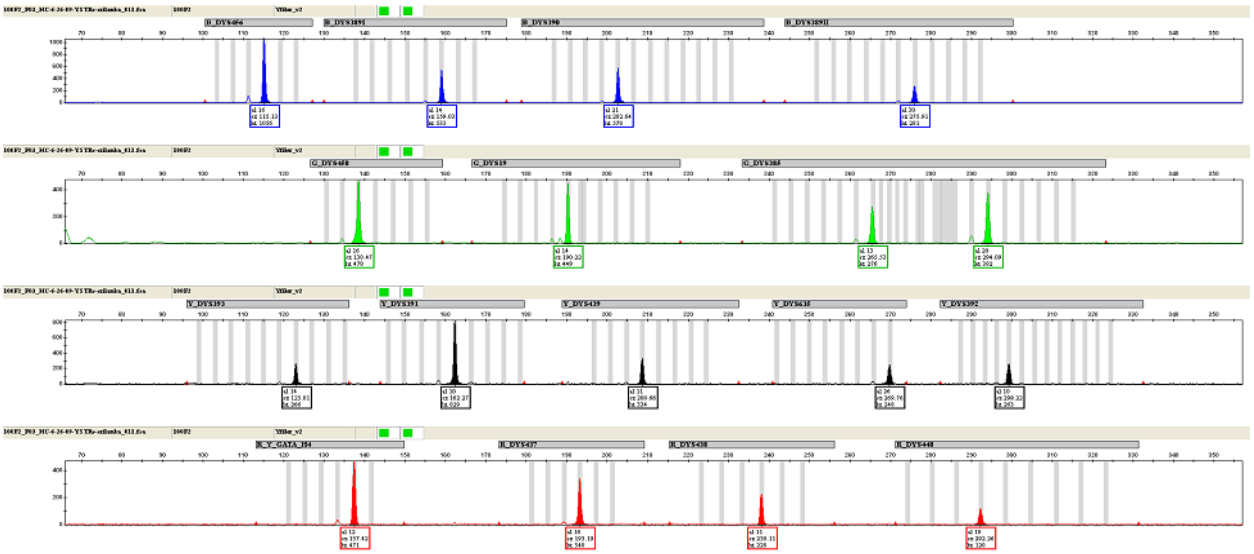


Figure 5: Y-Chromosomal STR Concordance Profiles

These figures reveal that the profiles obtained were exactly the same for both runs and without discrepancies. It was determined that previously obtained data as reliable for use in calculations for the Sri Lankan database because all samples checked were concordant.

The second phase of this project was to create an autosomal and Y-chromosomal DNA database for the country of Sri Lanka. First, profiles were determined for 377 autosomal STR samples and 171 Y STR samples. It was confirmed that none of the profiles were repeated within the sample set using DNA Typing software. For autosomal STRs, the profiles were entered into Genetic Data Analysis Version 1.1 software to determine the observed allele frequencies for each locus **(Table 1)**. Below is a table that reports the allele frequencies observed across all the loci for 377 autosomal STR samples:

ALLELE	D8S1179	D21S11	D7S820	CSF1PO	D3S1358	TH01	D13S317	D16S539	D2S1338	D19S433	vWA	TPOX	D18S51	D5S818	FGA
6	---	---	---	---	---	0.239	---	---	---	---	---	---	---	---	---
7	---	---	0.040	---	---	0.113	0.007	---	---	---	---	0.001	---	---	---
8	0.004	---	0.228	0.001	---	0.167	0.228	0.082	---	0.003	---	0.292	---	0.008	---
9	0.004	---	0.057	0.034	---	0.322	0.101	0.159	---	---	---	0.145	---	0.049	---
9.3	---	---	---	---	---	0.145	---	---	---	---	---	---	---	---	---
10	0.172	---	0.232	0.200	---	0.011	0.093	0.072	---	0.001	---	0.106	0.012	0.085	---
10.3	---	---	0.003	---	---	0.004	---	---	---	---	---	---	---	---	---
11	0.081	---	0.220	0.276	0.001	---	0.237	0.301	---	0.005	---	0.428	0.025	0.371	---
11.1	---	---	---	---	---	---	---	0.001	---	---	---	---	---	---	---
12	0.088	---	0.187	0.405	---	---	0.228	0.228	---	0.076	---	0.023	0.070	0.312	---
12.1	---	---	---	---	---	---	---	0.003	---	---	---	---	---	---	---
12.2	---	---	---	---	---	---	---	---	---	0.020	---	---	---	---	---
13	0.162	---	0.031	0.077	0.004	---	0.081	0.134	---	0.316	0.001	0.005	0.137	0.159	---
13.2	---	---	---	---	---	---	---	---	---	0.033	---	---	---	---	---
14	0.187	---	0.003	0.004	0.031	---	0.025	0.020	---	0.259	0.153	---	0.284	0.013	---
14.2	---	---	---	---	---	---	---	---	---	0.048	---	---	---	---	---
15	0.192	---	---	0.003	0.275	---	---	---	---	0.097	0.107	---	0.168	0.003	---
15.2	---	---	---	---	---	---	---	---	---	0.072	---	---	---	---	---
16	0.086	---	---	---	0.282	---	---	---	0.011	0.050	0.199	---	0.149	---	---
16.2	---	---	---	---	---	---	---	---	---	0.016	---	---	---	---	---
17	0.023	---	---	---	0.267	---	---	---	0.066	0.004	0.272	---	0.069	---	0.003
17.2	---	---	---	---	---	---	---	---	---	0.001	---	---	---	---	---
18	0.001	---	---	---	0.125	---	---	---	0.180	---	0.159	---	0.034	---	0.008
19	---	---	---	---	0.016	---	---	---	0.178	---	0.094	---	0.029	---	0.056
20	---	---	---	---	---	---	---	---	0.098	---	0.015	---	0.009	---	0.123
21	---	---	---	---	---	---	---	---	0.034	---	---	---	0.004	---	0.154
21.2	---	---	---	---	---	---	---	---	---	---	---	---	---	---	0.001
22	---	---	---	---	---	---	---	---	0.069	---	---	---	0.009	---	0.152
22.2	---	---	---	---	---	---	---	---	---	---	---	---	---	---	0.005
22.3	---	---	---	---	---	---	---	---	---	---	---	---	---	---	0.001
23	---	---	---	---	---	---	---	---	0.168	---	---	---	---	---	0.170
23.2	---	---	---	---	---	---	---	---	---	---	---	---	---	---	0.001
24	---	---	---	---	---	---	---	---	0.117	---	---	---	---	---	0.176
24.2	---	---	---	---	---	---	---	---	---	---	---	---	---	---	0.001
25	---	---	---	---	---	---	---	---	0.060	---	---	---	---	---	0.089
25.2	---	---	---	---	---	---	---	---	---	---	---	---	---	---	0.003
25.3	---	---	---	---	---	---	---	---	---	---	---	---	---	---	0.001
26	---	---	---	---	---	---	---	---	0.012	---	---	---	---	---	0.044
26.3	---	---	---	---	---	---	---	---	---	---	---	---	---	---	0.001
27	---	0.011	---	---	---	---	---	---	0.005	---	---	---	---	---	0.007
27.2	---	---	---	---	---	---	---	---	---	---	---	---	---	---	0.001
28	---	0.151	---	---	---	---	---	---	0.001	---	---	---	---	---	---
28.2	---	0.001	---	---	---	---	---	---	---	---	---	---	---	---	---
29	---	0.180	---	---	---	---	---	---	---	---	---	---	---	---	---
29.2	---	0.005	---	---	---	---	---	---	---	---	---	---	---	---	---
30	---	0.158	---	---	---	---	---	---	---	---	---	---	---	---	---
30.2	---	0.025	---	---	---	---	---	---	---	---	---	---	---	---	---
30.3	---	0.001	---	---	---	---	---	---	---	---	---	---	---	---	---
31	---	0.048	---	---	---	---	---	---	---	---	---	---	---	---	0.001
31.2	---	0.122	---	---	---	---	---	---	---	---	---	---	---	---	---
32	---	0.004	---	---	---	---	---	---	---	---	---	---	---	---	0.001
32.2	---	0.195	---	---	---	---	---	---	---	---	---	---	---	---	---
33	---	0.001	---	---	---	---	---	---	---	---	---	---	---	---	---
33.1	---	0.001	---	---	---	---	---	---	---	---	---	---	---	---	---
33.2	---	0.084	---	---	---	---	---	---	---	---	---	---	---	---	---
34.1	---	0.001	---	---	---	---	---	---	---	---	---	---	---	---	---
34.2	---	0.008	---	---	---	---	---	---	---	---	---	---	---	---	---
35.2	---	0.003	---	---	---	---	---	---	---	---	---	---	---	---	---

Table 1: Autosomal STR Allele Frequencies

Once allele frequencies were determined, further analysis was conducted on the data using Genetic Data Analysis® software, DNA Typing software, and Powerstats® Version 12 software. These additional tests were run to prove that the population is within Hardy Weinberg equilibrium **(Table 2)**. This infers that there is random mating occurring, the population is free of migration, there is a lack of mutation occurring, there is no inbreeding present, and no natural selection occurring. The observed heterozygosity was reported and it was seen that there was a high percentage of heterozygosity. Additional to observed heterozygosity, the expected heterozygosity for each locus was calculated. This is calculated by first determining the expected homozygosity, and then subtracting that value from one to obtain the expected heterozygosity. The expected homozygosity is determined by taking the sum of all the allele frequencies squared within each locus **(Table 2)**. This is what is expected within a population that is in Hardy Weinberg equilibrium, because it indicates there is high genetic variation within the population. On the contrary, observed homozygosity had a low percent of homozygosity across all loci. This in turn shows high genetic variation within the population and demonstrates that Hardy Weinberg equilibrium assumptions were followed. The power of exclusion was calculated using both the NRC II method and the Brenner method. The probability of exclusion is the probability of excluding a random individual from the population other than the tested individual. The primary difference between the NRC II method and Brenner method is that a theta correction is used for the NRC II method. A theta correction of 0.01 is used to correct for excess homozygosity within this population. The power of exclusion that was observed was relatively high, indicating there is a high discrimination across all loci because of the high number of alleles present and high genetic variation. Similarly a high power of discrimination was observed across all loci. Additional to these tests, the Fis was calculated which can indicate if inbreeding is

present within the population. All of the loci exhibited a very low Fis value, denoting that the population has very little to no inbreeding present. Lastly, Hardy Weinberg equilibrium was calculated using the χ^2 and Fisher method. All of the loci passed Hardy Weinberg equilibrium with an alpha level of 0.05 or higher, except at locus D21S11, D19S433, and D18S51 using the χ^2 method **(Table 2)**. However, after calculating a Bonferroni Correction, which changes the alpha level to 0.00064, this caused the locus D21S11, D19S433, and D18S51 to have a value above the alpha level, and therefore passing Hardy Weinberg equilibrium. Lastly, statistical tests were run using Genetic Data Analysis software and DNA Typing software to check if there is linkage present within this population. It was observed that there was a possible presence of linkage present between the D21S11 and vWA loci. However, observing linkage disequilibrium one time within the population is a normal phenomenon within a smaller population. In addition, below is a table that reports the additional tests performed to evaluate Hardy Weinberg equilibrium:

	D8S1179	D21S11	D7S820	CSF1PO	D3S1358	TH01	D13S317	D16S539	D2S1338	D19S433	vWA	TPOX	D18S51	D5S818	FGA
Hetero	0.838	0.817	0.822	0.698	0.719	0.745	0.814	0.790	0.870	0.756	0.833	0.682	0.801	0.719	0.854
Hetero (exp)	0.851	0.858	0.806	0.714	0.758	0.779	0.815	0.803	0.871	0.808	0.818	0.700	0.839	0.731	0.866
Homo	0.162	0.183	0.178	0.302	0.281	0.255	0.186	0.210	0.130	0.244	0.167	0.318	0.199	0.281	0.146
PE (NRC)	0.697	0.711	0.612	0.470	0.527	0.567	0.631	0.613	0.738	0.635	0.637	0.454	0.684	0.498	0.728
PD	0.958	0.963	0.931	0.871	0.902	0.918	0.938	0.931	0.967	0.940	0.940	0.860	0.954	0.886	0.965
PE (Mean)	0.672	0.631	0.641	0.425	0.458	0.502	0.626	0.581	0.735	0.520	0.661	0.401	0.601	0.458	0.703
HWE (CHI)	0.482	0.023	0.422	0.482	0.076	0.121	0.987	0.546	0.945	0.011	0.465	0.449	0.047	0.602	0.483
HWE (Fisher)	0.204	0.152	0.764	0.298	0.384	0.625	0.926	0.506	0.474	0.201	0.937	0.149	0.080	0.390	0.419
Fis	0.016	0.045	-0.020	0.024	0.046	0.044	-0.004	0.012	0.000	0.068	-0.017	0.021	0.045	0.016	0.018

Table 2: Autosomal STR Hardy Weinberg Equilibrium Tests (Hetero: heterozygosity; Hetero (exp): expected heterozygosity; Homo: homozygosity; PE (NRC): power of exclusion using the

NRCII method; PD: power of discrimination; PE (Mean): power of exclusion using the Brenner method; HWE (CHI): Hardy Weinberg Equilibrium χ^2 value; HWE (Fisher): Hardy Weinberg Equilibrium using the Fisher method; Fis: inbreeding coefficient.

Additional to creating an autosomal DNA database for the country of Sri Lanka, this database was compared to other Southeast Asian population databases. This was performed to ensure confidence within the database created for Sri Lanka, by comparing the values obtained with other Southeast Asian populations. The following population databases were used for comparison purposes: Tamilian population in South India, three various subpopulations of Bihar located in East India, the Bangladeshi population, the Indonesian population, and the East Timorian population (26, 27, 28, 29). After a general comparison of values was completed, it was concluded that all of these populations share similar statistical values and only approximately a 10% difference between values was observed which is negligible. Therefore, because the Sri Lankan population database has similar values as expected to that of other Southeast Asian populations, the database which was created can be used with confidence for forensic purposes.

A total of 171 samples were typed using Applied Biosystems Yfiler kit for 17 loci. Y-chromosomal STR allele frequencies were also determined using Arlequin Version 3.1 software. It was confirmed that duplicate samples were not present within the sample set. From a total of 171 samples analyzed, there were 168 different haplotypes observed. The table below reports all the haplotypes and the frequency that it was observed for this population.

Number	HAPLOTYPE																	Frequency
	DYS389I	DYS389II	DYS390	DYS456	DYS19	DYS385a	DYS385b	DYS458	DYS437	DYS438	DYS448	GATA_H4	DYS391	DYS392	DYS393	DYS439	DYS635	
1	14	32	24	15	15	13	17	15	16	10	19	11	11	13	11	11	23	0.011695906
2	13	30	25	17	15	13	15	17	14	9	19	12	10	11	13	11	23	0.011695906
3	13	29	20	17	15	16	17	18	14	9	19	11	10	11	12	11	20	0.011695906
4	14	30	21	16	14	13	20	16	16	11	19	12	10	10	14	11	26	0.005847953
5	14	30	23	14	15	15	20	19	14	9	21	11	9	11	12	12	23	0.005847953
6	14	31	23	15	15	12	19	20	16	11	19	11	10	10	14	11	24	0.005847953
7	13	29	20	16	15	15	15	16	15	10	18	11	10	11	14	11	21	0.005847953
8	12	28	22	16	14	13	19	16	15	10	19	12	10	14	11	12	24	0.005847953
9	13	32	24	16	16	11	14	15	14	11	20	12	11	11	13	10	23	0.005847953
10	13	29	23	16	15	15	15	17	14	9	19	12	10	11	12	11	21	0.005847953
11	13	29	24	16	15	12	18	17	14	11	17	10	10	11	13	8	22	0.005847953
12	12	28	22	15	14	13	19	15	15	10	18	12	10	14	11	12	25	0.005847953
13	13	31	24	15	14	16	17	15	14	10	20	12	10	11	14	11	18	0.005847953
14	14	30	24	15	15	15	17	18	14	9	19	11	10	11	12	11	20	0.005847953
15	14	30	22	16	15	15	17	16	14	9	19	11	10	11	12	11	22	0.005847953
16	14	32	26	15	15	11	14	17	14	11	20	13	10	11	13	10	23	0.005847953
17	14	30	23	14	15	11	19	17	16	11	19	12	9	10	14	11	25	0.005847953
18	12	28	22	15	14	14	18	14	15	11	19	11	10	14	11	11	24	0.005847953
19	12	28	22	15	14	13	17	15	15	10	19	11	10	14	11	13	22	0.005847953
20	13	31	25	17	15	11	14	15	14	11	20	13	11	11	12	10	23	0.005847953
21	13	31	25	15	16	11	14	17	14	11	20	11	11	11	13	10	23	0.005847953
22	14	29	23	15	16	12	20	19	16	11	19	11	10	10	14	10	24	0.005847953
23	12	28	24	14	14	12	16	18	14	9	19	11	10	12	12	14	21	0.005847953
24	12	30	24	16	16	11	14	16	14	11	20	13	11	11	13	10	23	0.005847953
25	13	29	22	17	15	16	17	17	14	9	19	12	10	11	12	11	22	0.005847953
26	13	29	25	15	15	11	14	17	14	11	20	13	10	11	13	11	23	0.005847953
27	13	29	22	15	14	14	15	13	14	10	18	12	10	11	12	10	21	0.005847953
28	13	29	23	15	14	13	17	16	16	11	19	12	10	10	14	10	25	0.005847953
29	14	30	23	15	14	13	16	16	15	9	21	11	10	11	13	12	24	0.005847953
30	14	30	23	14	15	13	20	18	16	11	19	12	10	10	14	10	25	0.005847953
31	13	29	24	16	14	12	16	18	16	11	19	11	10	10	14	10	24	0.005847953
32	13	29	24	13	15	12	16	14	15	9	18	11	10	11	12	13	20	0.005847953
33	14	30	22	16	16	16	18	17	14	7	19	12	10	11	12	11	19	0.005847953
34	12	28	22	15	14	11	16	15	15	10	19	12	10	14	11	12	23	0.005847953
35	14	31	25	15	16	11	14	16	14	11	20	12	10	11	13	11	24	0.005847953
36	12	28	22	15	14	13	16	14	15	10	19	12	10	14	11	12	24	0.005847953
37	14	31	22	15	17	15	18	19	15	10	19	11	10	11	13	12	20	0.005847953
38	13	30	21	15	15	16	16	18	14	11	18	11	10	11	14	11	21	0.005847953
39	14	32	25	15	15	11	14	16	14	11	20	13	10	11	13	10	23	0.005847953
40	13	28	21	15	14	13	20	16	14	9	20	12	10	11	14	11	21	0.005847953
41	13	31	22	16	14	15	20	17	15	9	20	12	10	11	13	11	22	0.005847953
42	13	30	23	15	14	14	19	16	15	9	20	11	11	12	13	11	21	0.005847953
43	12	28	22	15	14	13	18	14	15	10	19	12	9	15	11	12	24	0.005847953
44	13	29	22	15	15	16	16	16	14	9	19	12	10	11	12	12	20	0.005847953
45	13	29	23	15	14	13	17	16	16	11	20	12	10	10	14	10	24	0.005847953

46	14	30	23	15	15	15	19	17	14	9	19	12	10	12	12	11	20	0.005847953
47	12	28	22	17	14	13	16	15	15	10	19	12	10	14	11	12	23	0.005847953
48	12	29	24	13	15	14	18	15	15	9	19	12	10	11	12	13	21	0.005847953
49	14	31	21	16	14	14	17	20	14	10	19	11	11	11	14	12	21	0.005847953
50	13	31	23	16	12	13	18	15	14	10	19	11	10	14	11	12	25	0.005847953
51	12	30	23	15	17	12	14	15	14	11	20	13	10	11	13	10	23	0.005847953
52	13	32	25	15	15	11	14	16	14	9	20	12	11	11	13	10	23	0.005847953
53	13	29	22	15	14	13	17	15	15	10	19	12	10	14	11	11	23	0.005847953
54	14	30	23	15	14	14	17	18	14	11	19	11	10	10	14	11	26	0.005847953
55	13	29	22	15	15	14	16	16	14	9	20	13	10	11	12	11	19	0.005847953
56	12	28	24	13	15	11	17	16	15	9	20	11	11	11	13	12	22	0.005847953
57	13	29	22	16	15	15	16	19	14	9	18	11	11	12	12	11	20	0.005847953
58	13	30	23	15	15	11	19	19	16	11	19	12	10	10	13	10	25	0.005847953
59	13	29	24	15	15	13	16	17	15	10	19	12	10	12	13	13	24	0.005847953
60	13	30	24	15	15	13	15	17	14	10	20	12	9	11	13	11	24	0.005847953
61	13	29	23	16	14	12	18	18	16	11	18	12	10	10	15	11	28	0.005847953
62	14	30	22	15	16	9	16	18	16	10	19	11	10	14	12	11	21	0.005847953
63	14	30	22	16	15	15	16	17	14	9	19	12	10	11	12	11	20	0.005847953
64	12	28	22	16	14	13	16	14	15	10	20	12	11	14	11	12	24	0.005847953
65	14	31	22	15	16	19	19	17	14	9	19	12	10	11	12	12	20	0.005847953
66	12	28	22	15	14	14	17	13	15	10	19	12	10	14	11	13	24	0.005847953
67	12	28	24	13	15	13	18	15	14	9	19	11	10	11	12	12	21	0.005847953
68	13	31	25	16	16	11	14	15	14	11	20	12	11	11	13	10	23	0.005847953
69	13	30	23	15	13	16	18	18	14	10	20	12	9	11	14	12	21	0.005847953
70	13	31	24	15	15	11	14	16	14	11	20	13	10	11	13	11	23	0.005847953
71	13	30	24	15	16	11	14	16	14	11	20	12	11	11	13	10	23	0.005847953
72	11	27	23	15	14	12	19	16	14	12	19	12	11	14	13	11	23	0.005847953
73	14	29	23	15	13	14	19	18	16	11	19	12	10	10	13	11	25	0.005847953
74	14	30	22	16	15	16	18	18	14	9	19	12	10	11	12	11	20	0.005847953
75	13	29	21	15	13	14	18	16	15	10	18	12	10	13	13	11	21	0.005847953
76	12	28	22	16	14	14	17	15	14	11	19	11	10	15	11	12	22	0.005847953
77	12	28	22	15	14	14	17	16	15	11	19	12	10	14	11	12	23	0.005847953
78	12	28	22	16	14	13	17	14	15	10	20	12	10	14	11	13	22	0.005847953
79	13	29	24	15	14	11	15	16	15	12	20	13	11	13	12	12	23	0.005847953
80	13	30	24	15	17	11	14	16	14	11	20	12	11	11	13	10	23	0.005847953
81	13	28	24	14	14	15	15	15	14	10	18	11	10	11	13	11	20	0.005847953
82	13	31	25	15	15	11	14	17	14	11	20	13	10	11	13	10	23	0.005847953
83	13	30	24	15	15	13	13	15	15	9	19	11	10	11	12	11	21	0.005847953
84	13	32	25	15	15	11	14	16	14	11	20	13	11	11	13	10	23	0.005847953
85	12	28	23	15	14	13	15	14	15	10	19	12	10	14	11	12	23	0.005847953
86	13	29	22	16	15	16	16	15	14	9	18	12	10	11	12	12	20	0.005847953
87	12	29	22	15	15	9	15	19	16	10	20	11	10	14	12	11	22	0.005847953
88	12	28	24	15	16	15	16	19	14	10	22	11	10	12	13	12	21	0.005847953
89	12	28	23	13	15	12	15	19	16	11	20	12	10	11	13	13	21	0.005847953
90	11	27	23	13	15	12	18	16	15	9	20	11	11	11	12	12	21	0.005847953
91	13	30	25	14	15	13	15	17	14	9	20	12	11	11	13	11	22	0.005847953
92	13	31	24	15	14	14	17	18	14	10	19	12	10	11	14	11	20	0.005847953
93	13	29	25	15	15	12	17	18	14	9	19	12	10	11	13	11	21	0.005847953
94	13	29	21	15	16	16	18	16	14	8	19	11	10	11	12	11	20	0.005847953
95	12	28	22	15	14	13	13	16	16	10	20	11	10	11	14	12	21	0.005847953

96	14	32	24	15	15	11	14	16	14	9	20	12	10	11	13	10	23	0.005847953
97	14	31	23	16	15	13	21	17	16	11	19	12	10	10	15	12	26	0.005847953
98	14	30	23	15	14	13	19	18	16	11	19	12	10	10	14	10	25	0.005847953
99	12	28	22	16	14	13	17	14	15	10	20	12	10	14	11	13	23	0.005847953
100	13	30	25	15	15	11	15	18	14	9	20	12	10	11	13	10	25	0.005847953
101	12	28	24	13	15	13	16	15	15	9	19	11	10	11	12	11	21	0.005847953
102	13	29	24	14	16	12	16	17	14	10	19	12	10	11	12	12	21	0.005847953
103	14	29	25	15	14	13	14	17	16	11	19	12	10	10	14	11	25	0.005847953
104	14	31	25	15	15	11	14	16	14	11	20	12	10	11	13	10	23	0.005847953
105	12	28	22	16	14	13	17	15	15	10	18	12	10	14	11	13	24	0.005847953
106	12	28	22	17	14	14	17	14	15	10	19	12	10	14	11	12	24	0.005847953
107	13	30	22	16	15	16	16	16	14	9	19	12	10	11	12	11	21	0.005847953
108	13	29	22	16	15	16	17	16	14	9	19	12	10	11	12	11	21	0.005847953
109	14	32	22	15	16	11	14	17	14	11	20	13	11	11	13	10	23	0.005847953
110	14	33	24	14	15	11	14	16	14	11	20	14	11	11	13	10	24	0.005847953
111	13	30	22	15	16	13	14	17	14	10	19	12	10	11	12	12	21	0.005847953
112	13	30	25	15	16	11	14	16	14	11	20	13	11	11	13	10	23	0.005847953
113	14	32	27	16	15	11	14	16	14	11	20	14	11	11	13	10	23	0.005847953
114	13	28	23	15	14	12	20	17	16	11	19	11	11	10	14	11	24	0.005847953
115	13	31	22	17	15	15	14	20	16	10	19	12	9	11	15	11	21	0.005847953
116	13	29	22	16	15	13	18	17	14	9	19	12	10	11	12	12	20	0.005847953
117	14	32	24	15	13	14	16	16	14	10	19	12	10	11	14	10	18	0.005847953
118	13	29	24	14	15	15	16	16	14	10	18	11	10	11	13	11	20	0.005847953
119	12	27	22	16	14	13	16	14	15	10	19	12	10	14	11	13	23	0.005847953
120	12	28	22	16	14	13	16	14	15	10	20	12	11	14	11	12	22	0.005847953
121	14	31	24	15	14	13	19	17	16	11	19	11	10	10	13	11	25	0.005847953
122	13	30	25	15	16	11	14	15	14	11	20	12	10	11	14	10	23	0.005847953
123	13	30	24	15	14	15	17	16	14	10	19	11	10	11	14	11	18	0.005847953
124	12	28	23	15	14	14	19	17	16	11	19	11	11	10	15	12	24	0.005847953
125	13	30	23	16	14	15	18	18	14	10	19	12	10	11	14	11	18	0.005847953
126	13	29	22	16	15	7	16	17	16	10	19	11	10	14	12	12	22	0.005847953
127	12	29	24	14	16	13	18	16	15	9	19	11	10	12	12	12	23	0.005847953
128	13	30	25	15	16	11	14	16	14	12	20	13	10	11	13	10	23	0.005847953
129	12	28	22	15	14	13	16	15	15	10	19	11	10	14	11	12	21	0.005847953
130	13	31	23	16	16	11	14	16	14	11	20	11	10	11	14	10	23	0.005847953
131	13	29	22	16	16	14	16	17	14	9	19	11	11	11	13	11	20	0.005847953
132	13	32	21	15	17	15	20	18	15	10	18	11	10	11	13	12	24	0.005847953
133	13	29	22	17	15	16	17	17	14	9	19	11	10	11	12	11	20	0.005847953
134	12	29	24	13	15	12	17	16	15	9	20	11	10	11	12	12	19	0.005847953
135	13	32	25	15	15	11	14	15	14	11	19	12	11	11	13	11	23	0.005847953
136	14	30	22	16	15	16	17	16	14	9	20	12	10	11	12	12	20	0.005847953
137	12	28	22	17	14	13	17	14	15	10	19	12	10	14	11	13	21	0.005847953
138	11	27	24	16	14	12	18	17	14	10	17	11	10	11	13	8	21	0.005847953
139	14	32	25	14	15	11	14	17	14	11	20	11	10	11	14	10	23	0.005847953
140	14	30	23	15	14	13	16	18	15	11	17	11	10	10	14	11	26	0.005847953
141	13	31	24	16	16	11	14	15	14	11	20	13	10	11	13	10	23	0.005847953
142	12	28	22	15	14	13	17	15	15	10	19	12	10	14	11	13	24	0.005847953
143	12	28	22	16	15	13	17	14	15	10	19	11	10	14	11	12	23	0.005847953
144	13	31	24	16	15	13	14	16	14	12	19	12	10	11	14	13	22	0.005847953
145	14	31	23	15	14	14	17	18	16	11	20	12	10	10	14	11	25	0.005847953

146	14	30	23	15	14	13	18	16	16	11	19	11	10	9	13	10	26	0.005847953
147	12	30	22	15	12	13	17	15	15	10	19	13	10	14	11	12	24	0.005847953
148	13	29	22	15	14	13	18	15	15	10	19	12	10	15	11	12	24	0.005847953
149	12	28	26	13	14	13	17	17	15	9	19	11	11	11	12	11	21	0.005847953
150	13	29	25	15	15	11	14	16	14	11	20	12	10	11	13	10	24	0.005847953
151	12	29	23	17	16	13	14	16	14	10	16	12	10	14	13	11	21	0.005847953
152	13	31	22	18	15	15	15	17	16	10	19	11	10	11	15	11	22	0.005847953
153	14	31	23	15	14	13	17	18	16	11	19	12	10	10	15	11	24	0.005847953
154	14	32	26	15	15	11	14	15	14	11	20	12	10	11	13	10	22	0.005847953
155	13	30	25	15	16	11	14	16	14	11	20	12	10	11	13	10	23	0.005847953
156	13	28	22	16	14	13	20	16	15	10	19	13	10	14	11	12	23	0.005847953
157	13	29	24	14	14	14	14	16	14	10	19	12	10	11	14	11	18	0.005847953
158	12	27	23	15	15	13	19	18	15	11	20	11	10	14	12	13	20	0.005847953
159	13	29	25	17	16	11	13	17	14	11	20	12	10	11	13	11	23	0.005847953
160	14	30	22	15	16	14	17	18	14	9	19	11	10	11	12	11	24	0.005847953
161	14	32	25	15	16	11	14	18	14	11	20	13	10	11	13	11	23	0.005847953
162	12	29	22	14	15	14	15	16	15	10	21	12	10	11	13	11	22	0.005847953
163	12	28	22	16	14	12	14	14	15	10	19	11	10	14	11	13	22	0.005847953
164	12	27	23	15	15	12	17	15	14	11	20	11	10	11	13	13	20	0.005847953
165	13	29	20	14	16	15	20	17	14	8	19	11	10	11	12	11	20	0.005847953
166	14	30	22	15	14	7	16	18	16	10	19	11	10	14	12	11	22	0.005847953
167	13	29	23	13	14	15	19	18	14	9	21	11	10	11	12	11	23	0.005847953
168	13	30	22	15	15	15	15	19	16	10	19	12	11	11	14	12	22	0.005847953

Table 3: Y-Chromosomal STR Haplotype Frequencies

Population substructure affects Y-haplotypes more drastically than autosomal loci, because of its inheritance patterns. In this study, it was seen that there was fairly a large amount of genetic diversity because only three samples from a total of 171 samples had the same haplotype appear.

Additional to creating a Y-chromosomal DNA database, 10% of the Y-chromosomal samples were chosen randomly and entered into the Applied Biosystems Yfiler® haplotype database. This was completed to determine the frequency of these haplotypes within this database. The Applied Biosystems Yfiler® haplotype database has a sample size of 11,393 different haplotypes. After the database was searched, it was concluded that the 10% of samples had a frequency of zero within this database. This illustrates that, as expected, the haplotypes observed within the Sri Lankan population is unique to its own.

CHAPTER IV

CONCLUSIONS

In the first phase of this project, it was concluded that the initial data that was obtained was reliable and accurate. This was determined by processing 10% of both autosomal and Y-chromosomal DNA samples. Concordance of samples was determined and verified, which allowed for the use of all samples to be analyzed and included within the database.

In the second phase of this project, an autosomal and Y-chromosomal DNA database for the country of Sri Lanka was created. Profiles were determined for 377 autosomal and 171 Y-chromosomal DNA samples. Allele frequencies for autosomal STR loci were obtained by using Genetic Data Analysis Version 1.1 software. Additional tests were run using Genetic Data Analysis® software, DNA Typing software and Powerstats® Version 12 software, to determine if this population is within Hardy Weinberg equilibrium. It was observed that there was a high percentage of heterozygosity and a low percentage of homozygosity as expected within this particular population. This indicated that there is high genetic diversity within this population and it follows Hardy Weinberg assumptions. Also the observed and expected heterozygosity values were extremely similar as anticipated. Correspondingly, the power of exclusion and power of discrimination was calculated to determine if the population has high genetic variance present. It was seen that there was a relatively high power of exclusion and high power of discrimination present across all loci. The Fis was also calculated to determine if there is inbreeding present within the population. It was observed that there were very low Fis values

across all loci, which indicates very little inbreeding present within this population. Linkage disequilibrium tests were also run, and there was no linkage present within the population, other than what was predicted. Lastly, the Hard Weinberg equilibrium was calculated using the χ^2 and Fisher method. In conclusion, it was determined that after all statistical tests were run, the Sri Lankan population follows Hardy Weinberg equilibrium assumptions. Furthermore, when the Sri Lankan population's statistical values were compared to other Southeast Asian populations, it was concluded that the database can be used with confidence because the values are comparable as expected.

Similarly, haplotype frequencies for Y-chromosomal STR loci were obtained using Arlequin Version 3.1 software. From a total of 171 samples analyzed, there were 168 different haplotypes observed. In this study, it was seen that there was fairly a large amount of genetic diversity due to the high amount of various haplotypes observed relative to the total sample set size. Additionally, when 10% of the Sri Lankan Y-chromosomal DNA samples were searched against the Applied Biosystems Yfiler® haplotype database, it was concluded that the Sri Lankan population haplotypes unique to its own population because they were not observed within this database.

REFERENCES

1. Budowle, Bruce., Adamowicz, Mike., Aranda, Xavier G., Barna, Charles., Chakraborty, Ranajit., et al. "Twelve Short Tandem Repeat Loci Y Chromosomal Haplotypes: Genetic Analysis on Populations Residing in North America." Jouranl of Forensic Science International. May 2005. Volume 150(1). (1-15).
2. Budowle, Bruce., Moretti, TR., Baumstark, AL., Defenbaugh, DA., Keys, KM. "Population Data on the Thirteen CODIS Core Short Tandem Repeat Loci in African Americans, U.S. Caucasians, Hispanics, Bahamians, Jamaicans, and Trinidadians." Journal of Forensic Science. November 1999. Volume 44(6). (1277-1286).
3. Chakraborty, Ranajit., Jin, L., Zhong, Y., Srinivasan, MR., Budowle, B. "On Allele Frequency Computation from DNA Typing Data." International Journal of Legal Medicine. 1993. Volume 105(4). (233-238).
4. Butler, John M. Forensic DNA Typing : Biology, Technology, and Genetics of STR Markers. New York: Academic P, 2005
5. Butler, John M., Kline, Margaret C., Decker, Amy E. "Adressing Y-Chromosome Short Tandem Repeat Allele Nomenclature." Journal of Genetic Genealogy. 2008. Volume 4(2). 125-148.
6. Collins, PJ., Hennessy LK., Leibelt CS., Roby RK. "Developmental Validation of a Single-Tube Amplification of the 13 CODIS STR Loci, D2S1338, D19S433,and

Amelogenin: the AmpFℓSTR® Identifiler® PCR Amplification Kit."Journal of Forensic Science. 2004. Volume 49 (6). (11265-11277).

7. Applied Biosystems. "AmpFℓSTR® Identifiler™ PCR Amplification Kit Users Manual." Foster City, California. August 2006. (1-186).
8. Applied Biosystems. "AmpFℓSTR® Yfiler™ PCR Amplification Kit Users Manual." Foster City, California. August 2006. (1-234).
9. Krenke, BE., Viculis, L., Richard ML., Prinz, M., Milne, SC., Ladd, C., Gross, AM., et al. "Validation on Male Specific, 12-Locus Fluorescent Short Tandem Repeat Multiplex." Journal of Forensic Science International. June 2005. Volume 151(1). (111-124).
10. Ge, Jianye., Budowle, Bruce., Aranda, Xavier G., Planz, John V., Eisenberg, Arthur J., Chakraborty, Ranajit. "Mutation Rates at Y Chromosome Short Tandem Repeats in Texas Populations." Journal of Forensic Science International. June 2009. Volume 3(3). (179-184).
11. Shewale, Jaiprakash G., Bhushan, Anurag., Nasir, Huma., Schneida, Elaine. "Population Data for Four Population Groups from the United States for the Eleven Y-Chromosome STR Loci Recommended by SWGDAM." Journal of Forensic Sciences. May 2006. Volume 51(3). (700-702).
12. Mulero, JJ., Chang, CW., Calandro, LM., Green, RL. "Development and Validation of the AmpFℓSTR® Yfiler® PCR Amplification Kit." Journal of Forensic Sciences. 2006. Volume 51 (1). (64-75).
13. Hoff-Olsen, P., Mevog, B., Staalstrom, E., Hovde, B., Edeland, T., Olaisen, B. "Extraction of DNA from Decomposed Human Tissue. An Evaluation of Five

Extraction Methods for Short Tandem Repeat Typing." Journal of Forensic Sciences International. November 1999. Volume 105 (3). (1718-1783).

14. Walsh, PS., Metzger, DA., Higichi, R. "Chelex 100 as a Medium for Simple Extraction from DNA for PCR Based Typing from Forensic Material." Journal of Biotechniques. April 1991. Volume 10 (4). (506-513).
15. Belgrader P., Del Rios, SA., Turner, KA., Marino, MA., Waever, KR., Williams, PE. "Automated DNA Purification and Amplification from Blood Stained Cards Using a Robotic Workstation." Journal of Biotechniques. May 1995. Volume 19 (3). (427-432).
16. Aranda, XG., Campbell, RS., Planz, JV., Smith, MA., Igoe, F., Eisenberg, AJ. "FTA Technology. Unique Formats for the Collection Shipment, Archiving, and Processing of Biological Samples." Promega 12th International Symposium on Human ID. 2001.
17. Applied Biosystems. "AmpFℓSTR® Quantifiler™ Human DNA Quantification Kit and Quantifiler™ Y Human Male DNA Quantification Kit Users Manual." Foster City, California. April 2006. (1-205).
18. Green, R., Roinestad, I., Boland, C., Hennessy, L. "Developmental Validation of the Quantifiler Real-Time PCR Kits for the Quantification of Human Nuclear DNA Samples." Journal of Forensic Sciences. July 2005. Volume 50 (4). (809-825).
19. Applied Biosystems. "GeneMapper ID Software Version 3.1 Human Identification Analysis Users Manual." Foster City, California. December 2003. (1-440).

20. Hartl, D.L., Clark, A.G. Principles of Population Genetics. Sunderland, Massachusetts: Sinauer Associates, 1997.
21. Weir, Bruce. Genetic Data Analysis II. Sunderland, Massachusetts. Sinuar Associates Publishers. 1996.
22. Weir, Bruce. Population Genetics and Statistics for Forensic Biology. Raleigh, North Carolina. 1996.
23. National Research Council U.S. "The National Research Council Report on The Evaluation of Forensic DNA Evidence." 2006. Volume 2. (1-241)
24. National Center for Forensic Science. "Database Descriptive Statistics." March 1, 2009. <http://usystrdatabase.org//pdf/DatabaseDescriptiveStatistics.pdf>.
25. Roth, Andrea J.D., Schmechel, Richard J.D. "Presenting Mitochondrial DNA Statistics in the Courtroom." Promega. December 18, 2008. <http://www.promega.com/GENETICIDPROC/ussymp16proc/abstracts/14schmechel.pdf>.
26. Panneerchelvam, S., Vanaja, N., Baskar, D., Sivapriya, V., Damodaran, C. "Distribution of Alleles of 12 STR Loci in Tamil Population (South India)." Journal of Forensic Science International. October 2000. Volume 119. (126-128).
27. Ashma, R., Kashyap, V.K. "Genetic Polymorphism at 15 STR Loci among Three Important Subpopulation of Bihar, India." Journal of Forensic Science International. September 2002. Volume 130. (58-62).
28. Dobashi, Y., Kido, A., Fujitani, N., Susukida, R., Oya, M. "Population Data of Nine STR Loci, D3S1358, vWA, FGA, TH01, TPOX, CSF1PO, D5S818,

D13S317 and D7S820 in Bangladeshis and Indonesians." Journal of Forensic Science International. April 2003. Volume 135. (72-74).

29. Souto, L., Alves, C., Gusmão, L., Ferreira, E., Amorium, A., Côrte-Real, F., Vieira, D.N. "Population Data on 15 Autosomal STRs in a Sample from East Timor." Journal of Forensic Science International. January 2005. Volume 155. (77-80).